Fast Computation of the Temperature Distribution in VLSI Chips Using the Discrete Cosine Transform and Table Look-up

Yong Zhan and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering
University of Minnesota
{yongzhan, sachin}@ece.umn.edu

***Abstract*— Temperature-related effects are critical in determining both the performance and reliability of VLSI circuits. Accurate and efficient estimation of the temperature distribution corresponding to a specific circuit layout is indispensable in physical design automation tools. In this paper, we propose a highly accurate fast algorithm for computing the on-chip temperature distribution due to power sources located on the top surface of the chip. The method is a combination of several computational techniques including the Green function method, the discrete cosine transform (DCT), and the table look-up technique. The high accuracy of the algorithm comes from the fully analytical nature of the Green function method, and the high efficiency is due to the application of the fast Fourier transform (FFT) technique to compute the DCT and later obtaining the temperature field for any power source distribution using the pre-calculated look-up table. Experimental results have demonstrated that our method has a relative error of below 1% compared with commercial computational fluid dynamic (CFD) softwares for thermal analysis, while the efficiency of our method is orders of magnitude higher than the direct application of the Green function method.**

I. Introduction

The continuously increasing computational capability of modern VLSI circuits has resulted in higher and higher power consumption, and hence a significant increase in chip temperature. The temperature-related effects are very important in determining both the performance and reliability of the circuit. As pointed out in [1], the delay of aluminum interconnect goes up by 30% when the temperature rises from 25°C to 100°C, and it was reported in [2] that the electromigration-induced mean-time-to-failure of interconnect is reduced by 90% when the temperature increases from 25°C to 52.5°C. Thus, it is crucial to incorporate the thermal effects into the physical design tools of VLSI circuits, such as in [3] where a cell-level placement algorithm for improving the substrate thermal distribution was implemented.

The first step towards developing a powerful thermal-aware physical design tool is to be able to compute the temperature field on a chip quickly and accurately given a power source distribution. The efficiency of the temperature-computation algorithm is of paramount importance because it is often used as part of the simulation core of an optimization engine where a huge number of different physical layouts are compared and an independent temperature field computation has to be performed for each layout.

Generally speaking, there are three major types of methods that can be used to obtain the on-chip temperature distribution. The finite difference method (FDM) discretizes the differential operator of the governing equation of the thermal effect [3] [4]. This method is very robust and accurate, provided a fine discretization is used. However, since the volume meshing of the entire substrate is required even though the devices are fabricated only in a thin layer close to the top surface of the chip, the efficiency of the algorithm is relatively low. The second approach is the finite element method (FEM) where the field quantity is discretized [5]. The advantage of the FEM is its high flexibility because it can handle complicated geometric shapes and various kinds of boundary conditions effectively. However, the FEM also has the drawback of low efficiency because, as the FDM, it also requires the meshing of the entire chip volume. The third approach is based on the Green function method where the temperature field under a unity point power source is first obtained and then the temperature field under arbitrary power source distribution can be computed by integrating the corresponding Green function. The Green function method avoids the volume meshing problem by only focusing on the power-generating surfaces, which results in high efficiency. In addition, due to the fully analytical nature of the Green function, the accuracy of the computation is also maintained. In [6], the Green function approach was used to find the temperature field in rectangular substrates. However, since the underlying Green function is expressed as a multiple-infinite summation and it has to be truncated at high indices in real implementations to maintain reasonable accuracy, the efficiency of the method is not fully reflected. In [7], the method of images was used to obtain the Green function in closed form at the expense of relaxing the boundary conditions by assuming the chip is infinitely large horizontally. The advantage of this method is that the Green function can be computed on-line very quickly and thus it is suitable for optimization purposes. However, by assuming the chip is infinitely large horizontally, only the thermal effects that are far from the vertical chip boundaries can be accurately computed, which results in the method unsuitable for full-chip thermal simulations. In [8], an efficient algorithm for evaluating the temperature field in VLSI chips using the semi-analytical Green function of multilayered materials was proposed. However, this method also assumes that the chip is infinitely large horizontally, and hence it has the same problem as [7].

The computation of the steady state temperature field T in thermal problems is very similar to the computation of the potential field ϕ in electrical problems. Both T and ϕ satisfy the Poisson's equation, and the power source P in thermal problems corresponds to the charge q in electrical problems. However, the boundary conditions are different in the two problems, and therefore the solutions also differ. In [9] and [10], the discrete cosine transform (DCT) and a pre-calculated look-up table were used in a highly accurate and efficient Green function based algorithm to calculate the electric potential distribution in rectangular substrates. In this method, it is not necessary to compute the multiple-infinite summation in the Green function on-line. Instead, the potential can be computed by the simple summation of a few terms in the pre-calculated look-up table. Since the look-up table only needs to be computed once for each technology and substrate geometry, but is independent of where the devices are located on the chip, it can be used many times in the optimization process of the physical design, which results in a significant improvement in efficiency.

In this work, we adopted the DCT and table look-up approaches to calculate the steady state temperature distribution on rectangular VLSI chips. The Green function suitable for the special boundary conditions of the thermal problem was first developed. Then the DCT look-up table and vectors were set up to assist the evaluation of the temperature field. Experimental results show that our method has a relative error of below 1% compared with commercial computational fluid dynamic (CFD) softwares for thermal analysis, while the efficiency of our method is orders of magnitude higher than the direct application of the Green function method. The rest of the paper will be organized as follows. Section II formulates the temperature field computation problem under the appropriate boundary conditions. Section III develops the look-up table in detail using the Green function method and DCT. Section IV shows the experimental results. Section V discusses the extension of the proposed algorithm to the multilayered substrate and packaging structures, and the conclusions are given in section VI.

II. Problem formulation

Fig. 1 shows a schematic of a VLSI chip with the associated packaging. The shaded areas on the top surface of the chip represent the devices. The

This work was supported in part by DARPA under grant N66001-04-1-8909, SRC under contract 2003-TJ-1092, and NSF under award CCR-0205227.

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

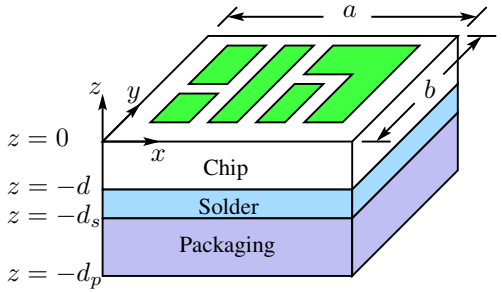


Fig. 1. Schematic of a VLSI chip with packaging.

heat conduction and temperature distribution inside the chip are governed by the heat diffusion equation

$$\rho c_p \frac{\partial T(x,y,z,t)}{\partial t} = \nabla \cdot [k(x,y,z,T)\nabla T(x,y,z,t)] + g(x,y,z,t) \quad (1)$$

where T is the temperature (°C), k is the thermal conductivity (W/(m·°C)), g is the volume power density (W/m^3), ρ is the density of the material (kg/m^3), and c_p is the specific heat (J/(kg·°C)) [11]. Since in microelectronics, all the devices that can dissipate power are fabricated in a very thin layer close to the top surface of the chip, it is reasonable to assume that there is no power source inside the chip and thus g can be set to zero. The power dissipation of the devices will be reflected in the top surface boundary condition. In addition, for steady state temperature distribution, the time derivative of T in equation (1) also vanishes. If we further assume that the thermal conductivity k is a constant, then equation (1) can be simplified to

$$\nabla^2 T(x,y,z) = 0 \quad (2)$$

which is the well-known Laplace's equation. The boundary conditions at the vertical surfaces of the chip are obtained by making the adiabatic assumption [12], which results in

$$\left.\frac{\partial T(x,y,z)}{\partial x}\right|_{x=0,a} = \left.\frac{\partial T(x,y,z)}{\partial y}\right|_{y=0,b} = 0 \quad (3)$$

For the bottom surface of the chip, we use the convection boundary condition

$$\left. k\frac{\partial T(x,y,z)}{\partial z}\right|_{z=-d} = h(T(x,y,z)|_{z=-d} - T_a) \quad (4)$$

where h is the combined effective heat transfer coefficient of the solder, packaging and ambient (W/(m^2·°C)), and T_a is the ambient temperature [13]. The top surface of the chip can be assumed to be adiabatic since a thick passivation layer is usually placed on the top of the IC chip in fabrication. The net result of the adiabatic assumption about the top surface is that all the power generated by the devices are conducted into the substrate, hence, the boundary condition at the top surface becomes

$$\left. k\frac{\partial T(x,y,z)}{\partial z}\right|_{z=0} = P_d(x,y) \quad (5)$$

where $P_d(x,y)$ is the surface power density (W/m^2) due to the power dissipation of devices [12]. Let $T' = T - T_a$, equation (2) and boundary conditions (3)-(5) then become

$$\nabla^2 T'(x,y,z) = 0 \quad (6)$$

$$\left.\frac{\partial T'(x,y,z)}{\partial x}\right|_{x=0,a} = \left.\frac{\partial T'(x,y,z)}{\partial y}\right|_{y=0,b} = 0 \quad (7)$$

$$\left. k\frac{\partial T'(x,y,z)}{\partial z}\right|_{z=-d} = hT'(x,y,z)|_{z=-d} \quad (8)$$

$$\left. k\frac{\partial T'(x,y,z)}{\partial z}\right|_{z=0} = P_d(x,y) \quad (9)$$

Equations (6)-(9) constitute the problem we are solving in this paper.

III. Development of the Look-up Table Using the Green Function Method and DCT

A. Green function method

Let $G(x,y,z,x',y')$ be the distribution of temperature above T_a in the chip when a unity point power source of 1W is placed at point (x',y') on the top surface of the chip. Then $G(x,y,z,x',y')$ satisfies the equation

$$\nabla^2 G(x,y,z,x',y') = 0 \quad (10)$$

and the boundary conditions

$$\left.\frac{\partial G(x,y,z,x',y')}{\partial x}\right|_{x=0,a} = \left.\frac{\partial G(x,y,z,x',y')}{\partial y}\right|_{y=0,b} = 0 \quad (11)$$

$$\left. k\frac{\partial G(x,y,z,x',y')}{\partial z}\right|_{z=-d} = hG(x,y,z,x',y')|_{z=-d} \quad (12)$$

$$\left. k\frac{\partial G(x,y,z,x',y')}{\partial z}\right|_{z=0} = \delta(x-x')\delta(y-y') \quad (13)$$

where $\delta(x-x')$ and $\delta(y-y')$ are the Dirac delta functions. $G(x,y,z,x',y')$ is called the Green function, and the temperature field under arbitrary surface power density distribution can be obtained easily by

$$T'(x,y,z) = \int_0^a dx' \int_0^b dy' G(x,y,z,x',y') P_d(x',y') \quad (14)$$

The validity of (14) can be proved as follows. Due to the linearity of integration, it is obviously true that $T'(x,y,z)$ satisfies equations (6)-(8) if $G(x,y,z,x',y')$ satisfies equations (10)-(12). Hence, we only need to show that the expression in (14) satisfies equation (9). Exchanging the order of partial derivative and integration, we obtain

$$\begin{aligned}
&\left. k\frac{\partial T'(x,y,z)}{\partial z}\right|_{z=0} \\
&= k\left[\frac{\partial}{\partial z}\int_0^a dx' \int_0^b dy' G(x,y,z,x',y') P_d(x',y')\right]\Bigg|_{z=0} \\
&= k\int_0^a dx' \int_0^b dy' \left.\frac{\partial G(x,y,z,x',y')}{\partial z}\right|_{z=0} P_d(x',y') \\
&= \int_0^a dx' \int_0^b dy' \delta(x-x')\delta(y-y') P_d(x',y') = P_d(x,y)
\end{aligned} \quad (15)$$

which proves that equation (9) is indeed satisfied.

B. Computation of the Green function

Now we compute the analytical form of the Green function. Using separation of variables, $G(x,y,z,x',y')$ can be written as

$$G(x,y,z,x',y') = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) Z_{mn}(z) \quad (16)$$

The x and y dependencies in (16) ensure that the boundary condition (11) is satisfied. The Laplace's equation (10) now becomes

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left(\frac{d^2 Z_{mn}(z)}{dz^2} - \gamma_{mn}^2 Z_{mn}(z)\right)\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) = 0 \quad (17)$$

where $\gamma_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2}$. Due to the orthogonality of the cosine functions in the integral sense, we obtain

$$\frac{d^2 Z_{mn}(z)}{dz^2} - \gamma_{mn}^2 Z_{mn}(z) = 0 \quad (18)$$

for arbitrary m and n. We require that each $Z_{mn}(z)$ satisfy

$$\left. k\frac{dZ_{mn}(z)}{dz}\right|_{z=-d} = hZ_{mn}(z)|_{z=-d} \quad (19)$$

such that the boundary condition (12) is satisfied. Boundary condition (13) can be cast into the form

$$k\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right)\left.\frac{dZ_{mn}(z)}{dz}\right|_{z=0}=\delta(x-x')\delta(y-y') \tag{20}$$

Thus, in the following derivations, it is only necessary to focus on equations (18)-(20) to find $Z_{mn}(z)$. We consider two different cases

1) $m=n=0$

Equation (18) becomes

$$\frac{d^2Z_{00}(z)}{dz^2}=0 \tag{21}$$

Thus,

$$Z_{00}(z)=\alpha_{00}z+\beta_{00} \tag{22}$$

From equation (19), we obtain

$$(k+hd)\alpha_{00}=h\beta_{00} \tag{23}$$

Integrating both sides of equation (20) over the top surface of the chip, we obtain

$$kab\cdot\alpha_{00}=1 \tag{24}$$

Hence

$$\alpha_{00}=\frac{1}{kab} \tag{25}$$

$$\beta_{00}=\frac{k+hd}{kabh} \tag{26}$$

2) m and n are not both equal to zero

From equation (18), we obtain

$$Z_{mn}(z)=\alpha_{mn}e^{\gamma_{mn}z}+\beta_{mn}e^{-\gamma_{mn}z} \tag{27}$$

From equation (19), we obtain

$$(h-k\gamma_{mn})e^{-\gamma_{mn}d}\alpha_{mn}+(h+k\gamma_{mn})e^{\gamma_{mn}d}\beta_{mn}=0 \tag{28}$$

Multiplying both sides of equation (20) by $\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right)$ and integrating over the top surface of the chip, we obtain

$$\frac{k}{s}ab(\alpha_{mn}-\beta_{mn})\gamma_{mn}=\cos\left(\frac{m\pi x'}{a}\right)\cos\left(\frac{n\pi y'}{b}\right) \tag{29}$$

where $s=4$ if $m\neq0$ and $n\neq0$, and $s=2$ if one of m and n is zero. Hence

$$\alpha_{mn}=\frac{\frac{s}{abk\gamma_{mn}}(h+k\gamma_{mn})}{(h+k\gamma_{mn})+(h-k\gamma_{mn})e^{-2\gamma_{mn}d}}\times\cos\left(\frac{m\pi x'}{a}\right)\cos\left(\frac{n\pi y'}{b}\right) \tag{30}$$

$$\beta_{mn}=\frac{-\frac{s}{abk\gamma_{mn}}(h-k\gamma_{mn})e^{-2\gamma_{mn}d}}{(h+k\gamma_{mn})+(h-k\gamma_{mn})e^{-2\gamma_{mn}d}}\times\cos\left(\frac{m\pi x'}{a}\right)\cos\left(\frac{n\pi y'}{b}\right) \tag{31}$$

Thus, we have obtained the analytical form of the Green function as a double-infinite summation.

Since all the devices and interconnects in a VLSI circuit are fabricated in a thin layer close to the top surface of the chip, we are only interested in the temperature distribution on the $z=0$ plane. Setting z to 0 in the Green function, we obtain

$$G'(x,y,x',y')\triangleq G(x,y,z=0,x',y')=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}C_{mn}\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right)\cos\left(\frac{m\pi x'}{a}\right)\cos\left(\frac{n\pi y'}{b}\right) \tag{32}$$

where

$$C_{mn}=\begin{cases}\frac{k+hd}{kabh} & \text{if } m=n=0\\ \frac{\frac{2}{abk\gamma_{mn}}((h+k\gamma_{mn})-(h-k\gamma_{mn})e^{-2\gamma_{mn}d})}{(h+k\gamma_{mn})+(h-k\gamma_{mn})e^{-2\gamma_{mn}d}} & \text{if } m\neq0, n=0 \text{ or } m=0, n\neq0\\ \frac{\frac{4}{abk\gamma_{mn}}((h+k\gamma_{mn})-(h-k\gamma_{mn})e^{-2\gamma_{mn}d})}{(h+k\gamma_{mn})+(h-k\gamma_{mn})e^{-2\gamma_{mn}d}} & \text{if } m\neq0, n\neq0\end{cases} \tag{33}$$

C. Building the look-up table using the DCT

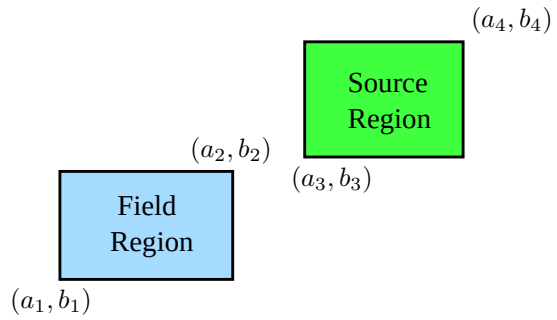


Fig. 2. Source and field regions for computing the surface temperature distribution.

As pointed out previously, for thermal problems, both the source region, where the power is generated, and the field region, whose temperature is to be computed, are located near the top surface of the chip. Thus, in the following analysis, we will focus on the $z=0$ plane and use the Green function given in equation (32). Fig. 2 shows two rectangular regions on the top surface of the chip. Assume the total power generated by the source region is P_s and it is uniformly distributed. Then the power density of the source region is given by

$$P_d(x',y')=\frac{P_s}{(a_4-a_3)(b_4-b_3)}\quad\text{where } (x',y')\in\text{source region} \tag{34}$$

The average temperature in the field region can be computed using

$$\overline{T_f}=\frac{1}{(a_2-a_1)(b_2-b_1)}\int_{a_1}^{a_2}dx\int_{b_1}^{b_2}dyT(x,y,z=0) \tag{35}$$

Applying equation (14) with z set to 0, we obtain

$$\begin{aligned}\overline{T_f}=&\frac{P_s}{(a_2-a_1)(b_2-b_1)(a_4-a_3)(b_4-b_3)}\times\\ &\int_{a_1}^{a_2}dx\int_{b_1}^{b_2}dy\int_{a_3}^{a_4}dx'\int_{b_3}^{b_4}dy'G'(x,y,x',y')=\\ &C_{00}P_s+\left\{\frac{P_s}{(a_2-a_1)(a_4-a_3)}\sum_{m=0}^{\infty}D_{m0}\left[\sin\left(\frac{m\pi a_2}{a}\right)-\sin\left(\frac{m\pi a_1}{a}\right)\right]\times\right.\\ &\left.\left[\sin\left(\frac{m\pi a_4}{a}\right)-\sin\left(\frac{m\pi a_3}{a}\right)\right]\right\}+\left\{\frac{P_s}{(b_2-b_1)(b_4-b_3)}\times\right.\\ &\left.\sum_{n=0}^{\infty}E_{0n}\left[\sin\left(\frac{n\pi b_2}{b}\right)-\sin\left(\frac{n\pi b_1}{b}\right)\right]\left[\sin\left(\frac{n\pi b_4}{b}\right)-\sin\left(\frac{n\pi b_3}{b}\right)\right]\right\}+\\ &\left\{\frac{P_s}{(a_2-a_1)(b_2-b_1)(a_4-a_3)(b_4-b_3)}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}F_{mn}\times\right.\\ &\left[\sin\left(\frac{m\pi a_2}{a}\right)-\sin\left(\frac{m\pi a_1}{a}\right)\right]\left[\sin\left(\frac{m\pi a_4}{a}\right)-\sin\left(\frac{m\pi a_3}{a}\right)\right]\times\\ &\left.\left[\sin\left(\frac{n\pi b_2}{b}\right)-\sin\left(\frac{n\pi b_1}{b}\right)\right]\left[\sin\left(\frac{n\pi b_4}{b}\right)-\sin\left(\frac{n\pi b_3}{b}\right)\right]\right\}\end{aligned} \tag{36}$$

where

$$D_{m0} = \begin{cases} C_{m0}\left(\frac{a}{m\pi}\right)^2 & \text{if } m{\neq}0 \\ 0 & \text{if } m = 0 \end{cases} \quad (37)$$

$$E_{0n} = \begin{cases} C_{0n}\left(\frac{b}{n\pi}\right)^2 & \text{if } n{\neq}0 \\ 0 & \text{if } n = 0 \end{cases} \quad (38)$$

$$F_{mn} = \begin{cases} C_{mn}\left(\frac{a}{m\pi}\right)^2\left(\frac{b}{n\pi}\right)^2 & \text{if } m{\neq}0, n{\neq}0 \\ 0 & \text{otherwise} \end{cases} \quad (39)$$

Using the identity

$$\sin(\theta_1)\sin(\theta_2) = \frac{1}{2}(\cos(\theta_1 - \theta_2) - \cos(\theta_1 + \theta_2)) \quad (40)$$

the first summation

$$\sum_{m=0}^{\infty} D_{m0}\left[\sin\left(\frac{m\pi a_2}{a}\right) - \sin\left(\frac{m\pi a_1}{a}\right)\right]\left[\sin\left(\frac{m\pi a_4}{a}\right) - \sin\left(\frac{m\pi a_3}{a}\right)\right] \quad (41)$$

can be re-written as a sum of eight terms in the form

$$\pm\frac{1}{2}\sum_{m=0}^{\infty} D_{m0}\cos\left(\frac{m\pi(a_i{\pm}a_j)}{a}\right) \quad (42)$$

where $i = 1, 2$ and $j = 3, 4$.

To utilize the DCT, we first discretize the top surface of the chip into M equal divisions along the x direction and N equal divisions along the y direction and put the resulting grid on the surface. Then we truncate the summation in equation (42) at index M. If we assume that all the vertices of the field and source regions are located on the grid points, i.e., $\frac{a_i}{a} = \frac{k_i}{M}$, $\frac{a_j}{a} = \frac{k_j}{M}$, where k_i and k_j are integers, and $0{\leq}k_i{\leq}M$, $0{\leq}k_j{\leq}M$, then equation (42) becomes

$$\pm\frac{1}{2}\sum_{m=0}^{M} D_{m0}\cos\left(\frac{m\pi(k_i{\pm}k_j)}{M}\right) \quad (43)$$

Let

$$k = \begin{cases} k_i{\pm}k_j & \text{if } 0{\leq}k_i{\pm}k_j{\leq}M \\ -(k_i{\pm}k_j) & \text{if } k_i{\pm}k_j < 0 \\ 2M - (k_i{\pm}k_j) & \text{if } k_i{\pm}k_j > M \end{cases} \quad (44)$$

then $0{\leq}k{\leq}M$ and equation (43) can be re-written as

$$\pm\frac{1}{2}\sum_{m=0}^{M} D_{m0}\cos\left(\frac{m\pi k}{M}\right) \quad (45)$$

This is exactly one term in the type-I DCT of the sequence D_{m0}, and the DCT sequence can be computed efficiently using the fast Fourier transform (FFT) in $O(M\log(M))$ time [14]. After the DCT sequence is obtained, it can be stored in a vector and used many times in future temperature simulations. As a result, the computation of the summation (41) is reduced to eight look-ups in the DCT vector in constant time and then adding up eight real numbers. Similarly, the summation involving E_{0n} in equation (36) can also be obtained efficiently using the DCT and table look-up.

The double summation in equation (36) can be re-written as a sum of 64 terms in the form

$$\pm\frac{1}{4}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_{mn}\cos\left(\frac{m\pi(a_i{\pm}a_j)}{a}\right)\cos\left(\frac{n\pi(b_p{\pm}b_q)}{b}\right) \quad (46)$$

where $i = 1, 2$, $j = 3, 4$, $p = 1, 2$, and $q = 3, 4$. Using a similar approach, equation (46) can be cast into

$$\pm\frac{1}{4}\sum_{m=0}^{M}\sum_{n=0}^{N} F_{mn}\cos\left(\frac{m\pi k}{M}\right)\cos\left(\frac{n\pi l}{N}\right) \quad (47)$$

where $0{\leq}k{\leq}M$ and $0{\leq}l{\leq}N$. This is one term in the 2-D type-I DCT of the matrix F_{mn}. The 2-D DCT matrix can be computed using the FFT in $O(M{\cdot}N{\cdot}max\{\log(M), \log(N)\})$ time, and after the 2-D DCT table is obtained, the double summation reduces to 64 table look-ups in constant time and then adding up 64 real numbers.

Note that for a general Manhattan-style layout, the power sources can always be divided into small rectangular regions where the computational techniques described here can be applied.

D. Selection of the discretization parameters M and N

The selection of the discretization parameters M and N deserves some more considerations. Assume the minimum feature size along the x and y directions that need to be resolved are x_{min} and y_{min} respectively, then M and N must satisfy

$$M{\geq}\frac{a}{x_{min}} \quad \text{and} \quad N{\geq}\frac{b}{y_{min}} \quad (48)$$

However, since M and N are also the truncation points of the summations in equation (36), they must be large enough to ensure the convergence of the summations. As pointed out in [15], the summations converge more slowly as x_{min} and y_{min} become smaller relative to the chip dimensions a and b. Thus, the actual values of M and N may be chosen to be larger than the lower bounds given in (48) for the convergence purpose. In addition, to assist the utilization of the FFT in the DCT computation, M and N are usually chosen to be powers of 2.

E. Time and storage complexity

As shown in the previous analysis, the overall algorithm is divided into two steps, i.e.,

1) DCT table and vector computation
2) Temperature field evaluation

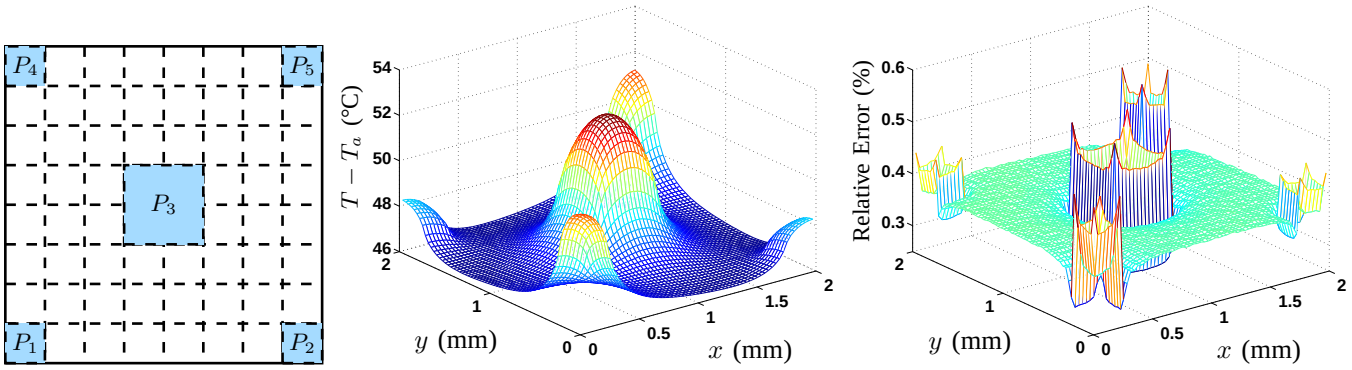
Using the FFT, the computation of the two DCT vectors involving M and N cost $O(M\log(M))$ and $O(N\log(N))$ time respectively, and the computation of the 2-D DCT table cost $O(M{\cdot}N{\cdot}max\{\log(M), \log(N)\})$ time. In the temperature evaluation step, the computation of the temperature rise in a field region due to the uniform power distribution in a source region involves total of 80 look-ups in the DCT table and vectors and the summation of the resulting numbers. As a comparison, the direct computation of equation (36) with the same accuracy without using the DCT and table look-up approach has a cost of $O(M{\cdot}N)$ and typically $M{\cdot}N$ is much larger than 80 [9]. We emphasize that since the DCT table and vectors only depend on the physical properties and the geometry of the chip but are independent of the locations of the field and source regions, they only need to be computed once for each fabrication technology and chip geometry. This fact makes the method presented in this paper very suitable for the optimization process in physical design such as the thermal-aware placement [3] where a large number of different heat source configurations are compared. Since the same DCT table and vectors can be used many times in temperature field evaluations, the amortized cost of obtaining the DCT table and vectors are extremely small and can usually be ignored in real optimization problems. The storage complexity of the DCT table and vectors are $O(M{\cdot}N)$, $O(M)$, and $O(N)$, respectively.

IV. Experimental results

A. Accuracy of the proposed algorithm

The first experiment is to demonstrate the accuracy of the proposed algorithm. Fig. 3(a) shows the top surface of a chip with dimensions of 2mm×2mm×0.5mm. The area is divided into 8×8 equal square sections and five power sources are placed in the corresponding sections as shown in the figure. The thermal conductivity k of the chip material is set to 148W/(m·°C) and the effective heat transfer coefficient h of the bottom surface of the chip is chosen to be 8700W/(m^2·°C), which is consistent with the value used in [13]. The strength of the five power sources are $(P_1, P_2, P_3, P_4, P_5)$ = (0.2W, 0.1W, 1W, 0.1W, 0.2W).

Fig. 3(b) shows the top surface temperature map obtained using the algorithm proposed in this paper, where $T - T_a$ is the temperature rise above the ambient. In obtaining the temperature map, the top surface of the chip was divided into 64×64 small square regions with equal size and the average temperature in each small square region was computed. The parameters M and N were both set to 64, the minimum required values from resolution considerations, because the convergence of the Green function has already been achieved with $M = N = 64$. Fig. 3(c) shows the relative error in the temperature map compared with the computation result obtained from a commercial CFD software for thermal analysis. We can see clearly that the error is below 1%, which demonstrates the accuracy of our method.



(a) (b) (c)

Fig. 3. Accuracy of the proposed algorithm (a) power source locations (b) computed temperature distribution above T_a using the proposed algorithm (c) relative error of the proposed algorithm compared with the result from a commercial CFD software.

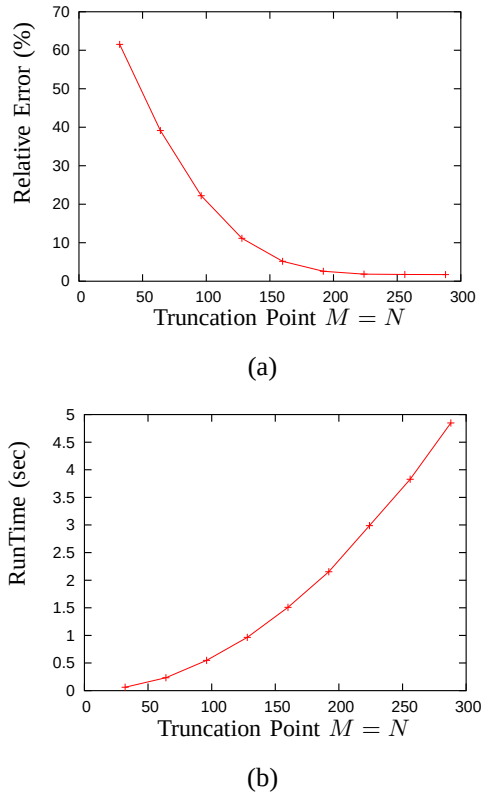
B. Efficiency of the proposed algorithm

The second experiment is to demonstrate the efficiency of the proposed algorithm compared with the direct application of the Green function method to compute the temperature distribution. We still use the same chip dimensions and physical properties as in the previous subsection. However, only one power source is used this time to make the presentation clearer. The power source occupies a square region with dimensions of $\frac{2}{128}\text{mm} \times \frac{2}{128}\text{mm}$ at the exact center of the chip. The strength of the power source is $P_s = 50\text{mW}$. The average temperature above T_a of the source region itself is computed. The parameters M and N are both chosen to be 512 in our algorithm from convergence considerations for the Green function, i.e., we require the truncation error to be within 1%. The infinite summations in the Green function is more difficult to converge in this example because the sizes of the source and field regions relative to the chip dimensions are smaller than those in the previous example.

Using the proposed algorithm, the average temperature of the source region itself above T_a is found to be 11.537°C. The total computation time of the DCT matrix and vectors using MATLAB is 1.39sec. Using the pre-calculated DCT coefficients for this technology and chip geometry, the evaluation of the average temperature only takes 2.25*m*sec.

As a comparison, we also computed the average temperature above T_a of the source region using equation (36) directly, which corresponds to the direct application of the Green function method. In the direct method, it is unnecessary to consider the resolution issue because equation (36) does not require the vertices (a_i, b_i) of the source and field regions to coincide with some grid points. So the parameters M and N are completely determined by the convergence consideration. Since the chip is square, we set $M = N$ in our analysis.

Fig. 4 shows the relative error and the corresponding runtime of the direct method. We can observe from the figure that even for a 5% relative error in $T - T_a$, the truncation point must be higher than 160. The runtime at this truncation point is 1.508sec, which is more than 600 times slower than our algorithm, and the accuracy of our algorithm is much higher.



(a)

(b)

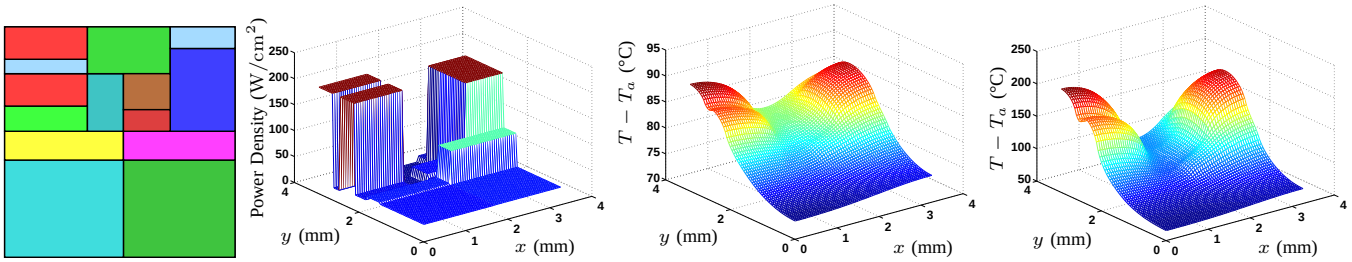
Fig. 4. Accuracy and computation time of the direct application of the Green function method (a) relative error in $T - T_a$ versus truncation point (b) runtime versus truncation point.

C. A realistic example

Fig. 5(a) shows the floorplan from [16], which is similar to that of the DEC Alpha 21264 processor but is scaled from the 350nm to the 65nm technology. The scaled chip dimensions are $3.3\text{mm} \times 3.3\text{mm} \times 0.5\text{mm}$, and we assume that the chip has the same physical properties as those used in the previous examples. Fig. 5(b) shows the power density distribution of the modules in W/cm^2. We divided the top surface of the chip into 64×64 small square regions with equal size and computed the temperature map, which is shown in Fig. 5(c). From the figure, we can see clearly that the temperature map is much smoother than the power density map, which is due to the relatively high thermal conductivity of the silicon substrate and the horizontal heat transfer [17]. To demonstrate this, we show in Fig. 5(d) the surface temperature map on quartz substrate when the same power density distribution is applied. Since quartz has a thermal conductivity of only 10.4W/(m·°C), which is much lower than that of silicon, the resulting temperature map tracks the power density map more closely. We point out that when the quartz substrate is used, the top surface of the chip can no longer be considered as adiabatic because of the low thermal conductivity of the substrate material. Fig. 5(d) is only provided to prove the validity of Fig. 5(c).

V. Extension to the Multilayered Substrate and Packaging Structures

The analysis in the previous sections are based on the assumption that the solder, packaging, and ambient can be characterized by a combined effective heat transfer coefficient h. This is a technique typically used in fast temperature estimations. If a more accurate temperature computation is required, the chip, solder, and packaging must be treated as a multilayered structure. It is very easy to extend the method presented in this paper



(a) (b) (c) (d)

Fig. 5. Power and temperature distribution of a realistic chip (a) floorplan (b) power distribution (c) temperature distribution (d) temperature distribution on quartz substrate.

to handle the rectangular shaped multilayers. Compared with the single-layered situation (equations (6)-(9)), equations (6) and (7) will be satisfied in each individual layer if a multilayered structure is studied. Equation (9) will be satisfied on the top surface of the chip and equation (8) becomes

$$k_p \frac{\partial T'(x,y,z)}{\partial z}\bigg|_{z=-d_p} = h_a T'(x,y,z)|_{z=-d_p} \tag{49}$$

where k_p is the thermal conductivity of the packaging and h_a is the heat transfer coefficient of the ambient. In addition, $T'(x,y,z)$ and the heat flow must be continuous at the boundary between two adjacent layers, which results in four more boundary conditions

$$T'(x,y,z)|_{z=-d+\epsilon} = T'(x,y,z)|_{z=-d-\epsilon} \tag{50}$$

$$T'(x,y,z)|_{z=-d_s+\epsilon} = T'(x,y,z)|_{z=-d_s-\epsilon} \tag{51}$$

$$k \frac{\partial T'(x,y,z)}{\partial z}\bigg|_{z=-d+\epsilon} = k_s \frac{\partial T'(x,y,z)}{\partial z}\bigg|_{z=-d-\epsilon} \tag{52}$$

$$k_s \frac{\partial T'(x,y,z)}{\partial z}\bigg|_{z=-d_s+\epsilon} = k_p \frac{\partial T'(x,y,z)}{\partial z}\bigg|_{z=-d_s-\epsilon} \tag{53}$$

where k, k_s, and k_p are the thermal conductivities of the chip, solder, and packaging respectively, and ϵ is an infinitely small quantity. The following steps of Green function analysis, building up the DCT table and vectors, and computing the temperature distributions are all very similar to the single-layered case except that each individual layer now has a different set of coefficients in the infinite series representation of the Green function. Interested readers are referred to the literature for the Green function of multilayered structures [9] [10].

VI. Conclusions

In this paper, we combined the Green function method with the DCT and table look-up techniques to accurately and efficiently compute the surface temperature distribution on VLSI chips due to surface power sources. The conventional Green function method used in temperature computations has the drawback of slow convergence in the infinite series representation of the temperature, hence, causing high computational cost in real implementations. By discretizing the chip surface and establishing the 2-D DCT table and 1-D DCT vectors, we were able to reduce the infinite summation to 64 look-ups in the DCT table and 16 look-ups in the DCT vectors. The following summation of the look-up results can be accomplished in constant time. Experimental results show that our method has a relative error of below 1% compared with commercial CFD softwares for thermal analysis and the runtime of our algorithm is orders of magnitude smaller than that of the conventional Green function method. The method demonstrated in this paper is very suitable for optimization problems such as the thermal-aware placement where a large number of different temperature distributions have to be computed. Since the DCT table and vectors only need to be computed once for each technology and chip geometry but are independent of the exact placement of cells, the amortized cost of establishing the DCT table and vectors is negligible. Finally, our method can be easily extended to simulate the temperature distributions when the multilayered substrate and packaging structures are considered, where a more accurate result can be obtained.

References

[1] D. Chen, E. Li, E. Rosenbaum, and S. M. Kang, "Interconnect Thermal Modeling for Accurate Simulation of Circuit Timing and Reliability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 2, pp. 197-205, Feb. 2000.

[2] S. Rzepka, K. Banerjee, E. Meusel, and C. Hu, "Characterization of Self-heating in Advanced VLSI Interconnect Lines Based on Thermal Finite Element Simulation," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, vol. 21, no. 3, pp. 406-411, Sept. 1998.

[3] 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.

[4] 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.

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

[6] A. Haji-Sheikh, "Peak Temperature in High-Power Chips," *IEEE Transactions on Electron Devices*, vol. 37, no. 4, pp. 902-907, Apr. 1990.

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

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

[9] R. Gharpurey and R. G. Meyer, "Modeling and Analysis of Substrate Coupling in Integrated Circuits," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 3, pp. 344-353, Mar. 1996.

[10] A. M. Niknejad, R. Gharpurey, and R. G. Meyer, "Numerically Stable Green Function for Modeling and Analysis of Substrate Coupling in Integrated Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 4, pp. 305-315, Apr. 1998.

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

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

[13] 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.

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

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

[16] W. Liao, L. He, and K. Lepak, "Temperature-Aware Performance and Power Modeling," Technical Report UCLA Engr. 04-250, UCLA, CA, 2004.

[17] K. Skadron, M. R. Stan, W. Huang, and S. Velusamy, " Temperature-Aware Microarchitecture," *Proceedings of the 30th International Symposium on Computer Architecture*, pp. 2-13, Jun. 2003.