Accurate Thermal Analysis Considering Nonlinear Thermal Conductivity

Anand Ramalingam[1] Frank Liu[2] Sani R. Nassif[2] David Z. Pan[1]
[1]Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712
[2]Austin Research Laboratory, IBM Research Division, Austin, TX 78758
{*anandram, dpan*}@*cerc.utexas.edu* and {*frankliu, nassif*}@*us.ibm.com*

Abstract

The increase in packing density has led to a higher power density in the chip which in turn has led to an increase in temperature on the chip. Temperature affects reliability, performance and power directly, motivating the need to accurately simulate the thermal profile of a chip. In literature, thermal conductivity is assumed to be a constant in order to obtain a linear system of equations which can be solved efficiently. But thermal conductivity is a nonlinear *function of temperature and for silicon it varies by* 22% *over the range* $27-80^{\circ}$ *C [1]. If the nonlinearity of the thermal conductivity is ignored the thermal profile might be off by* 10° *C. Thus to get an accurate thermal profile it is important to consider the nonlinear dependence of the thermal conductivity on temperature. In this paper the nonlinear system arising out of considering the nonlinear thermal conductivity is solved efficiently using a variant of Newton-Raphson. In this paper we also study the abstraction levels under which the approximation of a periodic source by a DC source is valid.*

1. Introduction

To achieve higher performance, technology is scaled aggressively. This leads to greater packing density which leads to higher power density in the chip. The increase in power density leads to increase in temperature on the chip. The increase in temperature leads to many challenges,

1. Reliability of transistors is exponentially dependent on the operating temperature of the junction. A difference of $10-15^{\circ}$ C can result in nearly $2\times$ difference in the lifespan of the devices [2].
2. At higher temperatures the electron mobility decreases due to the phonon scattering effect. The decrease in electron mobility leads to increased gate delay thus a lower clock frequency.
3. The leakage has super-linear dependency on the temperature. A difference of 30° C will affect the leakage by 30% [3].

Thus the temperature affects reliability, performance and power directly. Hence it is important to accurately simulate the thermal profile of a chip. Although the electrothermal simulation for analog circuits [4] is a well studied problem, those techniques are not applicable for simulating the thermal profile of a VLSI chip. This is because these electrothermal analysis are done in a SPICE like manner. Hence these techniques do not scale well making them unsuitable for the thermal simulation of a VLSI chip. Thus for the thermal analysis of VLSI circuits new methods have been proposed in the literature.

The simulation of the thermal profile essentially involves solving the heat equation. The heat equation is a partial differential equation (PDE) with boundary conditions which on discretization leads to a system of equations. The system of equations is *linear* if the thermal conductivity of the chip layers is assumed to be a constant. The early thermal simulators for the VLSI chips solved the linear system directly [5]. With the increase in package density, the number of variables to solve in the linear system increased. Model order reduction was used to reduce the system of equations to solve and increase computational efficiency [6]. To increase the efficiency of the transient solve, the ADI method was used [7]. The linear system resulting from discretizing the heat equation is a symmetric positive definite matrix, called the M-matrix. Multigrid technique can be used to accelerate solving the M-matrix and has been used for efficient full-chip thermal analysis [8].

The thermal simulators have provided the technology on which many thermal based applications spawned. Applications have been proposed in the areas of leakage analysis [3], reliability [9] among many others.

The *efficiency* of the solutions proposed in literature is due to the assumption that the thermal conductiv-

Proceedings of the 7th International Symposium on Quality Electronic Design (ISQED'06)
0-7695-2523-7/06 $20.00 © 2006 IEEE

ity is a constant which leads to a linear system of equations. But the thermal conductivity is a *nonlinear* function of temperature and for silicon it varies by 22% over the range $27-80^\circ$ C [1]. Thus ignoring the nonlinearity of the thermal conductivity might lead to a difference of temperature by 10° C as shown in Section 4. Thus to get an accurate thermal profile it is important to consider the nonlinear dependence of the thermal conductivity on temperature. This paper addresses the challenge of solving the nonlinear system of equations efficiently. A fast algorithm which is an variant of Newton-Raphson is proposed to solve the nonlinear system of equations.

This paper also studies the validity of approximating a periodic thermal source by a DC source. The value of the DC source is the root mean square (RMS) value of the periodic source. This paper provides the abstraction levels (for example, gate level or block level) under which this approximation is valid.

This paper makes the following contributions

- The nonlinear thermal conductivity is taken into account during thermal simulation.
- A fast algorithm is proposed to solve the nonlinear system of equations $A(x)x = b$.
- The abstraction levels under which the approximation of a periodic source by a DC source is valid.

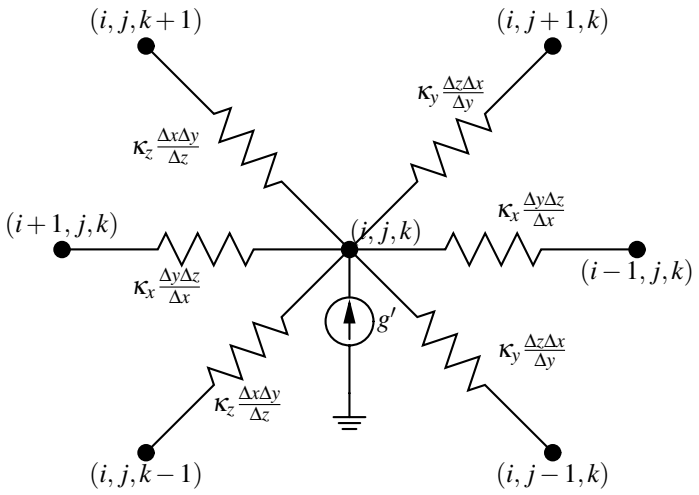
The paper is organized as follows. The mathematical model for nonlinear thermal simulation is presented in Section 2. A fast algorithm to solve the nonlinear system of equations is presented next in Section 3. The experimental results are presented in Section 4 and the validity of modeling the periodic source by a DC source is studied in Section 5 and we conclude in Section 6.

2. Thermal Modeling and Temperature Simulation

A general 3D thermal analysis involves solving the heat conduction equation which is discussed in detail in [3]. On discretizing the heat condition equation using finite differences and assuming steady-state conditions, the heat equation can be written as

$$\begin{aligned}\kappa_x(T_{i,j,k})\frac{\Delta y\Delta z}{\Delta x}\left(2T_{i,j,k}-T_{i-1,j,k}-T_{i+1,j,k}\right)\\ +\kappa_y(T_{i,j,k})\frac{\Delta z\Delta x}{\Delta y}\left(2T_{i,j,k}-T_{i,j-1,k}-T_{i,j+1,k}\right)\\ +\kappa_z(T_{i,j,k})\frac{\Delta x\Delta y}{\Delta z}\left(2T_{i,j,k}-T_{i,j,k-1}-T_{i,j,k+1}\right)\\ =\Delta x\Delta y\Delta z\times g(x,y,z)\end{aligned}\quad(1)$$

where $\kappa_x(T_{i,j,k})\frac{\Delta y\Delta z}{\Delta x}$ is interpreted as the thermal conductance in the x direction. The electrical interpretation of Eq. (1) is shown in Figure 1.



$$g' = \Delta x\Delta y\Delta z\times g(x,y,z)$$

Figure 1. Electrical interpretation of the 3-d heat equation. Note that the thermal conductivities k_x, k_y and k_z are functions of temperature $T_{i,j,k}$

Generalizing Eq. (1) leads to a matrix formulation

$$\mathbf{K}(\mathbf{T})\mathbf{T} = \mathbf{g} \quad (2)$$

similar to the circuit relation $\mathbf{GV} = \mathbf{i}$, thus capturing the analogy of voltage with temperature and current with heat sources. The challenge lies in solving this nonlinear system of equations efficiently. A fast algorithm to solve this nonlinear system is proposed in the next section.

3. A Fast Algorithm to solve $\mathbf{A}(\mathbf{x})\mathbf{x} = \mathbf{b}$

The nonlinear system of equations, $\mathbf{A}(\mathbf{x})\mathbf{x} = \mathbf{b}$, needs to be solved efficiently to make the problem of considering nonlinear thermal conductivity practical.

The widely used iterative solver for the nonlinear system of equations is the Newton-Raphson method. The Newton-Raphson iteration can be expressed as

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - [\mathbf{J}(\mathbf{x}^{(k-1)})]^{-1}\mathbf{f}(\mathbf{x}^{(k-1)}) \quad (3)$$

where

$$\mathbf{J}(\mathbf{x}) = \frac{\partial(\mathbf{A}(\mathbf{x})\mathbf{x}-\mathbf{b})}{\partial\mathbf{x}}$$
$$\mathbf{f}(\mathbf{x}) = \mathbf{A}_{\mathbf{i}}(\mathbf{x})\mathbf{x}-\mathbf{b}$$

and $\mathbf{A_{i.}}(\mathbf{x})$ denotes the row i of the matrix $\mathbf{A}$.

In each iteration of the Newton-Raphson in Eq. (3), the Jacobian needs to be inverted. In Figure 2, note that the tangent is evaluated each iteration. This tangent evaluation is equivalent to finding the inverse of a Jacobian matrix ($\mathbf{J} \in \mathbb{R}^{m \times m}$) in each iteration when solving the thermal circuit.

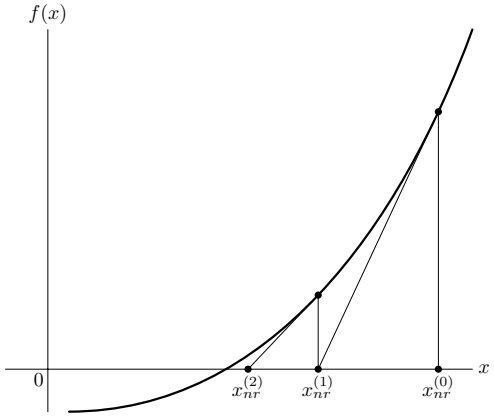


Figure 2. Newton-Raphson iteration. Note that the tangent is evaluated at every iteration. This is equivalent to finding the inverse of a Jacobian matrix ($\mathbf{J} \in \mathbb{R}^{m \times m}$) for every iteration when solving the thermal circuit.

The evaluation of an inverse has an asymptotic complexity $O(n^3)$. This is a time consuming operation considering the size of matrices is in the order of tens of thousands in thermal simulation. This motivates the need for a variant of the Newton-Raphson algorithm. Since the evaluation of the Jacobian inverse is the bottleneck in Newton-Raphson, the Jacobian inverse is evaluated once during the iteration and it is used in every iteration thereof. The Newton-Raphson iteration with a constant Jacobian inverse can be expressed as

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - [\mathbf{J}(\mathbf{x}^{(0)})]^{-1}\mathbf{f}(\mathbf{x}^{(k-1)}) \tag{4}$$

where $\mathbf{J}(\mathbf{x}^{(0)})$ is the Jacobian evaluated during the initial guess. The iteration in Eq. (4) is called the *constant Jacobian*. This will work if the initial guess is close to the final solution. If the initial guess is random, the solver might throw nonphysical temperatures or worse might not converge. Since the temperatures on the chip cannot go lower than the room temperature and usually in the range $[300, 400]$ K, the initial guess is set to the room temperature 300 K. In constant Jacobian, the slope of the tangent is the same for every iteration which leads to slower convergence. To accelerate the constant Jacobian, an approximation to the Jacobian named *reduced order* Jacobian is proposed and its algorithmic details is described next.

3.1. Evaluating the reduced order Jacobian

Let $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k-1)}$ be the first k iterations when solving the equation $\mathbf{A}(\mathbf{x})\mathbf{x} = \mathbf{b}$. Then $\mathbf{x}^{(k)}$ can be approximately calculated by fitting a plane through the vectors $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k-1)}$. The plane fitting is described next.

Let the vector $\mathbf{x}^{(j)}$ be divided into p partitions $\mathbf{x}_i^{(j)}, i = 0, \ldots, p-1$, where $k = p+1$. This procedure is repeated for all the vectors $\mathbf{x}^{(j)}$, $j = 0, \ldots, k-1$. Let the norm error be defined as $\varepsilon = \|\mathbf{Ax} - \mathbf{b}\|$. The error $\varepsilon_i^{(j)}$ can be thought of error in the partition i during the iteration j. The idea is to fit a plane through the vectors $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k-1)}$, such that the error in the $k+1$ iteration $\varepsilon^{(k)}$ goes to 0. Thus the following system of linear equations needs to be solved

$$\begin{pmatrix} \varepsilon_0^{(0)} & \cdots & \varepsilon_0^{(k-1)} \\ \cdots & \cdots & \cdots \\ \varepsilon_{p-1}^{(0)} & \cdots & \varepsilon_{p-1}^{(k-1)} \end{pmatrix} \quad \begin{pmatrix} \alpha^{(0)} \\ \cdots \\ \alpha^{(k-1)} \end{pmatrix} = \begin{pmatrix} 0 \\ . \\ 0 \end{pmatrix} \tag{5}$$

The solution looks trivial at the first look. But

$$\sum_{j=0}^{k-1} \alpha^{(j)} = 1 \tag{6}$$

This implies the RHS is non-zero and a non-trivial solution exists. Once $\alpha^{(j)}$ are determined, the $\mathbf{x}^{(k)}$ can be found using

$$\mathbf{x}^{(k)} = \alpha^{(0)}\mathbf{x}^{(0)} + \ldots + \alpha^{(k-1)}\mathbf{x}^{(k-1)} \tag{7}$$

The above equation can be thought of as *reduced order* Jacobian and can be rewritten as a recursion

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \underbrace{\begin{bmatrix} \mathbf{x}^{(k-1)} - \mathbf{x}^{(0)} \\ \cdots \\ \mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)} \end{bmatrix}^{\top} \begin{pmatrix} \varepsilon_0^{(k-1)} - \varepsilon_0^{(0)} & \cdots & \varepsilon_0^{(k-1)} - \varepsilon_0^{(k-2)} \\ \cdots & \cdots & \cdots \\ \varepsilon_{p-1}^{(k-1)} - \varepsilon_{p-1}^{(0)} & \cdots & \varepsilon_{p-1}^{(k-1)} - \varepsilon_{p-1}^{(k-2)} \end{pmatrix}^{-1}}_{\textit{reduced order} \text{ Jacobian}} \times \begin{pmatrix} \varepsilon_0^{(k-1)} \\ \cdots \\ \varepsilon_{p-1}^{(k-1)} \end{pmatrix} \tag{8}$$

The intuition for the *reduced order* Jacobian can be got by considering the number of partitions, $p = 1$. Thus atleast $k = p + 1 = 2$ iterations are needed before *reduced order* Jacobian algorithm can be applied. Recall that the norm error for iteration j is defined as

$\varepsilon^{(j)} = \|\mathbf{A}\mathbf{x}^{(j)} - \mathbf{b}\|$. The system of equations in Eq. (5) reduces to

$$\alpha^{(0)}\varepsilon_0^{(0)} + \alpha^{(1)}\varepsilon_0^{(1)} = 0$$
$$\alpha^{(0)}\varepsilon_0^{(0)} + (1-\alpha^{(0)})\varepsilon_0^{(1)} = 0 \qquad \text{using Eq. (6)}$$
$$\alpha^{(0)} = \frac{-\varepsilon_0^{(1)}}{\varepsilon_0^{(0)} - \varepsilon_0^{(1)}}$$

Thus $\mathbf{x}^{(2)}$ can be predicted as,

$$\mathbf{x}^{(2)} = \alpha^{(0)}\mathbf{x}^{(0)} + \alpha^{(1)}\mathbf{x}^{(1)}$$
$$= \alpha^{(0)}\mathbf{x}^{(0)} + (1-\alpha^{(0)})\mathbf{x}^{(1)}$$

The graphical illustration of the algorithm is shown next by considering a 1-dimensional ($m = 1$) root finding. Since the number of equations to be solved is one, the number of partitions is trivially one ($p = 1$).

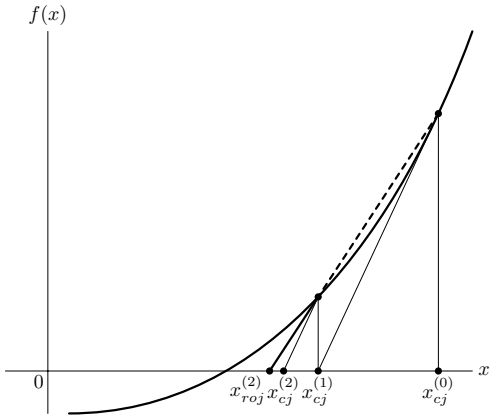


Figure 3. Constant Jacobian with speedup. Please observe that $x_{roj}^{(2)}$ is got by fitting a line through the previous two iterations. Also note that the $x_{roj}^{(2)}$ is closer to the root than $x_{cj}^{(2)}$ thus accelerating the constant Jacobian.

The speedup of the constant Jacobian is illustrated in Figure 3. Note that $x_{roj}^{(2)}$ is got by fitting a line through the previous two iterations. Also note that the $x_{roj}^{(2)}$ is closer to the root than $x_{cj}^{(2)}$ thus accelerating the constant Jacobian.

The *reduced order* Jacobian is applied every q iterations to accelerate the constant Jacobian and q is an integer usually between 2 to 5. The complete pseudocode is shown in Algorithm 1.

4. Experimental Results

The effect of nonlinear thermal conductivity was tested by considering the silicon layer of the chip. The simplified version has the silicon layer's boundaries at the room temperature 27° C. The dimension of the chip is $8mm \times 8mm$. For simplicity assume that the entire silicon layer dissipates 100 W uniformly. For comparison, the difference in temperatures obtained by assuming constant thermal conductivity and nonlinear thermal conductivity is studied.

In Figure 4, the constant thermal conductivity evaluated at 27° C, underestimates the peak temperature by 12% when compared to the thermal profile obtained by considering the nonlinear thermal conductivity. Similarly, if the constant thermal conductivity is evaluated at 127° C, the peak temperature is overestimated by 13%. This inaccuracy in the thermal profile demonstrates the need for considering the nonlinearity of the thermal conductivity while doing the thermal simulation.

Algorithm 1 Accelerated-Constant-Jacobian

Input: A system $\mathbf{F}(\mathbf{x}) \triangleq \mathbf{A}(\mathbf{x})\mathbf{x} - \mathbf{b}$ consisting of m nonlinear equations in m unknowns:

$$f_i(x_0, x_1, \ldots, x_{m-1}) = 0, \quad i = 0, 1, \ldots, m-1$$

Input: An initial guess $\mathbf{x}^{(0)} = (x_0^{(0)}, x_1^{(0)}, \ldots, x_{m-1}^{(0)})$ of the solution.

Input: p, the number of partitions $\mathbf{A}(\mathbf{x})$ is divided into.

Output: $\mathbf{x}^* = (x_0^*, \ldots, x_{m-1}^*)$ solving m nonlinear equations simultaneously.

```
1:  k ← 0
2:  repeat
3:     // Accelerate using reduced Jacobian
4:     // by using it every qth iteration
5:     // after the first k = p + 1 iterations
6:     if ((k > p) and !(k%q)) then
7:        Find x^(k+1) from the last k(= p + 1) iterations
          using Eq. (8)
8:     else
9:        // do constant Jacobian
10:       x^(k+1) = x^(k) − [J(x^(0))]^−1 f(x^(k))
11:    end if
12:    k ← k + 1
13: until ( ||x^(k) − x^(k−1)|| ≤ ε and ||F(x^(k))|| ≤ ε )
```

5. Validity of approximating a periodic source by a DC source

In the literature, the steady state response is modeled as a DC problem with the periodic sources being approximated by DC sources with the RMS value of the periodic source. In this section, we study under which

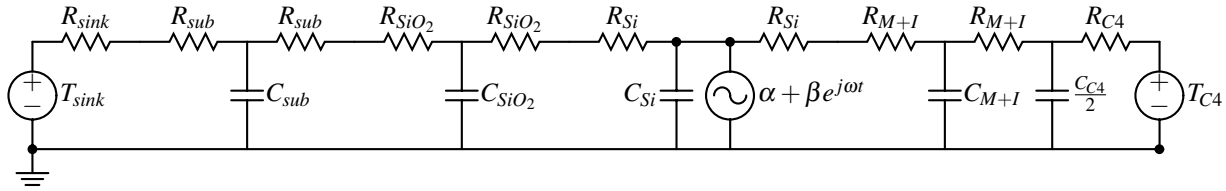


Figure 5. Thermal circuit to study the validity of approximating the periodic sources with DC sources. A simplified circuit is got by assuming uniform heat distribution along x, y directions in every layer. Hence the variation is along the z direction leading to a 1-d circuit which is used to study the problem. The switching of transistors is modeled as a sinusoidal heat source with frequency ω. In literature, instead of the sinusoidal source, a DC source is used with the amplitude equal to that of the RMS value of the sinusoidal source.

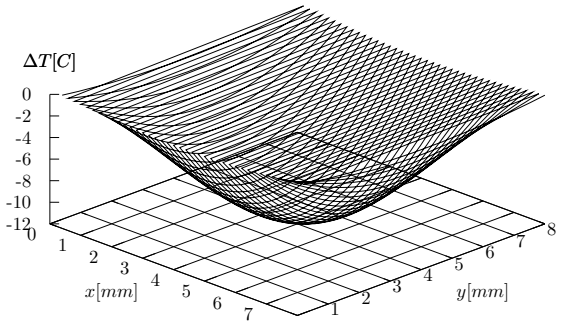


Figure 4. The difference in temperature profile in a silicon layer between having a constant thermal conductivity for silicon and incorporating nonlinear thermal conductivity for silicon. The chip dimension is $8mm \times 8mm$ and it dissipates 100 W uniformly. The constant thermal conductivity evaluated at 27° C and used in the thermal simulation underestimates the peak temperature by 12%. This is $\approx 12^\circ$ C in absolute value.

abstraction level this approximation is valid. First we study the problem at the transistor/gate level. The thermal model to study this problem is shown in Figure 5.

The techniques that were applied to simplify the model are described below:

1. A simplified thermal circuit is got by assuming uniform heat distribution along x, y directions in every layer. Hence the variation is along the z direction leading to a 1-d circuit which is used to study the problem.

2. The metal and ILD (inter-layer dielectric) layer are modeled using equivalent thermal resistance modeling. The equivalent thermal resistance in the metal and the ILD are adjusted according to the metal density of layers and according to the via density between adjacent metal layers [3].

3. The switching of the transistors is modeled as a sinusoidal heat source with the frequency equal to the clock frequency ω of the chip.

The following equations describe the resistance and capacitance along the z direction.

$$R_{\text{layer}} = \frac{\Delta z}{k_{\text{layer}} \times \Delta x \Delta y}$$
$$C_{\text{layer}} = \rho \times c_\rho \times \Delta x \Delta y \Delta z$$

To further simplify assume, $\Delta x = \Delta y = 1\mu m$. Since the heat generation is uniform along x, y directions the above assumption is justified. In Figure 6 steady state response of the circuit at the substrate is shown. The steady state values are zero due to the fact that the substrate has a high time constant $\tau = 4.4 \times 10^{-4} s$ compared to the clock frequency $\omega = 1$ Ghz which attenuates the temperature. This fact is brought out clearly in the frequency response plot in Figure 7. Note that the substrate response tails off near 1 Khz and starts attenuating around 1 Mhz due to the substrate time constant being $\tau = 4.4 \times 10^{-4} s$. The substrate time constant dominates the frequency response of the thermal circuit. The temperature rise at the substrate calculated using the RMS values for the ac sources was found to be 2.65° C. Thus using the RMS values instead of the steady state response does not cause a high estimation error. But this would change when the level of abstraction increases from transistors to blocks. As evident from Figure 7, the temperature change per unit increase of power dissipated is > 1 at lower frequencies. In blocks, the switching frequency might be much lower than the clock frequency due to decreased switching activity at a macroscopic level. Hence steady-state

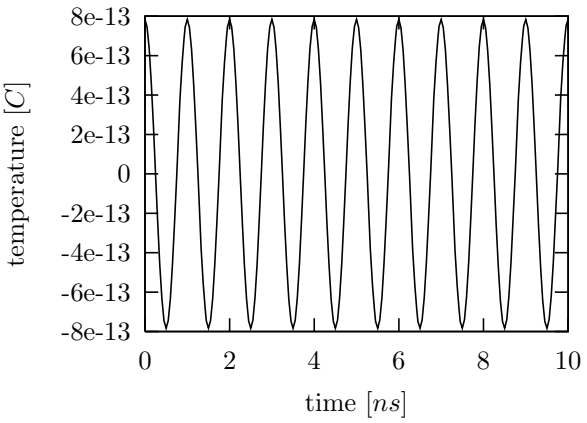


Figure 6. Steady state response at the substrate (the node between the resistors named R_{sub} in Figure 5). The temperature rise due to periodic source is *near* zero. This is due to the fact that the substrate has a high time constant (See Figure 7). The temperature rise at the substrate calculated using the RMS values for the ac sources is 2.65° C.

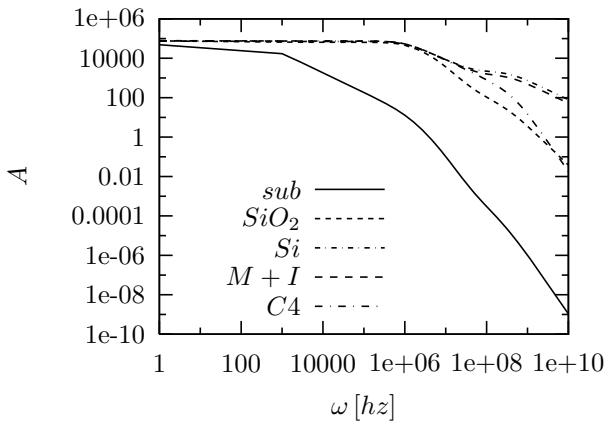


Figure 7. Frequency response of the thermal circuit in Figure 5. The y-axis represents the temperature change per unit increase of power dissipated ($A = \frac{\Delta T}{\beta}$). The substrate response tails off ≈ 1 Khz. This is because time constant for substrate is $4.4 \times 10^{-4}\, s$ Also this dominates the frequency response of thermal circuit.

response at block level would be attenuated much less than at a transistor level thus leading a significant estimation error if the RMS values are used.

6. Conclusion

In this paper we studied the effects of ignoring the nonlinearity of thermal conductivity. We have shown that the ignoring thermal conductivity might lead to temperature profile that is off by 10° C and might cause inaccurate results in reliability analysis. We also studied the abstraction levels under which the approximation of the steady state sources by DC sources with the RMS value is valid. We concluded that at a transistor level this is a valid approximation while it may not be valid at a block level.

Acknowledgment

Anand Ramalingam thanks IBM Austin Research Laboratory for providing an opportunity to intern in Summer 2005. Anand Ramalingam also thanks Ashish Kumar Singh for his valuable inputs on the algorithm. This work is partially sponsored by IBM Faculty Award.

References

[1] A. D. McConnell, S. Uma, and K. E. Goodson, "Thermal conductivity of doped polysilicon layers," *Journal of Microelectromechanical Systems*, vol. 10, no. 3, pp. 360–369, September 2001.

[2] R. Viswanath, V. Wakharkar, A. Watwe, and V. Lebonheur, "Thermal performance challenges from silicon to systems," *Intel Technology Journal*, no. Q3, Aug. 2000.

[3] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *ISLPED*, 2003, pp. 78–83.

[4] S.-S. Lee and D. J. Allstot, "Electrothermal simulation of integrated circuits," *JSSC*, pp. 1283–1293, 1993.

[5] 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," *TCAD*, pp. 668–681, 1998.

[6] C.-H. Tsai and S.-M. Kang, "Fast temperature calculation for transient electrothermal simulation by mixed frequency/time domain thermal model reduction," in *DAC*, 2000, pp. 750–755.

[7] T.-Y. Wang and C. C.-P. Chen, "Thermal ADI: A linear-time chip level dynamic thermal-simulation algorithm based on alternating-direction-implicit (ADI) method," *TVLSI*, pp. 691–700, 2003.

[8] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, "Efficient full-chip thermal modeling and analysis," in *ICCAD*, 2004, pp. 319–326.

[9] D. Chen, E. Li, E. Rosenbaum, and S.-M. Kang, "Interconnect thermal modeling for accurate simulation of circuit timing and reliability," *TCAD*, pp. 197–205, February 2000.