NUMANA: A Hybrid Numerical and Analytical Thermal Simulator for 3-D ICs

Yu-Min Lee[1], Tsung-Heng Wu[1], Pei-Yu Huang[2], Chi-Ping Yang[1]
[1]National Chiao Tung University, Taiwan, [2]Industrial Technology Research Institute, Taiwan
yumin@nctu.edu.tw, brian518094@gmail.com, chippy623.cm00g@nctu.edu.tw, pyhuang@itri.org.tw

Abstract— **By combining analytical and numerical simulation techniques, this work develops a hybrid thermal simulator, NUMANA, which can effectively deal with complicated material structures, to estimate the temperature profile of a 3-D IC. Compared with a commercial tool, ANSYS, its maximum relative error is only** 1.84%. **Compared with a well known linear system solver, SuperLU [1], it can achieve orders of magnitude speedup.**

I. Introduction

As the semiconductor industry strives to maintain the trend of Moore's Law, the 3-D IC design scheme is one way to achieve the interconnect density, manufacturing yields and cost targets. Moreover, it provides the flexibility for mixed signal design, the suitability for circuits operating on different supply voltages and the capability for heterogeneous integration. However, the operating temperature of 3-D ICs dramatically increases because of the high power density and the ill of heat dissipation capability. To verify the thermal reliability, thermal analyzers are needed for different physical design stages.

Thermal simulation techniques can be categorized into numerical and analytical methods. With the finite difference method (FDM), numerical methods spatially mesh a die into numerous control volumes, and each control volume is modeled as an equivalent thermal sub-circuit. Then, an equivalent thermal network is built, and the modified nodal analysis (MNA) method can be used to build its system equations.

However, to obtain accurate temperature distribution, direct solvers are runtime and memory inefficient since they require solving a huge MNA system with millions of unknowns. Therefore, several methods [2]–[7] have been developed to reduce the runtime and the memory usage. Since the detail thermal characteristics in control volumes can be taken into account, numerical methods are able to handle complex material structures. Therefore, although [2]–[7] are for 2-D ICs, they also provide the flexibility for 3-D ICs.

Different from the numerical methods [2]–[7], analytical thermal simulation methods [8], [9] construct the closed-form representation for spatial bases and apply them to explicitly approximate and fast evaluate the temperature distribution. Comparing with numerical methods, analytical methods gain the fundamental benefit that no MNA system needs to be solved. However, the assumption of homogeneous-material structure is required. This is a severe limitation for their applications on design stages of general 3-D ICs since the material structure of general 3-D ICs should not be viewed as a homogeneous medium. Though Huang and Lee [9] have extended their generalized integral transform based thermal simulator (GIT) to the structures of wire bonded, microbump-3D package or contactless-interconnection 3-D ICs, the material in a layer still needs to be homogeneous.

To take the advantages of both numerical and analytical methods, a hybrid thermal simulator, NUMANA, which finds out a practical scheme for combining numerical and analytical frameworks, will be developed for general structures of 3-D ICs. The major advantages of NUMANA are

1) From the aspect of simulation frameworks, NUMANA not only can handle complex material structures by using the FDM modeling technique, but also can avoid dealing with the large scale MNA system by adopting the analytical based approach to be the simulation kernel. Hence, NUMANA not only can be as flexible as numerical methods, but also can be as fast as analytical approaches.
2) From the aspect of practical applications, since no MNA system needs to be solved, NUMANA can efficiently update temperature variations corresponding to the modified material structure induced by reallocating TSVs/TTSVs, microbumps or wires.

This paper is organized as follows. Section II states the problem formulation, and the FDM based thermal modeling technique and GIT [9] are reviewed in section III. Then, NUMANA is detailed in section IV. Finally, experimental results and conclusion are given in sections V and VI, respectively.

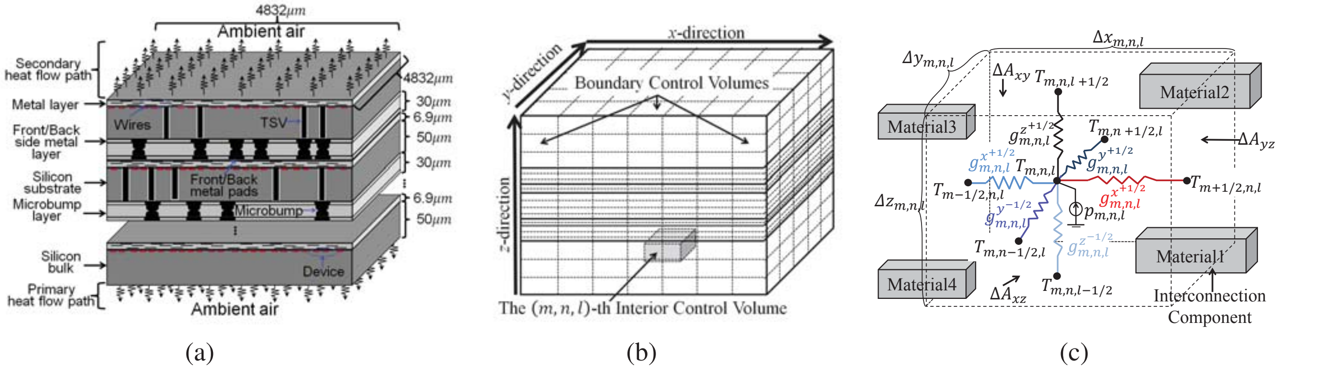
II. Problem Formulation

The physical model of a 3-D IC is exhibited in Fig. 1.(a). The primary heat flow path consists of heat spreader, heat sink and package. The secondary heat flow path is composed of I/O pads, the package substrate and PCB. Due to the small sidewall area, the lateral boundary surfaces are assumed to be adiabatic. Metal layers contain wires and dielectric material. Front/back side metal layers contain front/back side metal pads and wires. Microbump layers contain microbumps and dielectric material, and TSVs/TTSVs pass through silicon substrates. Devices are distributed in the junction region of active layers and are modeled as power sources. Interconnection components are also treated as power sources because the self-heating effect.

Generally, the steady-state temperature distribution, $T(\mathbf{r})$, is more concerned in physical design engines, and $T(\mathbf{r})$ can be governed by heat transfer equations as

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

This work was supported in part by National Science Council of Taiwan under Grant NSC 101-2221-E-009-168, and Industrial Technology Research Institute under Grant B5-10125-2J-3.
978-3-9815370-0-0/DATE13/©2013 EDAA



(a)

(b)

(c)

Fig. 1. Physical model and schematic plots of the FDM based thermal modeling technique for a 3-D IC. (a) Physical model of a TSV-based 3-D IC. (b) Mesh structure of layers. (c) The equivalent thermal resistance sub-circuit of an interior control volume.

subject to the boundary condition

$$\kappa(\mathbf{r}_{b_s})\frac{\partial T(\mathbf{r}_{b_s})}{\partial \vec{n}_{b_s}} + h_{b_s}(\mathbf{r}_{b_s})T(\mathbf{r}_{b_s}) = f_{b_s}(\mathbf{r}_{b_s}). \tag{2}$$

Here, $\mathbf{r} = (x, y, z)$ is an arbitrary position on the chip, $\kappa(\mathbf{r})$ is the distribution of thermal conductivities (W/m ·°C), and $p(\mathbf{r})$ is the power density function of heat sources (W/m^3). b_s is any specific boundary surface of the chip, $h_{b_s}(\mathbf{r}_{b_s})$ is the distribution of heat transfer coefficients on b_s, $\partial/\partial\vec{n}_{b_s}$ is the differentiation along the outward direction normalized to b_s, and $f_{b_s}(\mathbf{r}_{b_s})$ is an arbitrary function on b_s.

III. Review of Previous Works

A. FDM Based Thermal Modeling Technique

As shown in Fig. 1.(b), each layer is meshed separately to avoid control volumes crossing different layers. Mesh steps along the z-direction in different layers are different, and the chip is uniformly meshed along the x- and y-directions with different steps. After that, applying the FDM to (1), an equivalent thermal resistance network can be built. An interior control volume is not adjacent to any boundary surfaces, and a boundary control volume is adjacent to at least one boundary surface. Fig. 1.(c) shows the thermal model of an interior control volume, $V_{m,n,l}$. m, n and l are indices of $V_{m,n,l}$ in x-, y- and z-directions, respectively. Using the modeling technique in [2], each thermal conductance can be calculated as

$$g^{x^{-1/2}}_{m,n,l} = \frac{\kappa^{x^{-1/2}}_{m,n,l}\Delta A_{yz}}{\frac{1}{2}\Delta x_{m,n,l}},\ g^{y^{-1/2}}_{m,n,l} = \frac{\kappa^{y^{-1/2}}_{m,n,l}\Delta A_{xz}}{\frac{1}{2}\Delta y_{m,n,l}},\ g^{z^{-1/2}}_{m,n,l} = \frac{\kappa^{z^{-1/2}}_{m,n,l}\Delta A_{xy}}{\frac{1}{2}\Delta z_{m,n,l}},$$

$$g^{x^{+1/2}}_{m,n,l} = \frac{\kappa^{x^{+1/2}}_{m,n,l}\Delta A_{yz}}{\frac{1}{2}\Delta x_{m,n,l}},\ g^{y^{+1/2}}_{m,n,l} = \frac{\kappa^{y^{+1/2}}_{m,n,l}\Delta A_{xz}}{\frac{1}{2}\Delta y_{m,n,l}},\ g^{z^{+1/2}}_{m,n,l} = \frac{\kappa^{z^{+1/2}}_{m,n,l}\Delta A_{xy}}{\frac{1}{2}\Delta z_{m,n,l}}.$$

Here, $\Delta x_{m,n,l}$, $\Delta y_{m,n,l}$ and $\Delta z_{m,n,l}$ are mesh steps of $V_{m,n,l}$ in x-, y- and z-directions, respectively. ΔA_{xy}=$\Delta x_{m,n,l}\Delta y_{m,n,l}$, ΔA_{yz}=$\Delta y_{m,n,l}\Delta z_{m,n,l}$ and ΔA_{xz}=$\Delta x_{m,n,l}\Delta z_{m,n,l}$. $\kappa^{x^{-1/2}}_{m,n,l}$, $\kappa^{x^{+1/2}}_{m,n,l}$, $\kappa^{y^{-1/2}}_{m,n,l}$, $\kappa^{y^{+1/2}}_{m,n,l}$, $\kappa^{z^{-1/2}}_{m,n,l}$ and $\kappa^{z^{+1/2}}_{m,n,l}$ are weighted average thermal conductivities for heterogeneous materials from the central of $V_{m,n,l}$ to its relative borders, respectively. $p_{m,n,l}$ is the power of heat sources in $V_{m,n,l}$. By applying the continuity of heat flow on control-volume borders [10], we have

$$\begin{aligned}-g^{x^{-1}}_{m,n,l}T_{m-1,n,l} - g^{x^{+1}}_{m,n,l}T_{m+1,n,l} - g^{y^{-1}}_{m,n,l}T_{m,n-1,l} - g^{y^{+1}}_{m,n,l}T_{m,n+1,l}\\ -g^{z^{-1}}_{m,n,l}T_{m,n,l-1} - g^{z^{+1}}_{m,n,l}T_{m,n,l+1} + g_{m,n,l}T_{m,n,l} = p_{m,n,l}.\end{aligned}$$

Here, $T_{m-1,n,l}$, $T_{m+1,n,l}$, $T_{m,n-1,l}$, $T_{m,n+1,l}$, $T_{m,n,l-1}$ and $T_{m,n,l+1}$ are temperatures at central positions of control volumes adjacent to $V_{m,n,l}$, respectively. $g_{m,n,l}$=$g^{x^{-1}}_{m,n,l}$+$g^{x^{+1}}_{m,n,l}$ +$g^{y^{-1}}_{m,n,l}$+$g^{y^{+1}}_{m,n,l}$ +$g^{z^{-1}}_{m,n,l}$+$g^{z^{+1}}_{m,n,l}$, and each conductance can be calculated as

$$g^{x^{-1}}_{m,n,l} = g^{x^{-1/2}}_{m,n,l}||g^{x^{+1/2}}_{m-1,n,l},\ g^{x^{+1}}_{m,n,l} = g^{x^{+1/2}}_{m,n,l}||g^{x^{-1/2}}_{m+1,n,l},\ g^{y^{-1}}_{m,n,l} = g^{y^{-1/2}}_{m,n,l}||g^{y^{+1/2}}_{m,n-1,l},$$

$$g^{y^{+1}}_{m,n,l} = g^{y^{+1/2}}_{m,n,l}||g^{y^{-1/2}}_{m,n+1,l},\ g^{z^{-1}}_{m,n,l} = g^{z^{-1/2}}_{m,n,l}||g^{z^{+1/2}}_{m,n,l-1},\ g^{z^{+1}}_{m,n,l} = g^{z^{+1/2}}_{m,n,l}||g^{z^{-1/2}}_{m,n,l+1}.$$

where '||' is the operator for calculating the equivalent conductance of two conductances connected in series.

Equations for boundary control volumes are similar with interior control volumes. Detail models can be referred to [2].

Mapping the control volumes from the triple-index system ($V_{m,n,l}$) into a single-index system (V_i) and stamping related conductances to the system matrix, the temperature distribution can be obtained by solving the following MNA system.

$$\mathbf{G}\mathbf{T} = \mathbf{p}. \tag{3}$$

Here, $\mathbf{T}$ is the temperature vector, and $\mathbf{T}[i]$ is the temperature of V_i. $\mathbf{G}$ is the thermal conductance matrix. $\mathbf{p}$ is the power vector for representing equivalent power values of the control volumes, and $\mathbf{p}[i]$ is the equivalent power of V_i.

B. GIT: Generalized Integral Transform based Thermal Simulator

Given a homogeneous-materials 3-D IC[1], GIT first builds the closed-form expression of lateral bases $\varphi_{i_b}(x, y)$'s in x- and y-directions. Using these lateral bases, GIT constructs several one-dimensional (1-D) thermal problems for calculating the projected function $T^z_{i_b}(z)$ along the z-direction corresponding to each $\varphi_{i_b}(x, y)$. Using the FDM modeling technique, each 1-D thermal problem has a tri-diagonal thermal conductance matrix. Hence, each $T^z_{i_b}(z_l)$, which is the value of $T^z_{i_b}(z)$ at the center position z_l in the z-axis of l-th discretization along the z-direction, can be solved by the linear-time Thomas algorithm. Finally, with representing the steady-state temperature distribution of the homogeneous-materials 3-D IC, $T_h(x, y, z)$, as $\sum_{i_b} T^z_{i_b}(z)\varphi_{i_b}(x, y)$, 2D-STL and 2D-LTS FFT algorithms are utilized to evaluate the temperature distribution of $T(x, y, z_l)$ at each z_l. The computational cost of GIT is O($\mathcal{MNL}\log_2 \mathcal{B}_x\mathcal{B}_y$). $\mathcal{M}$, $\mathcal{N}$ and $\mathcal{L}$ are numbers of mesh steps in x-, y- and z-directions, respectively. $\mathcal{B}_x$ and $\mathcal{B}_y$ are numbers of the lateral bases in x- and y-directions, respectively.

[1] Each layer of a "homogeneous-materials 3-D IC" has an effective thermal conductivity, and effective thermal conductivities in different layers can be different.

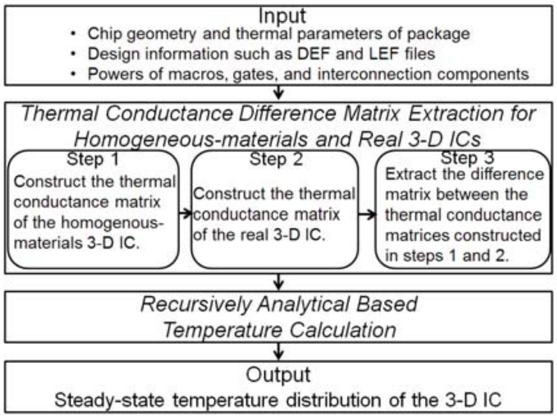

Input
• Chip geometry and thermal parameters of package
• Design information such as DEF and LEF files
• Powers of macros, gates, and interconnection components
Thermal Conductance Difference Matrix Extraction for Homogeneous-materials and Real 3-D ICs
Step 1
Construct the thermal conductance matrix of the homogenous-materials 3-D IC.
Step 2
Construct the thermal conductance matrix of the real 3-D IC.
Step 3
Extract the difference matrix between the thermal conductance matrices constructed in steps 1 and 2.
Recursively Analytical Based Temperature Calculation
Output
Steady-state temperature distribution of the 3-D IC

Fig. 2. Execution flow of the basic kernel of NUMANA.

IV. NUMANA

Instead of using submicrometer-scale resolution for metal wires, compact thermal models facilitate an efficient full-chip thermal analysis by appropriately modeling effective thermal conductivities for inhomogeneous materials. Though adopting compact thermal models does not give the exact result, it still provides the average on-chip thermal behavior.

Inspired by the above concept, the basic idea of NUMANA is that: Since compact thermal models can provide the average thermal behavior of chip, they can be used to supply a reasonable nominal temperature distribution. Then, NUMANA finds a mechanism to correct temperature variations between compact and FDM based thermal models for a real 3-D IC. The basic kernel of NUMANA is detailed in section IV-A and is enhanced by using the multi-point expansion technique in section IV-B.

A. Basic Kernel

The execution flow of the basic kernel of NUMANA is shown in Fig. 2. According to the chip geometry, thermal parameters of package, design information such as TSVs/microbumps included DEF and LEF files, and power consumption of macros/gates and interconnection components of a given real 3-D IC, NUMANA first chooses appropriate homogeneous thermal conductivities to build a homogeneous-materials 3-D IC and constructs its spatial bases for representing the temperature distribution. After that, using the FDM based modeling technique, NUMANA constructs thermal conductance matrices for the real 3-D IC and the homogeneous-materials 3-D IC. Then, NUMANA extracts their difference matrix by subtracting these two thermal conductance matrices. Finally, using GIT as the simulation kernel, NUMANA performs a recursive analytical-based calculation technique for the steady-state temperature distribution of the real 3-D IC.

A.1. Calculation Formula Construction

Using the FDM based modeling technique, the MNA system shown in (3) of a real 3-D IC can be built. $\mathbf{G}$ can be decomposed as $\mathbf{G}_{\mathrm{h}}+\Delta\mathbf{G}$. Here, $\mathbf{G}_{\mathrm{h}}$ is the thermal conductance matrix of a homogeneous-materials 3-D IC having the same stacked-layer structure as the real 3-D IC and having appropriately chosen thermal conductivities in different layers. Substituting $\mathbf{G} = \mathbf{G}_{\mathrm{h}} + \Delta\mathbf{G}$ into (3), the temperature distribution of a real 3-D IC can be calculated as

$$\mathbf{T} = (\mathbf{G}_{\mathrm{h}} + \Delta\mathbf{G})^{-1}\mathbf{p} = (\mathbf{I} + \mathbf{G}_{\mathrm{h}}^{-1}\Delta\mathbf{G})^{-1}\mathbf{G}_{\mathrm{h}}^{-1}\mathbf{p}. \tag{4}$$

Two theorems for developing NUMANA are stated as follows. Due to lack of space their proofs are omitted.

Theorem 1: Decomposing $\mathbf{G}$ as $\mathbf{G}_{\mathrm{h}} + \Delta\mathbf{G}$, $\mathbf{T}$ can be expressed as

$$\mathbf{T} = \sum_{q=0}^{\infty}(-1)^q \left(\mathbf{G}_{\mathrm{h}}^{-1}\Delta\mathbf{G}\right)^q \mathbf{G}_{\mathrm{h}}^{-1}\mathbf{p}, \tag{5}$$

where $\mathbf{G}_{\mathrm{h}}$ is constructed by setting the suitable thermal conductivity, κ_{h_l}, for each $\mathbb{V}_{z_l}$, and each κ_{h_l} satisfies

$$\kappa_{\mathrm{h}_l} > \kappa_{\mathrm{max}_l}/2, \text{ for } 1 \le l \le \mathcal{L}. \tag{6}$$

Here, l and $\mathcal{L}$ are the index and the number of discretization levels in z-direction, respectively. z_l is the central position of z-axis of the l-th discretization level along the z-direction, and $\mathbb{V}_{z_l}$ is the set of all lateral control volumes with their central position in the z-axis being equal to z_l. κ_{h_l} is the selected thermal conductivity in $\mathbb{V}_{z_l}$, and κ_{max_l} is the maximum thermal conductivity among $\mathbb{V}_{z_l}$. □

Theorem 2: The convergent rate of (5) is dominated by the maximum absolute eigenvalue of $\mathbf{G}_{\mathrm{h}}^{-1}\Delta\mathbf{G}$, $\sigma_{\mathrm{max}}\left(\mathbf{G}_{\mathrm{h}}^{-1}\Delta\mathbf{G}\right)$, and $\sigma_{\mathrm{max}}\left(\mathbf{G}_{\mathrm{h}}^{-1}\Delta\mathbf{G}\right)$ is bounded by

$$\sigma_{\mathrm{max}}\left(\mathbf{G}_{\mathrm{h}}^{-1}\Delta\mathbf{G}\right) \le \sigma_{\mathrm{max}}\left(\overline{\mathbf{G}_{\mathrm{h}}}^{-1}\right) \max_{i=1 \text{ to } K}\left(\max\left(\alpha\left[i\right], \beta\left[i\right]\right)\right), \tag{7}$$

where $K = \mathcal{MNL}$, $\overline{\mathbf{G}_{\mathrm{h}}}$ is the normalized matrix with its ij-th entry being $\mathbf{G}_{\mathrm{h}}[i][j]/\mathbf{G}_{\mathrm{h}}[i][i]$, and

$$\alpha[i] = \frac{\sum_{k\neq i}^{K}\left(|\Delta\mathbf{G}[i][k]| - \Delta\mathbf{G}[i][k]\right)}{\sum_{j\neq i}^{K}\mathbf{G}_{\mathrm{h}}[i][j]},$$

$$\beta\left[i\right] = \frac{\sum_{k\neq i}^{K}\left(|\Delta\mathbf{G}[i][k]| + \Delta\mathbf{G}[i][k]\right)}{\sum_{j\neq i}^{K}\mathbf{G}_{\mathrm{h}}[i][j]}.$$

□

From *Theorem 1*, the Q-th order approximation of $\mathbf{T}$ is

$$\mathbf{T} \approx \mathbf{T}^Q \overset{\mathrm{def}}{=} \mathbf{m}_0 + \sum_{q=1}^{Q}\mathbf{m}_q, \tag{8}$$

where $\mathbf{m}_0 = \mathbf{G}_{\mathrm{h}}^{-1}\mathbf{p}$, $\mathbf{m}_q = \mathbf{G}_{\mathrm{h}}^{-1}\mathbf{b}_{q-1}$, and $\mathbf{b}_{q-1} = -\Delta\mathbf{G}\mathbf{m}_{q-1}$. Here,each $\mathbf{m}_q$ can be calculated by using the same $\mathbf{G}_{\mathrm{h}}^{-1}$, and $\mathbf{T}^Q$ can be computed recursively. To efficiently compute $\mathbf{T}^Q$, NUMANA develops a recursive analytical-based calculation technique for solving each $\mathbf{m}_q$.

A.2. Recursive Analytical-based Temperature Calculation

The key idea of the recursive analytical-based temperature calculation is conducted by the following observation. Since $\mathbf{G}_{\mathrm{h}}$ is a thermal conductance matrix of a homogeneous-materials 3-D IC, $\mathbf{m}_0$ and $\mathbf{m}_q$'s can be viewed as the temperature vectors for a homogeneous-materials 3-D IC by treating entries of $\mathbf{p}$ and $\mathbf{b}_{q-1}$'s as heat sources in the corresponding control volumes. Since GIT supports such physical model and can analytically solve the solution, it is suitable for calculating $\mathbf{m}_0$ and $\mathbf{m}_q$'s.

```
Procedure  T = NUMANA-SP ($\mathbf{p}$, $\boldsymbol{\kappa}$, $\vec{\kappa}_{\mathrm{h}}$, $\varepsilon$)
Input: $\mathbf{p}$: power distribution vector,
       $\boldsymbol{\kappa}$: thermal conductivity set of a real 3-D IC,
       $\vec{\kappa}_{\mathrm{h}}$: selected homogeneous thermal conductivity vector,
       $\varepsilon$: threshold.
Output: $\mathbf{T}$: Estimated temperature distribution.
 1  Begin
 2    Construct spatial bases and 1-D MNA system of GIT($\cdot$) by using $\vec{\kappa}_{\mathrm{h}}$;
 3    Construct $\Delta\mathbf{G}$ by using $\vec{\kappa}$ and $\vec{\kappa}_{\mathrm{h}}$;
 4    $\mathbf{m}_0 \leftarrow$ GIT($\mathbf{p}$);
 5    $\mathbf{T}^0 \leftarrow \mathbf{m}_0$;
 6    $q \leftarrow 1$;
 7    While $\max\left\{\left|\mathbf{m}_{q-1}[i]/\mathbf{T}^{q-1}[i]\right|\right\} > \varepsilon$
 8      $\mathbf{b}_{q-1} \leftarrow -\Delta\mathbf{G}\mathbf{m}_{q-1}$;
 9      $\mathbf{m}_q \leftarrow$ GIT($\mathbf{b}_{q-1}$);
10      $\mathbf{T}^q \leftarrow \mathbf{T}^{q-1} + \mathbf{m}_q$;
11      $q \leftarrow q + 1$;
12    $\mathbf{T} \leftarrow \mathbf{T}^{q-1}$;
13  End
```

Fig. 3. NUMANA-SP: Basic kernel of NUMANA.

The basic kernel of NUMANA is shown in Fig. 3. Since this procedure only utilizes *one* homogeneous-materials, we call it NUMANA-SP (*single-point expansion*). Given $\vec{\kappa}_{\mathrm{h}}$, first, the spatial bases and 1-D MNA system of GIT($\cdot$) are constructed. Here, $\vec{\kappa}_{\mathrm{h}} = [\kappa_{\mathrm{h}_1}, \cdots, \kappa_{\mathrm{h}_{\mathcal{L}}}]^T$. $\Delta\mathbf{G}$ is constructed by utilizing $\vec{\kappa}_{\mathrm{h}}$ and $\boldsymbol{\kappa}$ (the thermal conductivity set of control volumes of the real 3-D IC). After that, the recursive analytical-based temperature calculation stated in *Lines* 4~11 repeatedly computes $\mathbf{m}_0$ and each $\mathbf{m}_q$ until each absolute ratio of the i-th entry of $\mathbf{m}_{q-1}$ and $\mathbf{T}^{q-1}$, $\max\{\left|\mathbf{m}_{q-1}[i]/\mathbf{T}^{q-1}[i]\right|\}$, is less than the threshold ε. According to section III-B, the complexity for solving $\mathbf{T}^Q$ is $\mathrm{O}\left((Q+1)\mathcal{MNL}\log_2 \mathcal{B}_x\mathcal{B}_y\right)$.

Choosing appropriate values of $\vec{\kappa}_{\mathrm{h}}$ for construing $\Delta\mathbf{G}$ is important, because it leads to the different convergent properties of (8). As indicated by *Theorem 2*, an appropriate selection of $\vec{\kappa}_{\mathrm{h}}$ should try to minimize absolute values of entries in $\Delta\mathbf{G}$. Notice that large differences between thermal conductivities in the real 3-D IC and $\vec{\kappa}_{\mathrm{h}}$ will result in large absolute values of entries in $\Delta\mathbf{G}$. Therefore, the minimization of absolute values of entries in $\Delta\mathbf{G}$ can be adequately approximated by minimizing the differences between thermal conductivities in control volumes and the corresponding entries in $\vec{\kappa}_{\mathrm{h}}$. To ensure a fast convergent rate for a large portion of temperatures, the minimization process should take into account the statistic information of thermal conductivities in control volumes. With the above discussions, each κ_{h_l} can be appropriately selected by solving the convex optimization problem as

$$\text{minimize} \qquad \sum_{g=1}^{g=\mathcal{N}_g^l} \mathrm{Prob}_g \left(\kappa_g^l - \kappa_{\mathrm{h}_l}\right)^2, \tag{9}$$

$$\text{subject to} \qquad \kappa_{\max_l}/2 - \kappa_{\mathrm{h}_l} \leq 0. \tag{10}$$

Here, κ_g^l's are the effective thermal conductivities for constructing conductances of control volumes in $\mathbb{V}_{z_l}$. $\mathcal{N}_g^l$ is the number of different conductances of control volumes in $\mathbb{V}_{z_l}$. Prob_g is the occurrence probability of κ_g^l.

This optimization problem can be solved as

$$\kappa_{\mathrm{h}_l} = \max\left(\sum_{g=1}^{g=\mathcal{N}_g^l} \mathrm{Prob}_g \kappa_g^l, \left(\frac{\kappa_{\max_l}}{2}\right)^+\right) \tag{11}$$

Here, $(\kappa_{\max_l}/2)^+$ is a value that is slightly larger than $\kappa_{\max_l}/2$.

With (11), the temperatures in the large portion of control volumes can be well approximated by a low truncation order. However, the control volumes with large differences of thermal conductivities have a low convergent rate, and the maximum error usually occurs in those control volumes. The multi-point expansion technique will be adopted for amending this issue.

```
Procedure  T = NUMANA-MP($\mathbf{p}, \boldsymbol{\kappa}, \mathbf{N}, \varepsilon, \varepsilon'$)
Input: $\mathbf{p}$: power distribution vector,
       $\boldsymbol{\kappa}$: thermal conductivity set of a real 3-D IC,
       $N$: maximum expanding point number,
       $\varepsilon$, $\varepsilon'$: threshold.
Output: $\mathbf{T}$: Estimated temperature distribution.
Begin
  Compile $\mathbf{CDF}_l(\kappa)$ of thermal conductivity for each layer;
  Select expanding point as $\vec{\kappa}_{\mathrm{h}}^{(1)}$;
  $[\mathbf{T}, \mathbf{R}]$ = RGIT($\mathbf{p}, \boldsymbol{\kappa}, \vec{\kappa}_{\mathrm{h}}^{(1)}, \varepsilon'$);
  $\varepsilon' \leftarrow c \cdot \varepsilon'$;
  $n \leftarrow 1$;
  While ($n \leq N$ & !Conv)
    Select expanding point as $\vec{\kappa}_{\mathrm{h}}^{(n)}$ ;
    $\left[\mathbf{T}^{(n)}, \mathbf{R}^{(n)}\right]$ = RGIT($\mathbf{p}, \boldsymbol{\kappa}, \vec{\kappa}_{\mathrm{h}}^{(n)}, \varepsilon'$);
    $[\mathbf{T}, \mathbf{R}, \mathbf{Conv}] \leftarrow$ Point-wise best result extractor ($\mathbf{T}, \mathbf{T}^{(n)}, \mathbf{R}, \mathbf{R}^{(n)}, \varepsilon$);
    $n \leftarrow n + 1$;
End
```

Fig. 4. Procedure of NUMANA-MP.

```
Procedure  $[\mathbf{T}, \mathbf{R}]$ = RGIT($\mathbf{p}$, $\boldsymbol{\kappa}$, $\vec{\kappa}_{\mathrm{h}}$, $\varepsilon'$)
Input: $\mathbf{p}$: power distribution vector,
       $\boldsymbol{\kappa}$: thermal conductivity set of a real 3-D IC,
       $\vec{\kappa}_{\mathrm{h}}$: selected homogeneous thermal conductivity vector,
       $\varepsilon'$: threshold.
Output: $\mathbf{T}$: Estimated temperature distribution,
        $\mathbf{R}$: Estimated point-wise ratio of redundancy.
Begin
  Construct spatial bases and 1-D MNA system of GIT($\cdot$) by using $\vec{\kappa}_{\mathrm{h}}$;
  Construct $\Delta\mathbf{G}$ by using $\vec{\kappa}_{\mathrm{h}}$ and $\boldsymbol{\kappa}$;
  $\mathbf{m}_0 \leftarrow$ GIT($\mathbf{p}$);
  $\mathbf{T}^0 \leftarrow \mathbf{m}_0$;
  $q \leftarrow 1$;
  While $\|\mathbf{m}_{q-1}\|_2/\|\mathbf{T}^{q-1}\|_2 > \varepsilon'$
    $\mathbf{b}_{q-1} \leftarrow -\Delta\mathbf{G}\mathbf{m}_{q-1}$;
    $\mathbf{m}_q \leftarrow$ GIT($\mathbf{b}_{q-1}$);
    $\mathbf{T}^q \leftarrow \mathbf{T}^{q-1} + \mathbf{m}_q$;
    $q \leftarrow q + 1$;
  For all $i$
    $\mathbf{R}[i] \leftarrow \left|\mathbf{m}_{q-1}[i]/\mathbf{T}^{q-1}[i]\right|$;
  $\mathbf{T} \leftarrow \mathbf{T}^{q-1}$;
End
```

Fig. 5. Procedure of RGIT.

B. Multi-Point Expansion (NUMANA-MP)

As variations of thermal conductivities in a layer are quite large, NUMANA-SP might need dozens of truncation orders to approximate the temperature profile because some entries of $\Delta\mathbf{G}$ and $\mathbf{b}_q$'s would be large. To reduce truncation orders, we can appropriately build several different homogeneous-materials 3-D ICs (i.e. *multi-point expansion*) to ease the effects of large thermal conductivity variations. We name it NUMANA-MP (*multi-point expansion*).

The procedure of NUMANA-MP is shown in Fig. 4. NUMANA-MP first estimates the cumulative distribution function $\mathbf{CDF}_l(\kappa)$ of thermal conductivities for each $\mathbb{V}_{z_l}$, and a suitable expanding-point vector $\vec{\kappa}_{\mathrm{h}}^{(1)}$ is chosen by using (11). Next, a recursive analytical temperature calculation procedure (RGIT) shown in Fig. 5 is executed. RGIT is similar to NUMANA-SP except that RGIT records the redundancy ratio $\mathbf{R}[i] = \left|\mathbf{m}_{q-1}[i]/\mathbf{T}^{q-1}[i]\right|$ for each control volume i and computes $\left\|\mathbf{m}_{q-1}\right\|_2 / \left\|\mathbf{T}^{q-1}\right\|_2$. If $\left\|\mathbf{m}_{q-1}\right\|_2 / \left\|\mathbf{T}^{q-1}\right\|_2$ is equal or less than a threshold ε', it means that the truncation order is large enough for the most control volumes.

After obtaining the estimated temperature profile and redundancy ratios, ε' is loosened by $c \cdot \varepsilon'$ for the rest expanding-point vectors because NUMANA-MP tries to avoid spending too much time on those convergent control volumes by the first expanding-point vector. Then, NUMANA-MP processes the rest expanding-point vectors. To ensure efficient simulations, a chosen expanding-point vector must differ from those expanding-point vectors that we have picked. NUMANA-MP sets that the difference between the thermal conductivities of different expanding-point vectors in $\mathbb{V}_{z_l}$ must be larger than $\alpha \cdot \kappa_{max_l}$ (α<1). As each $\vec{\kappa}_{\text{h}}^{(n)}$ is well picked (i.e., a new homogeneous-materials 3-D IC is built.), NUMANA-MP executes RGIT and *the point-wise best result extractor* to output the temperature profile.

The point-wise best result extractor compares the values of $\mathbf{R}[i]$ got by different expanding-point vectors with $\mathbf{R}^{(n)}[i]$, and replaces $\mathbf{T}[i]$ with $\mathbf{T}^{(n)}[i]$ if $\mathbf{R}^{(n)}[i]$ is less than $\mathbf{R}[i]$. As all control volumes are compared, it checks whether any existing $\mathbf{R}[i]$'s are larger than ε. $\mathbf{T}$ is converged if no one is larger than ε. Otherwise, NUMANA-MP selects another expanding-point vector and continues the simulation procedure.

V. Experimental Results

MUMANA is implemented in C++ language and tested on Intel Xeon Processor E5620 2.4-GHz CPU with 96GB RAM. Thermal conductivities (W/m·K) of silicon, copper, microbump and oxide are 148, 406, 60 and 0.83 under 27°C, respectively. The physical model of stacked-layer 3-D IC is shown in Fig. 1.(a). The material of TSV is copper. The routing density of each interconnect layer is 0.4~0.6, and the diameter of micro solder bump is 30μm. The heat transfer coefficients of primary and secondary heat flow path, $h_{b_s^p}(\vec{r}_{b_s^p})$ and $h_{b_s^s}(\vec{r}_{b_s^s})$ are 42829.81W/m^2·K and 2000W/m^2·K. The number of lateral spatial bases is 32×32 for executing GIT(·). c is 100, and α is 0.05 for NUMANA-MP. Maximum relative error $e_{\max}$ and average relative error e_{avg} are used as measures of difference between estimated and exact thermal profiles.

$$e_{\max} = \max_{\forall v \in \mathbb{V}} \left(\left| \left(\mathbf{T}_v^{\text{ref}} - \mathbf{T}_v^{\text{est}} \right) / \mathbf{T}_v^{\text{ref}} \right| \right), \tag{12}$$

$$e_{\text{avg}} = \frac{1}{N} \sum_{\forall v \in \mathbb{V}} \left| \left(\mathbf{T}_v^{\text{ref}} - \mathbf{T}_v^{\text{est}} \right) / \mathbf{T}_v^{\text{ref}} \right|, \tag{13}$$

where $\mathbb{V}$ is the control volume set, N is the number of control volumes, $\mathbf{T}_v^{\text{ref}}$ (°C) is the temperature of v-th control volume obtained by a reference solver, and $\mathbf{T}_v^{\text{est}}$ (°C) is the temperature of v-th control volume obtained by a estimated method.

A. Accuracy Verification

To verify the accuracy of NUMANA, NUMANA-SP is used to simulate a three stacked-layer 3-D IC, and its result is compared with that of ANSYS. The diameter of TSV is 45μm, and 200 TSVs are distributed between each two stacked layers. To execute ANSYS, first, the geometry of 3-D IC is built to fit the form of ANSYS. After building the thermal model, ANSYS spatially discretizes the model. As the mesh procedure is done, boundary conditions are set, heat sources are inserted, and the temperature distribution is calculated by ANSYS.

The test chip is discretized to 32, 32, 14 steps along x-, y- and z-directions, respectively, for NUMANA-SP. Hence, the control-volume number is $32 \times 32 \times 14$. In aspect of accuracy, the temperature profile got by NUMANA, plotted in Fig. 6.(b), conforms to the result got by ANSYS, shown in Fig. 6.(a). Its maximum relative error is 1.84%, and its average relative error is 0.54%. In aspect of simulation speed, ANSYS spent half a day to estimate the temperature profile but NUMANA only took 0.15 second to estimate the temperature profile.

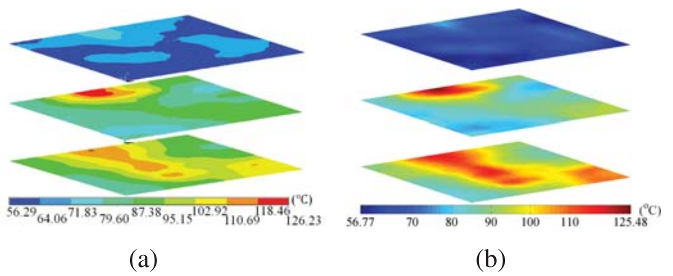


Fig. 6. Temperature profiles of a three stacked-layer chip by (a) ANSYS and (b) NUMANA-SP.

B. Robustness and Efficiency Demonstration

SuperLU (version V4.3) [1] is a well known matrix solving tool. It provides an efficient method (we name it SuperLU-Direct) to decompose symmetric positive definite matrix and directly solve matrix equations. In addition, it also provides a generalized minimal residual method (GMRES) solver (we name it SuperLU-GMRES) with an incomplete LU factorization pre-conditioner [1], which is faster than directly solving the problem. The stopping criterion of SuperLU-GMRES is set to achieve the same accuracy level as NUMANA. To demonstrate efficiency and robustness of NUMANA, NUMANA is compared with SuperLU-Direct and SuperLU-GMRES. ε and ε' for NUMANA are 10^{-6} and 10^{-4}, respectively.

B.1. Comparison between NUMANA and SuperLU with Different Amounts of TSVs

With different amounts of TSVs, the comparison between NUMANA and SuperLU is shown in TABLE I. The reference solver is SuperLU-Direct. Each test chip has five stacked layers with different amounts of TSVs. The diameter of TSV is 6μm. TSVs are placed randomly, and micro solder bumps are placed uniformly. The control-volume number is 256×256×24. "Cell Cnt." is the number of cells, "TSV Cnt." is the number of TSVs, and "Power" is the chip power consumption.

From TABLE I, the maximum relative error of NUMANA is less than 1.82%, and its average relative error is less than 0.97%. Comparing with SuperLU-Direct and SuperLU-GMRES, NUMANA-SP is at least 856× speedup and 18× speedup, respectively. NUMANA-MP uses two expansion-point vectors, and their truncation orders are shown in the 8-th column. As shown in the 7-th and 8-th columns, the truncation number of NUMANA-MP is less than NUMANA-SP. Hence, NUMANA-MP further saves the runtime and can achieve at least 1162.8× speedup and 25.1× speedup over SuperLU-Direct and SuperLU-GMRES, respectively.

TABLE I also shows that as the number of TSVs increases, the relative errors and truncation orders of NUMANA decrease. The reason is as follows. Since the portions of the

TABLE I

Comparison between NUMANA and SuperLU [1] with different amounts of TSVs.

Test Chip	Tier Cnt.	Cell Cnt.	TSV Cnt.	Power	Truncation order/ #iteration			Maximum Relative Error (%)			Average Relative Error (%)			†Runtime (s)				NUMANA Speedup Ratio			
		($\times10^6$)	($\times10^3$)	(W)	SuperLU	NUMANA		SuperLU	NUMANA		SuperLU	NUMANA		SuperLU		NUMANA		over -Direct		over -GMRES	
					-GMRES	SP	MP	-GMRES	SP	MP	-GMRES	SP	MP	-Direct	-GMRES	SP	MP	SP	MP	SP	MP
chip 1	5	16.0	3.4	68.49	339	40	23,5	1.9094	1.7991	1.8142	1.5161	0.9472	0.9603	9860.86	212.84	11.52	8.48	856.0	1162.8	18.48	25.10
chip 2	5	15.5	6.8	66.37	294	32	19,4	1.6000	1.5590	1.5727	1.3017	0.7539	0.7626	9889.07	190.70	9.25	7.09	1069.1	1392.8	20.62	26.90
chip 3	5	15.0	9.8	64.37	298	24	14,4	1.2480	1.1556	1.1657	1.0003	0.5446	0.5505	9875.09	198.50	6.97	5.68	1416.8	1738.6	28.48	34.95
chip 4	5	14.5	13.3	62.21	325	17	11,3	0.9149	0.9114	0.9142	0.7017	0.3769	0.3784	9928.32	204.13	4.99	4.56	1989.6	2177.3	40.91	44.77

† The runtime does not include the execution time for parsing files.

TABLE II

Comparison between NUMANA and SuperLU [1] with different amounts of control volumes.

Test Chip	Control Volume Cnt.	Truncation Order/ #iteration			Maximum Relative Error (%)			Average Relative Error (%)			†Runtime (s)				NUMANA Speedup Ratio			
		SuperLU	NUMANA		SuperLU	NUMANA		SuperLU	NUMANA		SuperLU		NUMANA		over -Direct		over -GMRES	
		-GMRES	SP	MP	-GMRES	SP	MP	-GMRES	SP	MP	-Direct	-GMRES	SP	MP	SP	MP	SP	MP
Chip 1	98.3K	219	8	4,4	0.6013	0.5499	0.5499	0.4830	0.1324	0.1324	62.44	7.89	0.23	0.29	271.5	215.3	34.30	27.21
	393.2K	286	24	15,4	1.4062	1.3612	1.3654	1.0561	0.6687	0.6726	842.93	44.36	2.00	1.69	421.5	498.2	22.18	26.25
	1.6M	339	40	23,5	1.9094	1.7991	1.8142	1.5161	0.9472	0.9603	9860.86	212.84	11.52	8.48	856.0	1162.8	18.48	25.10
	♣6.3M	NA	41	24,4	NA	2.0109	2.0301	NA	1.0480	1.0608	NA	NA	46.10	33.16	NA	NA	NA	NA
	♣25.2M	NA	42	24,4	NA	2.1111	2.1333	NA	1.0791	1.0938	NA	NA	191.57	135.02	NA	NA	NA	NA
Chip 4	98.3K	199	5	3,3	0.6753	0.6529	0.6529	0.4771	0.1803	0.1803	62.58	7.44	0.15	0.24	417.2	260.8	49.60	31.00
	393.2K	330	8	6,3	0.7095	0.6701	0.6703	0.5700	0.2614	0.2728	842.36	49.53	0.70	0.89	1203.4	946.5	70.76	55.65
	1.6M	325	17	11,3	0.9149	0.9114	0.9142	0.7017	0.3769	0.3784	9928.32	204.13	4.99	4.56	1989.6	2177.3	40.91	44.77
	♣6.3M	NA	18	11,3	NA	1.0680	1.0721	NA	0.4172	0.4196	NA	NA	20.63	17.94	NA	NA	NA	NA
	♣25.2M	NA	18	11,3	NA	1.1463	1.1513	NA	0.4313	0.4342	NA	NA	83.85	73.29	NA	NA	NA	NA

† The runtime does not include the execution time for parsing files.
♣ The reference solution is the result of SuperLU-GMRES.

left-half side of thermal conductivity distribution of silicon substrate and micro solder bump layers are decreased as the number of TSVs increases, the number of non-zero entries of $\Delta\mathbf{G}$ is fewer, and their values would be smaller.

B.2. Comparison between NUMANA and SuperLU with Different Amounts of Control Volumes

With different amounts of control volumes, temperature profiles of chip 1 and chip 4 are calculated by NUMANA and SuperLU. The control-volume number is varied from 98.3K ($64 \times 64 \times 24$) to 25.2M ($1024 \times 1024 \times 24$), and TABLE II shows the comparison. The reference solver is SuperLU-Direct. However, the reference solution is obtained by executing SuperLU-GMRES with $10,000$ iterations for the cases of 6.3M and 25.2M control volumes since SuperLU-Direct runs out of memory.

The maximum relative error of NUMANA is less than 2.1%, and its average relative error is less than 1.1%. The 4-th and 5-th columns of TABLE II reveal that the truncation order is raised as the control-volume number increases. Since more control volumes in a chip lead to more expressions of composed materials, the thermal conductivity distribution of each layer varies more wildly. The 3-rd column shows that the iteration number of SuperLU-GMRES also increases as the control-volume number increases.

The 14-th and 15-th columns show the runtime of NUMANA-SP and NUMANA-MP. Since NUMANA-MP uses multiple-point vectors to expand the system, it is more robust than NUMANA-SP. Hence, as the control-volume number increases, the runtime of NUMANA-MP is less than that of NUMANA-SP. As shown in the 16-th and 17-th columns of TABLE II, the speedup ratio of NUMANA over SuperLU-Direct grows as the control-volume number increases. Finally, the 18-th and 19-th columns show that NUMANA can be an order of speedup over SuperLU-GMRES.

VI. Conclusion

NUMANA, an efficient thermal simulator for 3-D ICs, has been developed. By combing both advantages of analytical and numerical techniques, NUMANA can effectively estimate the temperature profile of a 3-D IC even with complicated material structures.

References

[1] SuperLU, `http://crd-legacy.lbl.gov/~xiaoye/SuperLU/`.

[2] T. Y. Wang and C. C. P. Chen. 3-D thermal-ADI: A linear-time chip level transient thermal simulator. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 21(12):1434–1245, 2002.

[3] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan. HotSpot: A compact thermal modeling methodology for early-stage VLSI design. *IEEE Transactions on Very Large Scale Integration Systems*, 14(5):501–513, 2006.

[4] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra. IC thermal simulation and modeling via efficient multigrid-based approaches. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(9):319–326, 2006.

[5] Y. Yang, Z. Gu, C. Zhu, R. P. Dick, and L. Shang. ISAC: Integrated space-and-time-adaptive chip-package thermal analysis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(1):86–99, 2007.

[6] C. Xu, L. Jiang, S. K. Kolluri, B. J. Rubin, A. Deutsch, H. Smith, and K. Banerjee. Fast 3-D thermal analysis of complex interconnect structures using electrical modeling and simulation methodologies. In *Proceedings of International Conference on Computer Aided Design-Digest of Technical Papers*, pages 658–665, 2009.

[7] Z. Hassan, N. Allec, F. Yang, L. Shang, R. P. Dick, and X. Zeng. Full-spectrum spatial–temporal dynamic thermal analysis for nanometer-scale integrated circuits. *IEEE Transactions on Very Large Scale Integration Systems*, 19:2276–2289, 2011.

[8] Y. Zhan and S. S. Sapatnekar. High-efficiency Green function-based thermal simulation algorithms. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(9):1661–1675, 2007.

[9] P. Y. Huang and Y. M. Lee. Full-chip thermal analysis for the early design stage via generalized integral transforms. *IEEE Transactions on Very Large Scale Integration Systems*, 17(5):613–626, 2009.

[10] A. Kariant, M. Pramono, and Y. Zhan. Method and apparatus for thermal analysis, Jan. 2012. US Patent 8,104,007.