A High Efficiency Full-Chip Thermal Simulation Algorithm

Yong Zhan and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering
University of Minnesota

***Abstract*— Thermal simulation has become increasingly important in chip design especially in the nanometer regime, where the on-chip hot spots severely degrade the performance and reliability of the circuit and increase the leakage power. In this paper, we present a highly efficient and accurate thermal simulation algorithm that is capable of performing full-chip temperature calculations at the cell level. The algorithm is a combination of several important numerical techniques including the Green function method, the discrete cosine transform (DCT), and the frequency domain computations. Experimental results show that our algorithm can achieve orders of magnitude speedup compared with previous Green function based algorithms while maintaining the same accuracy.**

I. Introduction

Thermal simulation algorithms in chip design can be roughly divided into two categories based on whether the meshing of the entire substrate is necessary during the simulation process. The generic thermal simulation algorithms such as the finite difference method (FDM) and the finite element method (FEM) used in [1] and [2] enjoy the advantages of high flexibility in handling different kinds of boundary conditions in thermal problems and the capability of achieving high accuracy. However, the requirement of meshing the entire substrate and later solving a large system of linear equations in the simulations using these methods makes them relatively inefficient. In [3], the thermal-ADI algorithm was proposed to efficiently solve the transient thermal problems using a meshing scheme similar to that used in the FDM. However, for steady-state analysis, the ADI algorithm can also become slow if the initial guess of the temperature distribution is far from the final solution.

The boundary element method (BEM) constitutes another class of thermal simulation algorithms in which the volume meshing of the substrate is completely avoided. An important underlying concept in the BEM is the Green function which describes the temperature distribution in the chip when a unit point power source is present. In [4] and [5], the analytical forms of the Green function were derived by assuming that the chip was infinitely large horizontally. One significant advantage of the analytical forms of the Green function is that they are very cheap to evaluate and hence can be easily incorporated into optimization procedures where the Green function needs to be evaluated many times. However, by assuming that the chip is infinitely large horizontally, the derived Green functions tend to severely underestimate the temperature, although they can correctly identify the locations of the hot spots as shown in [4]. In [6], a Green function that is suitable for the rectangular shaped chip geometry was presented and look-up tables were established based on the Green function to assist the efficient evaluation of the temperature field. Nevertheless, the cost of this algorithm can become prohibitive for cell level full-chip thermal simulations where both the number of heat sources and the number of field regions are large.

In [7], an efficient thermal simulation algorithm based on the solution of the finite difference equations using the multigrid approach was proposed, and it has the capability of performing the full-chip thermal analysis. In this paper, we present another highly efficient and accurate full-chip thermal simulation algorithm that is based on the Green function method, the discrete cosine transform (DCT), and the frequency domain computations. Since the temperature field can be obtained by convolving the power distribution with the underlying Green function, using the frequency domain computations in conjunction with the DCT will result in a significant improvement in efficiency as can be achieved in many signal processing works where the time or space domain convolution is replaced by the frequency domain analysis. The functional eigen-decomposition approach used in the substrate parasitic extraction work in [8] is also a specific implementation of the frequency domain analysis. Our algorithm takes a piece-wise constant power density map as the input and generates a piece-wise constant temperature map as the output. The primary steps of the algorithm include

1) Obtain the frequency domain representation of the power density map using the 2D DCT.
2) Calculate the frequency domain representation of the temperature map by multiplying each frequency component of the power density map by the corresponding frequency response of the linear system determined by the Green function.
3) Use a 2D inverse discrete cosine transform (IDCT) to obtain the temperature map from its frequency domain representation.

Both the 2D DCT and the 2D IDCT can be calculated efficiently using the 2D fast Fourier transform (FFT) in $O((M\cdot N)\times\log(M\cdot N))$ time, where $M\cdot N$ is the total number of grid cells in the power density map, which is also the total number of grid cells in the resulting temperature map. This is a significant improvement over the $O((M\cdot N)^2)$ time complexity of the algorithm presented in [6]. Experimental results show that the algorithm proposed in this paper can achieve orders of magnitude speedup over the algorithm in [6], while still maintaining the same accuracy.

This work was supported in part by DARPA under grant N66001-04-1-8909 and NSF under award CCR-0205227.

The authors thank the University of Minnesota Supercomputing Institute for providing the computing facilities.

II. Problem Formulation

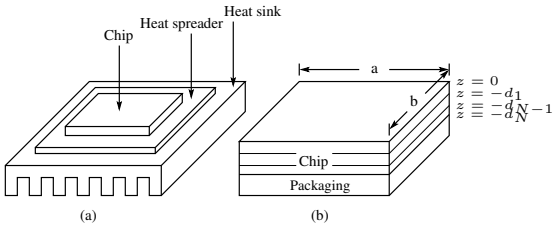


Fig. 1. Schematic of a VLSI chip with packaging (a) IC chip and the packaging structure (b) simplified model of the chip and packaging.

Fig. 1(a) shows an IC chip with the associated packaging, and Fig. 1(b) shows a schematic of the structure in Fig. 1(a) where the packaging including the heat spreader and the heat sink has been simplified but the multilayered structure of the chip is explicitly shown. The steady state temperature distribution inside the chip are governed by the Poisson's equation

$$\nabla^2 T(\mathbf{r}) = -\frac{g(\mathbf{r})}{k_{l(\mathbf{r})}} \tag{1}$$

where $\mathbf{r} = (x, y, z)$, $T(\mathbf{r})$ is the temperature (°C) distribution inside the chip, $g(\mathbf{r})$ is the volume power density (W/m^3), and $k_{l(\mathbf{r})}$ is the thermal conductivity (W/(m·°C)) of the layer where point $\mathbf{r}$ is located [9]. The vertical surfaces and the top surface of the chip are assumed to be adiabatic [10], and the bottom surface of the chip is assumed to be convective, with an effective heat transfer coefficient h (W/(m^2·°C)) [11]. In mathematical forms, these boundary conditions can be expressed as

$$\left.\frac{\partial T(\mathbf{r})}{\partial x}\right|_{x=0,a} = \left.\frac{\partial T(\mathbf{r})}{\partial y}\right|_{y=0,b} = 0 \tag{2}$$

$$\left.\frac{\partial T(\mathbf{r})}{\partial z}\right|_{z=0} = 0 \tag{3}$$

$$k_N \left.\frac{\partial T(\mathbf{r})}{\partial z}\right|_{z=-d_N} = h(T(\mathbf{r})|_{z=-d_N} - T_a) \tag{4}$$

where T_a is the ambient temperature, and k_N is the thermal conductivity of the bottom layer of the chip. In addition, we enforce the continuity conditions at the interface between adjacent layers within the multilayered chip, i.e.,

$$T(\mathbf{r})|_{z=-d_i+\epsilon} = T(\mathbf{r})|_{z=-d_i-\epsilon} \tag{5}$$

$$k_i \left.\frac{\partial T(\mathbf{r})}{\partial z}\right|_{z=-d_i+\epsilon} = k_{i+1} \left.\frac{\partial T(\mathbf{r})}{\partial z}\right|_{z=-d_i-\epsilon} \tag{6}$$

where ϵ is an infinitely small quantity and k_i is the thermal conductivity of the i^{th} material layer in the multilayered chip structure.

III. Full-Chip Thermal Simulation Algorithm

A. The Green function of the rectangular-shaped multilayered structure

Let $G(\mathbf{r}, \mathbf{r}')$, with $\mathbf{r} = (x, y, z)$ and $\mathbf{r}' = (x', y', z')$, be the distribution of temperature above T_a in the multilayer when a unit point power source of 1W is placed at position $\mathbf{r}'$. Then $G(\mathbf{r}, \mathbf{r}')$ satisfies the equation

$$\nabla^2 G(\mathbf{r}, \mathbf{r}') = -\frac{\delta(\mathbf{r} - \mathbf{r}')}{k_{l(\mathbf{r})}} \tag{7}$$

and the boundary conditions

$$\left.\frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial x}\right|_{x=0,a} = \left.\frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial y}\right|_{y=0,b} = 0 \tag{8}$$

$$\left.\frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial z}\right|_{z=0} = 0 \tag{9}$$

$$k_N \left.\frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial z}\right|_{z=-d_N} = hG(\mathbf{r}, \mathbf{r}')|_{z=-d_N} \tag{10}$$

$$G(\mathbf{r}, \mathbf{r}')|_{z=-d_i+\epsilon} = G(\mathbf{r}, \mathbf{r}')|_{z=-d_i-\epsilon} \tag{11}$$

$$k_i \left.\frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial z}\right|_{z=-d_i+\epsilon} = k_{i+1} \left.\frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial z}\right|_{z=-d_i-\epsilon} \tag{12}$$

where $\delta(\mathbf{r}, \mathbf{r}') = \delta(x - x')\delta(y - y')\delta(z - z')$ is the three-dimensional Dirac delta function, and $G(\mathbf{r}, \mathbf{r}')$ is the Green function. The temperature field under an arbitrary power density distribution can be obtained easily as

$$T(\mathbf{r}) = T_a + \int_0^a dx' \int_0^b dy' \int_{-d_N}^0 dz' G(\mathbf{r}, \mathbf{r}') g(\mathbf{r}') \tag{13}$$

Using a derivation similar to that presented in [12], the Green function can be written in the form

$$G(\mathbf{r}, \mathbf{r}') = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) \cos\left(\frac{m\pi x'}{a}\right) \cos\left(\frac{n\pi y'}{b}\right) Z'_{mn}(z, z') \tag{14}$$

where $Z'_{mn}(z, z')'s$ are functions of only the z coordinates of the source and field points.

B. Full-chip thermal simulation algorithm

In the following analysis, we assume that both the heat sources and the field regions are located on discrete horizontal planes. Since the vertical dimensions of the devices are much smaller than that of the silicon chip, this assumption is reasonable for most practical purposes. For a particular pair of source and field planes, i.e., for a particular z and z', the Green function can be written as

$$G(x, y, x', y') = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} C_{mn} \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) \cos\left(\frac{m\pi x'}{a}\right) \cos\left(\frac{n\pi y'}{b}\right) \tag{15}$$

The temperature distribution on the field plane due to the heat sources on the source plane is given by

$$T(x, y) = T_a + \int_0^a dx' \int_0^b dy' G(x, y, x', y') P_d(x', y') \tag{16}$$

where $P_d(x', y')$ is the power density distribution on the source plane. The convolution integral in (16) can be considered as the governing equation of a linear system determined by the Green function $G(x, y, x', y')$.

As stated previously, the first step of our algorithm is to obtain the frequency domain representation of the power density map in the form

$$P_d(x', y') = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_{ij} \phi_{ij}(x', y') \tag{17}$$

where

$$\phi_{ij}(x, y) = \cos\left(\frac{i\pi x}{a}\right) \cos\left(\frac{j\pi y}{b}\right) \tag{18}$$

It is easy to show that $\phi_{ij}(x, y)$ satisfies the equation

$$\lambda_{ij} \phi_{ij}(x, y) = \int_0^a dx' \int_0^b dy' G(x, y, x', y') \phi_{ij}(x', y') \tag{19}$$

where

$$\lambda_{ij} = \begin{cases} abC_{ij} & \text{if } i = j = 0 \\ \frac{1}{2} abC_{ij} & \text{if } i = 0, j \neq 0 \text{ or } i \neq 0, j = 0 \\ \frac{1}{4} abC_{ij} & \text{if } i \neq 0, j \neq 0 \end{cases} \tag{20}$$

is the response of the linear system to the frequency component $\phi_{ij}(x, y)$. After the frequency domain representation of the power density distribution in the source plane is obtained, the temperature distribution in the field plane can be calculated easily by

$$T(x, y) = T_a + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda_{ij} a_{ij} \phi_{ij}(x, y) \tag{21}$$

As will be shown next, both the frequency decomposition in (17) and the double-summation in (21) can be calculated efficiently using the DCT and IDCT through the FFT.

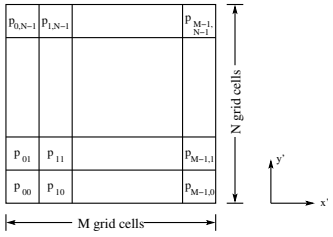


Fig. 2. The arrangement of the $M \times N$ grid cells on the source plane.

Now we assume that both the source plane and the field plane are divided into $M \times N$ rectangular grid cells of equal size as shown in Fig. 2, and the power density in each grid cell on the source plane is uniform, i.e., the power density distribution can be written in the piece-wise constant form

$$P_d(x', y') = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{mn} \Theta(x' - (m + \frac{1}{2})\Delta x, y' - (n + \frac{1}{2})\Delta y) \tag{22}$$

where

$$\Theta(x', y') = \begin{cases} 1 & \text{if } |x'| \leq \frac{1}{2}\Delta x \text{ and } |y'| \leq \frac{1}{2}\Delta y \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

and $\Delta x = \frac{a}{M}, \Delta y = \frac{b}{N}$. P_{mn} is the power density of the mn^{th} grid cell. Substituting (22) into (17) and using the orthogonality property of the cosine functions in the integral sense, we obtain

$$a_{ij} = A_{ij} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{mn} \cos\left(\frac{i\pi(2m+1)}{2M}\right) \cos\left(\frac{j\pi(2n+1)}{2N}\right) \tag{24}$$

where

$$A_{ij} = \begin{cases} \frac{1}{MN} & \text{if } i = j = 0 \\ \frac{4}{iN\pi} \sin\left(\frac{i\pi}{2M}\right) & \text{if } i \neq 0, j = 0 \\ \frac{4}{Mj\pi} \sin\left(\frac{j\pi}{2N}\right) & \text{if } i = 0, j \neq 0 \\ \frac{16}{ij\pi^2} \sin\left(\frac{i\pi}{2M}\right) \sin\left(\frac{j\pi}{2N}\right) & \text{if } i \neq 0, j \neq 0 \end{cases} \tag{25}$$

Note that to accurately represent the power density distribution $P_d(x', y')$ using (17), the theoretical upper limit of the double summation should be infinity. In practical implementations, however, the summation must be truncated to ensure a reasonable runtime. Since (17) is essentially the Fourier expansion of $P_d(x', y')$, a natural criterion for determining the truncation point is that enough "energy" contained in $P_d(x', y')$ is covered by the truncated Fourier expansion. Mathematically, we have

$$\int_0^a dx' \int_0^b dy' P_d^2(x', y') = ab \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} s_{ij} a_{ij}^2 \tag{26}$$

where

$$s_{ij} = \begin{cases} 1 & \text{if } i = j = 0 \\ \frac{1}{2} & \text{if } i = 0, j \neq 0 \text{ or } i \neq 0, j = 0 \\ \frac{1}{4} & \text{if } i \neq 0, j \neq 0 \end{cases} \tag{27}$$

Substituting (22) into the left hand side of (26), we obtain

$$\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{mn}^2 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} s_{ij} a_{ij}^2 \tag{28}$$

which can be considered as a form of the Parseval's theorem. The truncation points M' and N' are then determined by

$$\sum_{i=0}^{M'-1} \sum_{j=0}^{N'-1} s_{ij} a_{ij}^2 \geq \eta \left(\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{mn}^2 \right) \tag{29}$$

where η is the proportion of the "energy" of the space domain signal $P_d(x', y')$ that must be covered by the truncated Fourier expansion. In practice, we found that setting η to 90% will usually be enough to obtain very accurate results in temperature calculations.

Note that for $0 \le i < M$ and $0 \le j < N$, the double summation in (24) can be considered as a term in the 2D type-II DCT [13] of the power density matrix P. For $i \ge M$ or $j \ge N$, we can always find integers s_1 and s_2 such that $i = 2s_1M \pm \hat{i}$ and $j = 2s_2N \pm \hat{j}$ where $0 \le \hat{i} < M$ and $0 \le \hat{j} < N$. Hence, for any i and j, we always have

$$a_{ij} = \pm A_{ij} \widetilde{P}_{\hat{i}\hat{j}} \tag{30}$$

where

$$\widetilde{P}_{\hat{i}\hat{j}} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{mn} \cos\left(\frac{\hat{i}\pi(2m+1)}{2M}\right) \cos\left(\frac{\hat{j}\pi(2n+1)}{2N}\right) \tag{31}$$

with $0 \le \hat{i} < M$ and $0 \le \hat{j} < N$ is the 2D type-II DCT of the P matrix and the sign of (30) is determined by whether s_1 and s_2 are even or odd numbers. Equation (31) can be calculated efficiently using the 2D FFT in $O((M \cdot N) \times \log(M \cdot N))$ time. After the 2D DCT matrix $\widetilde{P}$ is obtained, the calculation of a_{ij} simply involves computing the coefficient A_{ij} and finding the corresponding term $\widetilde{P}_{\hat{i}\hat{j}}$.

From (18) and (21), the temperature distribution $T(x, y)$ can now be written as

$$T(x,y) = T_a + \sum_{i=0}^{M'-1} \sum_{j=0}^{N'-1} \lambda_{ij} a_{ij} \cos\left(\frac{i\pi x}{a}\right) \cos\left(\frac{j\pi y}{b}\right) \tag{32}$$

and the average temperature of the mn^{th} grid cell can be obtained by

$$\begin{aligned} T_{mn} &= \frac{1}{\Delta x \Delta y} \int_{m\Delta x}^{(m+1)\Delta x} dx \int_{n\Delta y}^{(n+1)\Delta y} dy T(x,y) \\ &= T_a + \frac{MN}{ab} \sum_{i=0}^{M'-1} \sum_{j=0}^{N'-1} B_{ij} \cos\left(\frac{i\pi(2m+1)}{2M}\right) \cos\left(\frac{j\pi(2n+1)}{2N}\right) \end{aligned} \tag{33}$$

where

$$B_{ij} = \begin{cases} \lambda_{ij} a_{ij} \frac{ab}{MN} & \text{if } i = j = 0 \\ 2\lambda_{ij} a_{ij} \frac{ab}{iN\pi} \sin\left(\frac{i\pi}{2M}\right) & \text{if } i \ne 0, j = 0 \\ 2\lambda_{ij} a_{ij} \frac{ab}{Mj\pi} \sin\left(\frac{j\pi}{2N}\right) & \text{if } i = 0, j \ne 0 \\ 4\lambda_{ij} a_{ij} \frac{ab}{ij\pi^2} \sin\left(\frac{i\pi}{2M}\right) \sin\left(\frac{j\pi}{2N}\right) & \text{if } i \ne 0, j \ne 0 \end{cases} \tag{34}$$

As stated previously, any $i \ge M$ and $j \ge N$ can be written as $i = 2s_1M \pm \hat{i}$ and $j = 2s_2N \pm \hat{j}$ such that $0 \le \hat{i} < M$, $0 \le \hat{j} < N$ and s_1,s_2 are integers. Using the periodicity of the cosine function, we can finally cast T_{mn} into the form

$$T_{mn} = T_a + \sum_{\hat{i}=0}^{M-1} \sum_{\hat{j}=0}^{N-1} L_{\hat{i}\hat{j}} \cos\left(\frac{\hat{i}\pi(2m+1)}{2M}\right) \cos\left(\frac{\hat{j}\pi(2n+1)}{2N}\right) \tag{35}$$

where $L_{\hat{i}\hat{j}} = K_{\hat{i}\hat{j}} \widetilde{P}_{\hat{i}\hat{j}}$ and $K_{\hat{i}\hat{j}}$ can be calculated as follows

1) if $\hat{i} = \hat{j} = 0$

$$K_{\hat{i}\hat{j}} = \frac{\lambda_{00}}{MN} \tag{36}$$

2) if $\hat{i} \ne 0, \hat{j} = 0$

$$K_{\hat{i}\hat{j}} = \frac{8M}{N\pi^2} \sin^2\left(\frac{\hat{i}\pi}{2M}\right) \sum_{\substack{i < M' \\ i = 2s_1M \pm \hat{i}}} \frac{\lambda_{i0}}{i^2} \tag{37}$$

3) if $\hat{i} = 0, \hat{j} \ne 0$

$$K_{\hat{i}\hat{j}} = \frac{8N}{M\pi^2} \sin^2\left(\frac{\hat{j}\pi}{2N}\right) \sum_{\substack{j < N' \\ j = 2s_2N \pm \hat{j}}} \frac{\lambda_{0j}}{j^2} \tag{38}$$

4) if $\hat{i} \ne 0, \hat{j} \ne 0$

$$K_{\hat{i}\hat{j}} = \frac{64MN}{\pi^4} \sin^2\left(\frac{\hat{i}\pi}{2M}\right) \sin^2\left(\frac{\hat{j}\pi}{2N}\right) \sum_{\substack{i < M' \\ i = 2s_1M \pm \hat{i}}} \sum_{\substack{j < N' \\ j = 2s_2N \pm \hat{j}}} \frac{\lambda_{ij}}{i^2 j^2} \tag{39}$$

Input:
- Chip geometry and physical properties of the material layers.
- Power density map - matrix P.

Output: Temperature distribution map - matrix T.
Algorithm:
1) Calculate the Green function coefficients ${C_{ij}}'s$;
2) Calculate the frequency responses of the system ${\lambda_{ij}}'s$;
3) Calculate the type-II 2D DCT of the power density matrix $\widetilde{P} = \text{2DDCT}(P)$;
4) $TSE = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{mn}^2$;
5) $M' = M,\ N' = N$;
   $ASE = \sum_{i=0}^{M'-1} \sum_{j=0}^{N'-1} s_{ij} a_{ij}^2$;
   while ($ASE < \eta \times TSE$)
   $M' = M' + M,\ N' = N' + N$;
   Update ASE;
   end while;
6) Calculate the matrix K;
7) Calculate the matrix L with $L_{\hat{i}\hat{j}} = K_{\hat{i}\hat{j}} \widetilde{P}_{\hat{i}\hat{j}}$;
8) Calculate the temperature distribution map using the type-II 2D IDCT $T = T_a + \text{2DIDCT}(L)$;

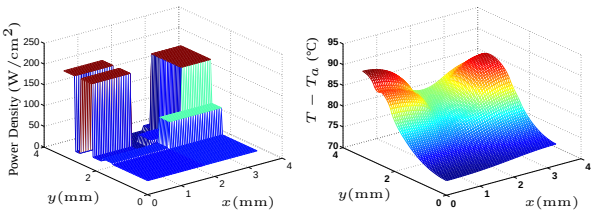
Fig. 3. Thermal simulation algorithm using the Green function method, the DCT, and the frequency domain computations.

After the coefficients $K'_{\hat{i}\hat{j}}s$ are calculated, the matrix L can be easily obtained by point-wise multiplying the matrices K and $\widetilde{P}$. The double summation in (35) can then be calculated efficiently using the 2D IDCT.

The complete thermal simulation algorithm using the Green function method, the DCT, and the frequency domain computations is shown in Fig. 3. The asymptotic time complexity of the algorithm is $O((M \cdot N) \times \log(M \cdot N))$ where $M \cdot N$ is the total number of grid cells. This is a significant improvement over the $O((M \cdot N)^2)$ complexity of the algorithm given in [6]. Note that up to now, we have focused on the effect of one source layer on the temperature distribution of the field layer. When multiple source layers are present, such as in the emerging 3D IC technology, their effects can be calculated individually and summed up. The ambient temperature T_a should only appear once in the final summation.

IV. Experimental Results

A. Accuracy and Efficiency of the Algorithm



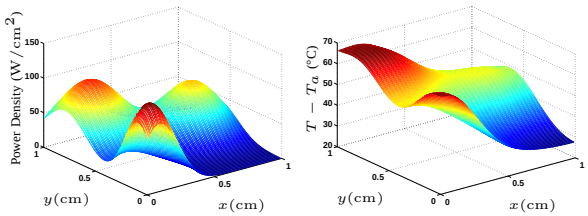
(a) (b)

Fig. 4. Power density and temperature distribution of an example chip (a) power density distribution (b) temperature distribution.

Because the method presented in [6] is accurate except for the very small truncation error of the Green function, we use the result in [6] as the benchmark to test the accuracy and efficiency of our algorithm. Fig. 4 shows the power distribution of an example chip and the calculated temperature map using the algorithm proposed in this paper. The heat sources are assumed to be located on the top surface of the chip and the size of the temperature map is 64×64. The chip has dimensions of 3.3mm×3.3mm×0.5mm. The thermal conductivity of silicon is 148W/(m·°C) and the bottom surface of the chip has an effective heat transfer coefficient of 8700W/(m^2·°C). We require our algorithm to achieve a similar accuracy as that in [6] by choosing η in (29) appropriately, i.e., within 1% error compared with the results from commercial computational fluid dynamic softwares, and the runtimes of the two algorithms are compared. Each runtime is divided into two parts, i.e., the time spent on the steps that are independent of the input power density matrix and hence can be pre-calculated outside the optimization loop in thermal-aware designs, and the time spent on the steps that depend on the input power density matrix and hence must be executed within the optimization loop. For both algorithms, the Green function coefficients ${C_{ij}}'s$ can always be pre-calculated and stored. In addition, the look-up tables in the algorithm in [6] can be pre-calculated while the frequency responses ${\lambda_{ij}}'s$ in our algorithm can be pre-calculated. For the pre-calculated steps, the runtime is dominated by the computation

of the coefficients $C_{ij}{}'s$ in both algorithms, which may take about 95sec to obtain a 2048×2048 C matrix. However, since these steps only need to be executed once for each die geometry and then used many times in later thermal simulations, the amortized cost is usually extremely small and we will ignore this part of the runtime in further analysis. Experimental results show that for the steps that are not pre-calculated, the runtime of our algorithm using MATLAB is 0.09sec while that of the algorithm in [6] is 128sec. Note that the runtime of the algorithm in [6] is linear with respect to the number of heat sources and there are only 13 heat sources in the example given here. For cell level full-chip simulations where the number of heat sources is significantly larger, the advantages of our algorithm will become even more obvious.

B. Cell Level Full-Chip Thermal Simulation



(a) (b)

Fig. 5. Cell level power density and temperature distribution of a 1cm×1cm chip (a) power density distribution (b) temperature distribution.

In this subsection, we show an example of cell level full-chip thermal simulations. We consider a chip with dimensions of 1cm×1cm×0.5mm and the same physical properties as the chip used in the previous example. There are 1024×1024 square grid cells of equal size located on the top surface of the chip and a 1024×1024 temperature distribution map of the cell layer is calculated. Fig. 5 shows the input power density map and the resulting temperature map. The time it takes for MATLAB to obtain this temperature map containing 1.05M grid cells is only 6.4sec excluding the time for the pre-calculations while the runtime of the algorithm in [6] becomes intractable.

V. Discussions - Strategies for Performing the Thermal Simulations with Local High Accuracy Requirements

The situation frequently arises in real design environments where the accuracy requirements on the thermal simulation differ from place to place on the same chip. For example, in mixed signal designs where the analog circuits are fabricated on the same chip as the digital circuits, the analog blocks often have more stringent accuracy requirements on the thermal simulation because the operations of the analog circuits are more sensitive to temperature. For these kinds of problems, a better strategy can be adopted to accelerate the runtime of the algorithm further. The key idea is to use a coarse grid to divide the source and field layers such that each grid cell can contain several logic gates or analog functional units. The power density of each grid cell is calculated by summing up the power dissipations of all the logic gates and analog functional units located in it and divide the sum by the area of the grid cell. A coarse temperature map is then obtained from the coarse power density map using the algorithm presented in section III and is used for the digital blocks on the chip. Note that we must ensure that the coarse grid is fine enough for the digital blocks but they may not achieve the accuracy requirements of the analog blocks.

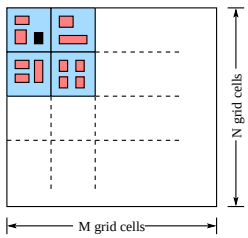


Fig. 6. A mixed signal chip where the analog block has higher requirement on the accuracy of the thermal simulation.

Fig. 6 shows a chip that is divided into $M \times N$ coarse grid cells each of which contains several logic gates or analog functional units, and let the shaded area represent the analog block. An $M \times N$ temperature map is first obtained. The inaccuracies in the temperature calculations, besides that due to the truncation of the eigen-decomposition of the power density map, will come from two sources which include

- Assuming that the power density in each grid cell is uniform.
- Only the average temperature of each grid cell is calculated, i.e., all the logic gates and analog functional units inside the same grid cell obtain the same calculated temperature.

Now assume that we need to calculate the temperature of the analog functional unit located in the ij^{th} grid cell and represented by the black rectangle more accurately. Let T_{ij} and $T_{ij,kl}$ be the average temperature of the ij^{th} grid cell and the contribution of the average power density of the kl^{th} grid cell to the average temperature of the ij^{th} grid cell in the coarse grid temperature calculations respectively, and let $T_{a.c.}$ represent the more accurate average temperature of the analog functional unit. We divide the grid cells into two categories, i.e., those with close interactions with the ij^{th} grid cell (denoted by CI_{ij}) and those without close interactions with the ij^{th} grid cell. The effects of the logic gates and analog functional units that are contained in the grid cells belonging to CI_{ij} are re-calculated for increased accuracy. For example, we can put the ij^{th} grid cell and all the grid cells surrounding it into the first category and all the other grid cells into the second category. If higher accuracy is required, then more grid cells should be put into the first category. The temperature $T_{a.c.}$ can then be calculated using

$$T_{a.c.} = T_{ij} - \sum\nolimits_{kl \in CI_{ij}} T_{ij,kl} + \sum\nolimits_{s} T_s^{\text{gate and functional unit}} \tag{40}$$

where $T_s^{\text{gate and functional unit}}$ is the contribution to $T_{a.c.}$ from the s^{th} logic gate or analog functional unit in the grid cells that have close interactions with the ij^{th} grid cell. Both $T_{ij,kl}$ and $T_s^{\text{gate and functional unit}}$ can be calculated efficiently using the table look-up approach given in [6] and will not be reiterated here.

VI. Conclusions

In this paper, we presented a cell level full-chip thermal simulation algorithm that is a combination of the Green function method, the DCT, and the frequency domain computations. Experimental results show that our algorithm can achieve orders of magnitude speedup compared with previous Green function based thermal simulation algorithms while still maintain the same accuracy. The simulation of a chip containing 1.05M grid cells only takes about 6.4sec after the pre-calculations have been performed. In addition, the strategies that can be used for the problems that have local high accuracy requirements on temperature calculations are also discussed.

References

[1] C. H. Tsai and S. M. Kang, "Cell-Level Placement for Improving Substrate Thermal Distribution," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 2, pp. 253-266, Feb. 2000.

[2] B. Goplen and S. S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs Using a Force Directed Approach," *Digest of Technical Papers, IEEE/ACM International Conference on Computer-Aided Design*, pp. 86-89, Nov. 2003.

[3] T. Y. Wang and 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*, vol. 21, no. 12, pp. 1434-1445, Dec. 2002.

[4] Y. K. Cheng and S. M. Kang, "An Efficient Method for Hot-Spot Identification in ULSI Circuits," *Digest of Technical Papers, IEEE/ACM International Conference of Computer-Aided Design*, pp. 124-127, Nov. 1999.

[5] B. Wang and P. Mazumder, "Fast Thermal Analysis for VLSI Circuits via Semi-analytical Green's Function in Multi-layer Materials," *IEEE International Symposium on Circuits and Systems*, pp. 409-412, May 2004.

[6] Y. Zhan and S. S. Sapatnekar, "Fast Computation of the Temperature Distribution in VLSI Chips Using the Discrete Cosine Transform and Table Look-up," *Proceedings of the 2005 Asia and South Pacific Design Automation Conference*, pp. 87-92, Jan. 2005.

[7] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, "Efficient Full-Chip Thermal Modeling and Analysis," *Digest of Technical Papers, IEEE/ACM International Conference of Computer-Aided Design*, pp. 319-326, Nov. 2004.

[8] J. P. Costa, M. Chou, and L. M. Silveira, "Efficient Techniques for Accurate Modeling and Simulation of Substrate Coupling in Mixed-Signal IC's," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, no. 5, pp. 597-607, May 1999.

[9] M. N. Ozisik, "Boundary Value Problems of Heat Conduction," Oxford University Press, Oxford, UK, 1968.

[10] A. G. Kokkas, "Thermal Analysis of Multi-Layer Structures," *IEEE Transactions on Electron Devices*, vol. 21, no. 11, pp. 674-681, Nov. 1974.

[11] Y. K. Cheng, P. Raha, C. C. Teng, E. Rosenbaum, and S. M. Kang, "ILLIADS-T: An Electrothermal Timing Simulator for Temperature-Sensitive Reliability Diagnosis of CMOS VLSI Chips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 8, pp. 668-681, Aug. 1998.

[12] R. Gharpurey, "Modeling and Analysis of Substrate Coupling in Integrated Circuits," Ph. D. Thesis, UC Berkeley, CA, 1995.

[13] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, " Discrete-Time Signal Processing," Prentice Hall, Upper Saddle River, NJ, 1999.