A Self-Consistent Substrate Thermal Profile Estimation Technique for Nanoscale ICs—Part II: Implementation and Implications for Power Estimation and Thermal Management

Sheng-Chih Lin, *Student Member, IEEE*, Greg Chrysler, *Member, IEEE*, Ravi Mahajan, *Senior Member, IEEE*, Vivek K. De, *Senior Member, IEEE*, and Kaustav Banerjee, *Senior Member, IEEE*

***Abstract*—As transistors continue to evolve along Moore's Law and silicon devices take advantage of this evolution to offer increasing performance, there is a critical need to accurately estimate the silicon-substrate (junction or die) thermal gradients and temperature profile for the development and thermal management of future generations of all high-performance integrated circuits (ICs) including microprocessors. This paper presents an accurate chip-level leakage-aware method that self-consistently incorporates various electrothermal couplings between chip power, junction temperature, operating frequency, and supply voltage for substrate thermal profile estimation and also employs a realistic package thermal model that comprehends different packaging layers and noncubic structure of the package, which are not accounted for in traditional analyses. The evaluation using the proposed methodology is efficient and shows excellent agreements with an industrial-quality computational-fluid-dynamics (CFD) based commercial software. Furthermore, the methodology is shown to become increasingly effective with increase in leakage as technology scales. It is shown that considering electrothermal couplings and realistic package thermal model not only improves the accuracy of estimating the heat distribution across the chip but also has significant implications for precise power estimation and thermal management in nanometer-scale CMOS technologies.**

***Index Terms*—Integrated circuits, leakage, performance, power, temperature gradient, thermal management.**

I. Introduction

HIGHLY integrated circuits, including system-on-chips (SoCs) with different functional blocks, blocks with different activity rates (e.g., logic versus memory), and clock/power-gating techniques, essentially create nonuniform temperature distributions across the chip substrate [1]. Regions with higher temperature are commonly referred to as hot-spots. Hot-spots simultaneously lead to temperature gradients that affect performance (including delay and timing) and reliability among a host of other issues and also result in a general overdesign in the microprocessor packaging and cooling solutions. In fact, previous publications have identified large thermal gradients in high-performance microprocessor chips [2]–[6].

A. Implications of Substrate Temperature Rise and Nonuniform Thermal Profile

High temperature not only leads to the onset and acceleration of reliability problems at the device and interconnect level but also impacts circuit- and system-level metrics (Fig. 1). Elevated temperature deteriorates circuit performance by degrading the device carrier mobility and increasing the interconnect metal resistivity. Nonuniform temperature distribution (thermal gradient) across the chip substrate causes the thermal profile of interconnects to be nonuniform that severely affects the integrity of the clock signal and interconnect performance [7], [8]. In addition, buffer-insertion and gate-sizing schemes (in the physical design process) are influenced by thermal gradients because interconnect and gate delays are strongly dependent on temperature [9]. Moreover, nonuniform temperature distribution impacts placement and routing algorithms that are employed to ensure acceptable voltage-drop levels [10]–[13]. At the system level, thermal-management (packaging and cooling) solutions are also affected by the substrate temperature profile because they have to meet the maximum heat-flux requirements at the silicon-package interface [14], [15] (Fig. 1).

B. Measurement and Modeling of Substrate Thermal Profile: Prior Work

As elevated and nonuniform substrate temperature extensively impacts the chip reliability, performance, and thermal management, acquiring accurate thermal profiles is necessary in the early design stage (before the chip is fabricated).

While thermal infrared (IR) imaging system can be used for acquiring thermal profiles, this system offers a limited resolution of substrate thermal profiles of chips with sophisticated packaging structures. Although a technique using an

Manuscript received October 6, 2006. This work was supported in part by Intel Corporation, by the University of California—MICRO Program (03–004), and by the National Science Foundation under Award CCF-0541465. The review of this paper was arranged by Editor H. S. Momose.

S.-C. Lin and K. Banerjee are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: sclin@ece.ucsb.edu; kaustav@ece.ucsb.edu).

G. Chrysler and R. Mahajan are with the Assembly and Test Technology Development, Intel Corporation, Chandler, AZ 85226 USA.

V. K. De is with the Circuit Research Laboratory, Intel Corporation, Hillsboro, OR 97124 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2007.909038

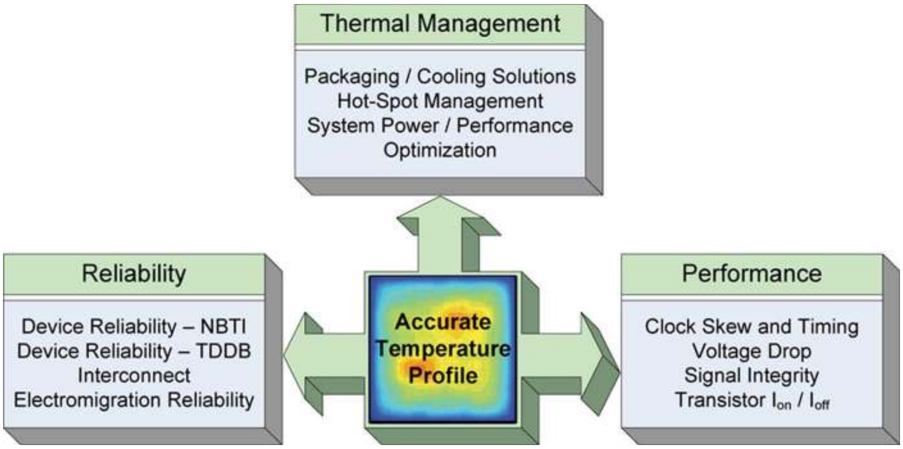

Thermal Management
Packaging / Cooling Solutions
Hot-Spot Management
System Power / Performance Optimization
Reliability
Device Reliability – NBTI
Device Reliability – TDDB
Interconnect
Electromigration Reliability
Accurate Temperature Profile
Performance
Clock Skew and Timing
Voltage Drop
Signal Integrity
Transistor I_on / I_off

Fig. 1. Schematic illustrating the significance and implications of accurate chip-substrate thermal profile estimation on IC reliability issues, performance, and thermal management.

infrared-transparent heatsink has been recently demonstrated to capture the thermal profile by IR imaging [3], [4], it can only be used after chip fabrication. Moreover, such techniques may only be available to some select researchers. Similarly, integrated thermal sensors are commonly employed to ensure that hot-spots do not exceed the specified maximum temperature criteria in high-performance microprocessors. However, only a rudimentary thermal profile with low resolution can be detected by these sensors since the number of sensors that can be integrated into a chip is limited by routing and pin-out constraints.

In order to predict the thermal gradients as well as the temperature profile of high-performance ICs in the early design stage, several methodologies have been developed to perform a full-chip thermal analysis. In [16] and [17], a chip-level temperature profile is generated by a numerical finite-difference approach incorporating temperature-dependent device models and lumped equivalent R–C network models. This approach solves the partial differential equations (PDEs) of heat transfer by a direct matrix factorization, which becomes complicated for large-scale problems. Different thermal-simulation algorithms have been proposed for improving computational efficiency. A chip-level 2-D and 3-D thermal simulator is presented in [18] and [19]. Instead of direct matrix solving, the simulator solves the similar heat-diffusion PDE by employing the alternating-direction-implicit (ADI) method with higher efficiency [20], [21]. In [22], a multigrid method, along with coarsening grid process, is presented to reduce the memory usage for computation. In [23], a combination of Green's function method and transformation is proposed for efficient steady-state thermal analysis. A full-chip thermal-simulation methodology using a precalculated constant power dissipation at the functional block level is proposed in [24]. In [25], the analysis for a full-chip and a cooling-system thermal model is presented.

However, all these analyses are mainly focused on either of the following: 1) improving the algorithms for solving heat-transfer equations to accelerate computational speed for substrate temperature estimations (improvement in simulation runtime) or 2) based on a cubic (unrealistic) package thermal model for the entire chip packaging stack-up, which in turn, compromises the accuracy of thermal estimation because the unrealistic package thermal model neglects the effect of heat spreading at different packaging layers. Most importantly, these works do not account for electrothermal couplings between the leakage power and the substrate temperature, as described in the companion paper [26], thereby making them ineffective for emerging nanometer scale designs.

C. Scope of This Paper

Prior substrate temperature-profile-estimation methods either employ an overly simplistic package thermal model or ignore these electrothermal couplings that are an inseparable aspect of the nanometer-scale chip operation. Hence, unlike previous works that target only the computational efficiency, a novel full-chip thermal analysis methodology is proposed. The method incorporates these electrothermal couplings, as well as a realistic package thermal model, to improve the accuracy of the substrate thermal-profile estimation, and it is implemented via one of the widely used efficient algorithms for solving heat-diffusion equations.

The rest of this paper is organized as follows. In Section II, numerical approaches for solving parabolic PDEs for heat diffusion are briefly discussed. The overview of the leakage-aware self-consistent method for estimating the chip-level substrate thermal gradient and temperature profile is presented. The methodology is then verified by comparing the generated temperature profile against the one generated by a computational-fluid-dynamics (CFD) based commercial software [27] under an identical simulation condition for a case where leakage is negligible. In Section III, the setup and implementation of the proposed methodology for a generic high-performance microprocessor with typical thermal (packaging and cooling) solutions are outlined. The impact of the package thermal model on temperature-gradient estimation is then discussed. Furthermore, the implications of the temperature profiles generated by the proposed methodology for power estimation and thermal management are presented. Finally, concluding remarks are made in Section IV.

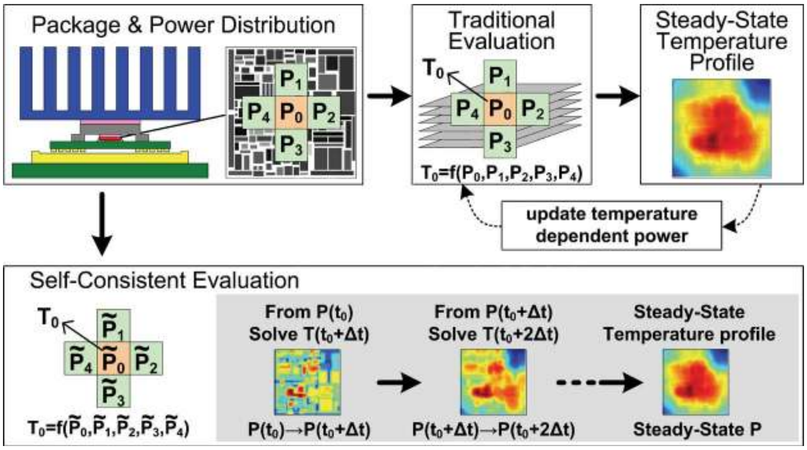

Package & Power Distribution
P1
P4 P0 P2
P3
Traditional Evaluation
T0
P1
P4 P0 P2
P3
T0=f(P0,P1,P2,P3,P4)
Steady-State Temperature Profile
update temperature dependent power
Self-Consistent Evaluation
T0
P̃1
P̃4 P̃0 P̃2
P̃3
T0=f(P̃0,P̃1,P̃2,P̃3,P̃4)
From P(t0) Solve T(t0+Δt)
P(t0)→P(t0+Δt)
From P(t0+Δt) Solve T(t0+2Δt)
P(t0+Δt)→P(t0+2Δt)
Steady-State Temperature profile
Steady-State P

Fig. 2. Schematic diagram showing the difference between traditional evaluation and the proposed self-consistent substrate thermal profile estimation methodology. Due to the strong interdependence of temperature and leakage power, the temperature at the central block (T_0) is not simply a function of the nominal power dissipation within and adjacent to the center block (P_0, P_1, P_2, P_3, and P_4) as per traditional analysis. Nominal power distribution should be updated self-consistently with the temperature evaluation (e.g. P_i is updated to $\widetilde{P_i}$). The proposed method evaluates the temperature by incorporating the correlation between power and temperature at each time step (Δt).

II. Self-Consistent Substrate Temperature Profile Estimation Methodology

A. Numerical Approach

PDEs of the general form shown in (1) are classified as parabolic PDEs [28], [29] and can be solved using the finite-difference approximation by two well-known approaches: explicit and implicit methods.

$$\frac{\partial\varphi(x,y,z,t)}{\partial t} = \alpha\left(\frac{\partial^2\varphi(x,y,z,t)}{\partial x^2} + \frac{\partial^2\varphi(x,y,z,t)}{\partial y^2} + \frac{\partial^2\varphi(x,y,z,t)}{\partial z^2}\right). \quad (1)$$

The explicit method is simple and straightforward [28], [29]. The explicit method calculates the state of a system at the next time step from the state of the system at the current time. However, in many cases, time steps must be very small to maintain stability; this results in long computation time for a steady-state analysis. In order to overcome the aforementioned disadvantages of the explicit method, the implicit method considers both the current state and the state at the next time step [28], [29], and the stability can be maintained over much larger values of time step. However, this method is more complicated to setup, and massive matrix manipulations require a considerable amount of computation memory and runtime for each time step.

The ADI method is a widely used algorithm for the numerical solution of parabolic PDEs involving multiple spatial variables [20], [21]. The advantage of applying this method arises from transferring a multiple dimensional parabolic PDE into a succession of 1-D problems (see the Appendix for more details). Therefore, no large-scale matrix has to be computed, and it is easy to implement. Thus, the ADI method is employed as the core algorithm to solve the heat PDEs for achieving higher computation efficiency. It is important to note that although other computationally efficient methods exist, choosing any one of them over the others does not affect the core results of the proposed methodology.

Fig. 2 shows the key aspect of the proposed approach as compared to traditional methods. Although the entire thermal profile can be obtained by the traditional evaluation, the traditional method is apparently misleading because it ignores the correlation between power and temperature. While one might think of applying the traditional evaluation iteratively by updating the temperature-dependent power (as shown by the dotted arrows), however, this dramatically increases the computation time. In addition, once the steady-state temperature is evaluated without considering the electrothermal couplings, the iterations (as shown by the dotted arrows) based on inaccurate information are meaningless. On the other hand, the proposed self-consistent approach evaluates the steady-state temperature profile by employing the ADI method such that the correlation between the power and the temperature can be incorporated at each time step (see the Appendix for more details of the numerical approach). Hence, the self-consistent method inherently generates a more accurate power profile, which can then be used to generate an accurate temperature profile by efficient PDE solvers.

B. Overview of the Methodology

In order to accurately estimate on-chip thermal gradients and the full-chip power dissipation profile, the methodology outlined in [30] is further improved with the capability of incorporating a precise layout geometry and the power dissipation of individual circuit blocks in a chip.

Fig. 3 shows the overview of the proposed leakage-aware self-consistent methodology for silicon-substrate temperature-profile estimation [31]. The chip (target simulation domain) is partitioned into a mesh according to the information provided by the layout geometry (block positions) and power-distribution map. Nominal power dissipation (including switching and leakage power) for each functional block is used as initial value according to its activity, depending on the specific circuit implementation and application. Physical parameters, such as specific heat, thermal conductivity, and heat-transfer coefficient, depend on the specific packaging material properties

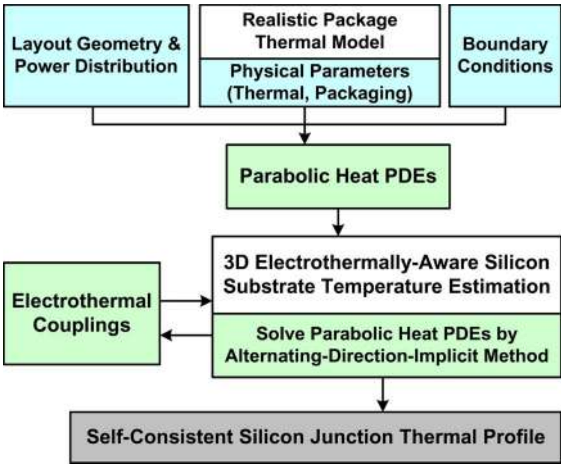

Layout Geometry & Power Distribution
Realistic Package Thermal Model
Physical Parameters (Thermal, Packaging)
Boundary Conditions
Parabolic Heat PDEs
Electrothermal Couplings
3D Electrothermally-Aware Silicon Substrate Temperature Estimation
Solve Parabolic Heat PDEs by Alternating-Direction-Implicit Method
Self-Consistent Silicon Junction Thermal Profile

Fig. 3. Overview of the electrothermally-aware methodology for silicon-substrate temperature-profile estimation. The methodology has been implemented as a simulation tool [31].

and applied cooling techniques. The full-chip realistic package thermal model (introduced in the companion paper [26]) is then incorporated, which comprehends both vertical and lateral heat-transfer paths. Boundary conditions are determined by the operating environment. The simulator uses the layout geometry, nominal power dissipation, boundary conditions, and physical thermal/packaging parameters as initial values to formulate parabolic partial differential equations and then solves these equations in a self-consistent manner using the ADI method for every mesh element. The algorithm converts a multiple dimensional parabolic PDE into a succession of 1-D linear equations. The electrothermal couplings are also embedded in the core of the simulator that simultaneously estimates the temperature-dependent quantities for each simulation step. Once the difference of the temperature evaluation between two steps is within a certain range, the evaluation stops, and the steady-state temperature profile is obtained. However, if the temperature exceeds the maximum criteria (defined by reliability constraints) for certain extreme cases due to poor packaging solutions or high power dissipation, the evaluation stops, and thermal runaway is reported.

C. PDE Solver Verification

As discussed in Section I-B, in the early design stage, it is not feasible to obtain the substrate temperature profile by a direct measurement such as the thermal IR-imaging technique or via integrated thermal sensors. In this paper, an industrial-quality CFD-based commercial software [27] is used to verify the substrate (junction) temperature profile generated by the proposed methodology. The CFD software has been verified in the past against direct measurements in chips where leakage was not significant [32]. Since the CFD software does not incorporate electrothermal couplings, the verification is carried out for a case with negligible leakage.

In this study, a die with a simple layout geometry (25 functional blocks with identical dimension but different power dissipation) is considered. It is assumed that the laterals (four sides of each layer) and one of the surfaces of the die [toward the printed circuit board (PCB)] are adiabatic. Heat can only be transferred into the ambient (45 °C) by conduction through various packaging layers and by convection from the heatsink. Under the same heat flux of all the functional blocks, material properties, and boundary conditions, Figs. 4 and 5 compare the steady-state substrate temperature profile evaluated by these two different techniques (commercial software and the proposed method) when the heat-transfer coefficients are 2000 and 4000 $\mathrm{W/m^2\,^\circ C}$, respectively. Note that the range of temperature shown in Figs. 4 and 5 is different, but the scales of the axes are kept constant to provide a better visual comparison.

Table I shows the estimated maximum and minimum junction temperature obtained using these two techniques for different heat-transfer coefficients. Moreover, quantitatively, 10 000 data points from identical locations of these two profiles are evenly selected (over the entire substrate) and compared to calculate the maximum deviation. As shown in Table I, the maximum and minimum temperature values, as well as the small values of the maximum deviation (obtained by comparing 10 000 data points), indicate a good agreement between the proposed methodology and the commercial software. Moreover, the simulation runtime of the proposed method is comparable or even less (less than 5 minutes to converge) than that of the commercial software. It is instructive to note that other previous works might have reported a higher computational efficiency by employing a cubic package thermal model or by neglecting the electrothermal couplings. However, it is more meaningful to have an accurate temperature-profile evaluation than to only reduce the simulation runtime.

III. Implementation and Implications of Self-Consistent Thermal-Profile Estimation

A. Setup and Implementation

The methodology is further implemented on a PC (3.06-GHz Pentium 4 processor, 1-GB memory) using C++ language. A microprocessor design with a die size of 10 mm × 10 mm (discretized into 100 × 100 grids) and with power densities per functional block is shown in Fig. 6. The power dissipation of the chip or each functional block depends on the application (workload, activity, etc.). However, in this analysis, the power-distribution map is known. The nominal total power consumption of the chip at ambient temperature (45 °C) is 96 W (nominal active power = 93.1 W; leakage power = 2.9 W). The short-circuit component is relatively small; therefore, it is neglected for simplicity. The physical and thermal properties of all packaging layers are evaluated according to a practical packaged high-performance microprocessor (Table II).

B. Implications of Self-Consistent Thermal-Gradient Estimation

In order to demonstrate the importance of incorporating electrothermal couplings and realistic package thermal model for estimating the substrate temperature profile, four different simulation scenarios are compared using the design shown in Fig. 6. Although the results of the proposed methodology have not been verified against direct measurements, the method

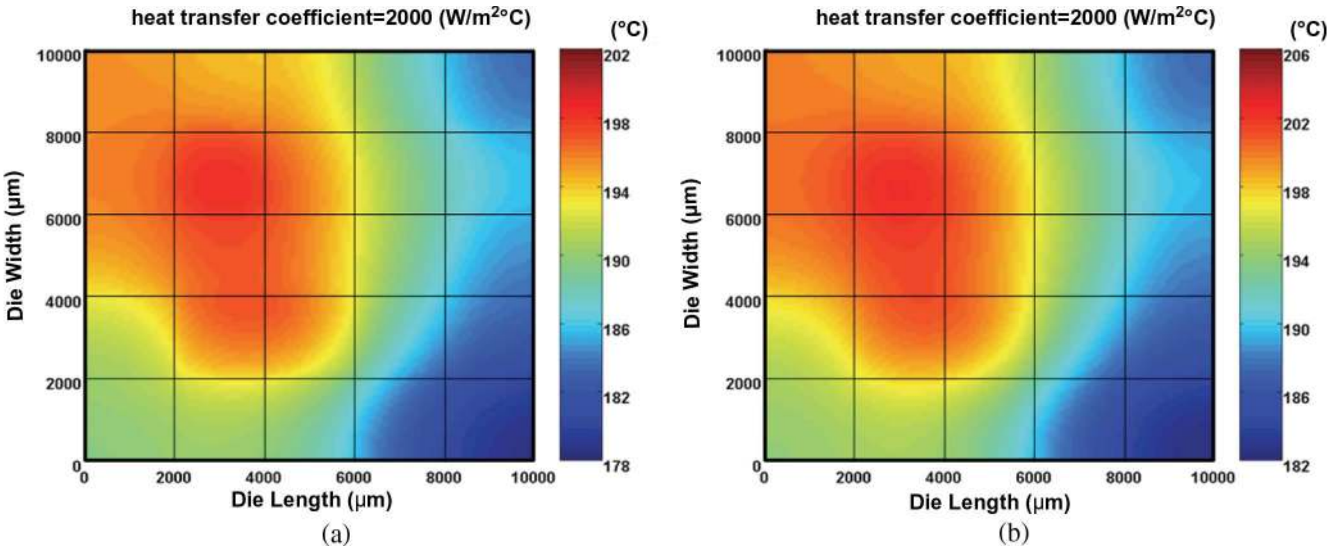


Fig. 4. Substrate temperature profile when the heat-transfer coefficient is 2000 $\mathrm{W/m^2\,°C}$. (a) Commercial software. (b) Proposed methodology.

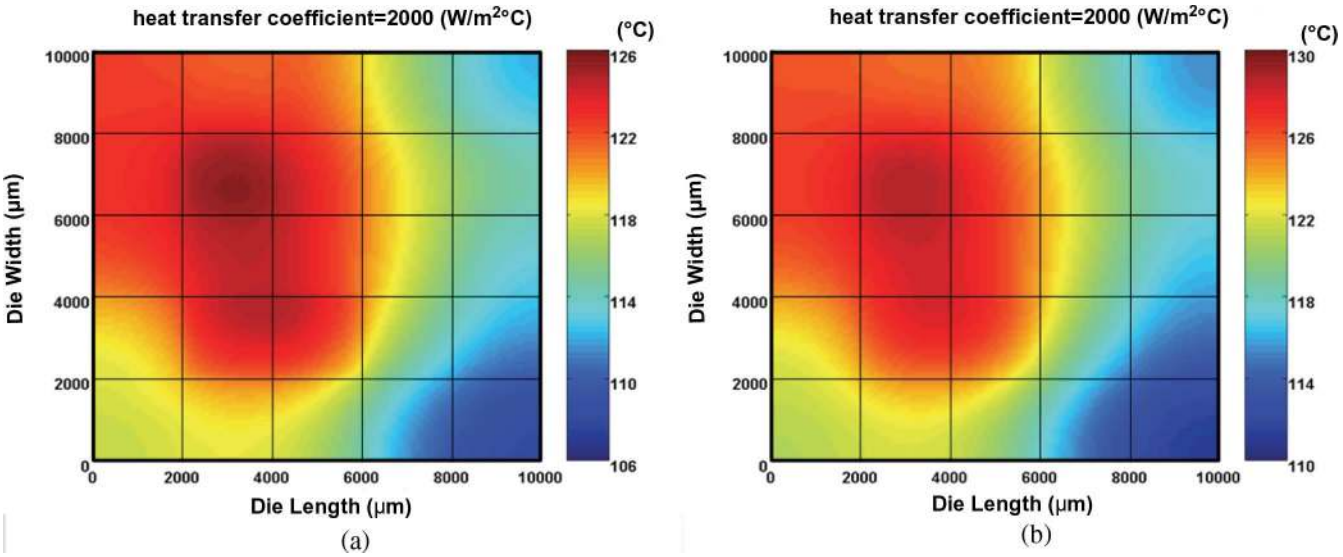
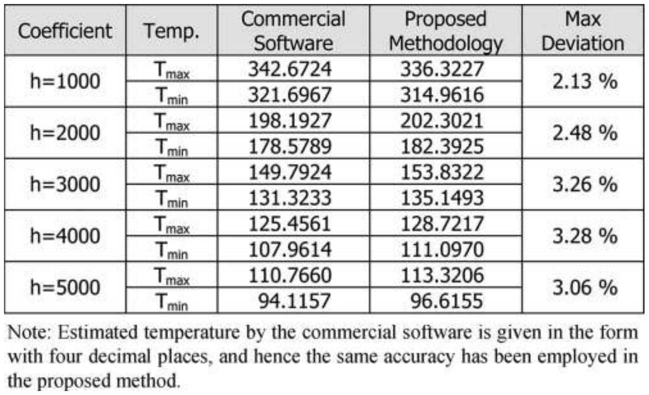


Fig. 5. Substrate temperature profile when the heat-transfer coefficient is 4000 $\mathrm{W/m^2\,°C}$. (a) Commercial software. (b) Proposed methodology.

TABLE I
COMPARISON BETWEEN THE COMMERCIAL SOFTWARE AND THE PROPOSED METHODOLOGY FOR DIFFERENT HEAT-TRANSFER COEFFICIENTS



Coefficient	Temp.	Commercial Software	Proposed Methodology	Max Deviation
h=1000	T_{max}	342.6724	336.3227	2.13 %
	T_{min}	321.6967	314.9616	
h=2000	T_{max}	198.1927	202.3021	2.48 %
	T_{min}	178.5789	182.3925	
h=3000	T_{max}	149.7924	153.8322	3.26 %
	T_{min}	131.3233	135.1493	
h=4000	T_{max}	125.4561	128.7217	3.28 %
	T_{min}	107.9614	111.0970	
h=5000	T_{max}	110.7660	113.3206	3.06 %
	T_{min}	94.1157	96.6155	

Note: Estimated temperature by the commercial software is given in the form with four decimal places, and hence the same accuracy has been employed in the proposed method.

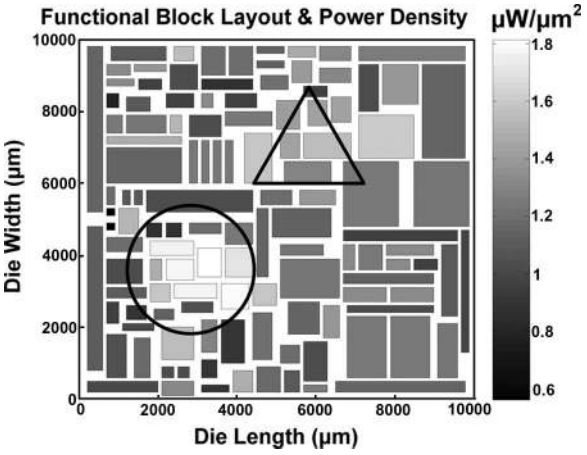
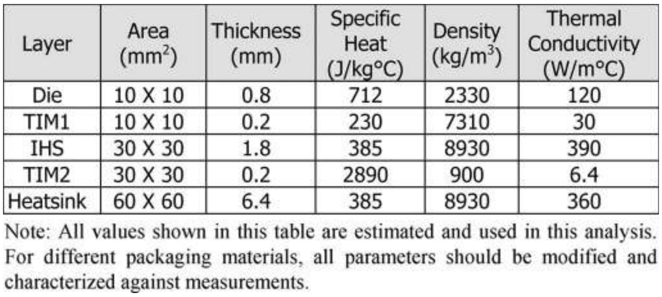


Fig. 6. Functional block layout of a microprocessor test chip [31]. Power densities associated with functional blocks are also shown. The circle encloses a region where the functional blocks have the highest power density. The triangle encloses the functional blocks that have higher leakage-power dissipation than all other blocks.

TABLE II
DIMENSIONS AND THERMAL PROPERTIES OF DIFFERENT PACKAGE LAYERS



Layer	Area (mm^2)	Thickness (mm)	Specific Heat (J/kg°C)	Density (kg/m^3)	Thermal Conductivity (W/m°C)
Die	10 X 10	0.8	712	2330	120
TIM1	10 X 10	0.2	230	7310	30
IHS	30 X 30	1.8	385	8930	390
TIM2	30 X 30	0.2	2890	900	6.4
Heatsink	60 X 60	6.4	385	8930	360

Note: All values shown in this table are estimated and used in this analysis. For different packaging materials, all parameters should be modified and characterized against measurements.

simply ensures the self-consistency between power and temperature during each iteration of the PDE solver, which has been validated against an industrial-quality CFD software. The same heat equations are employed, and the inclusion of the electrothermal couplings does not change the fundamental equations governing the thermal transport via heat conduction and convection but provides an algorithm to self-consistently solve the temperature and leakage power. Hence, once the core of the solver has been validated against the CFD, the results of the methodology can be trusted even with the inclusion of the electrothermal couplings.

Although the results are specific to the aforementioned IC, the conclusions are more generic. It can be observed that there is a region indicated by a circle in Fig. 6 where blocks have the highest power density. In addition, there is a region indicated by a triangle in Fig. 6 where blocks have 10× leakage-power dissipation with respect to the values of other functional blocks. However, the average power density of the circuit blocks in the triangle is around 60% of the average-power-density value in the circle.

Figs. 7(a) and (b) and 8(a) and (b) show the silicon-substrate temperature profiles generated under four different scenarios, respectively:

1) traditional method + cubic package thermal model;
2) traditional method + realistic package thermal model;
3) self-consistent method + cubic package thermal model;
4) self-consistent method+realistic package thermal model.

Note that the cubic and realistic package thermal models refer to [26, Fig. 11], which is the companion paper. In addition, in all these figures, the temperature profiles are shown using a constant temperature range (56 °C–66 °C) for ease of comparison.

The impact of electrothermal couplings on the substrate temperature evaluation can easily be observed by comparing Figs. 7(b) and 8(b), which both employ the realistic package thermal model [26, Fig. 11(b)] and the same cooling conditions. The substrate thermal profile in [Fig. 7(b)] is generated using a traditional thermal simulator, without considering electrothermal couplings. The highest temperature (hot-spot) is approximately 64.23 °C and is located in a region with the highest power density (indicated by the circle in Fig. 6). However, a different substrate temperature profile [Fig. 8(b)] is obtained by employing the proposed self-consistent methodology. From the temperature profile in Fig. 8(b), two hot-spots can be observed: one in the region with the highest power density and the other in the region with a higher percentage of leakage power. Unlike the traditional evaluation, the highest temperature is around 63.81 °C and is located in the region with a higher percentage of leakage power (indicated by the triangle in Fig. 6). Note that the self-consistent methodology comprehends the couplings between power (active and leakage) and temperature. The steady-state power dissipation (active and leakage) is self-consistent with the temperature and may not be equal to the nominal power dissipation.

As explained in [26], the regions with higher switching power density do not necessarily yield a higher temperature due to the various electrothermal couplings. Although the highest temperature values are similar in Figs. 7(b) and 8(b), the silicon temperature profile obtained by the self-consistent evaluation shows an additional hot-spot and, thus, a different temperature distribution. The traditional estimation is clearly misleading in terms of hot-spot count, location, and the overall spatial temperature profile as it neglects the electrothermal couplings between power dissipation and temperature.

Moreover, Fig. 9 shows the increase in the maximum substrate temperature ($\Delta T_{\max}$) with an increase in the leakage-power consumption (P_{total} is constant). It can be observed that the traditional evaluation, which does not capture the electrothermal couplings, results in a constant maximum temperature rise, which is certainly misleading. Since leakage is known to exacerbate with scaling, the significance of employing the proposed methodology is therefore expected to increase, as technology scales.

Hot-spots are known to determine the system-level thermal-management choices since the packaging and cooling solutions have to meet the maximum heat-flux requirements at the silicon-package interface. As already shown in Fig. 9, two curves with different θ_{ja} have different slopes as the technology becomes leaky, i.e., the impact of lowering θ_{ja} on the hot-spot temperature by packaging and cooling solutions will increase for leakage dominant technologies.

The impact of employing two different package thermal models for the cooling path on the substrate temperature-profile estimation can be observed by comparing Fig. 8(a) and (b). For fair comparison, the layout, power-density distribution, and discretization of the die are kept identical. In addition, the physical and thermal properties of each packaging-layer material are kept constant in both models. Fig. 8(a) shows the estimated substrate temperature profile by using a cubic (unrealistic) package thermal model. Although the electrothermal couplings are considered, unrealistic package thermal model underestimates the lateral heat spreading of the packaging layers (particularly in integrated heat spreader and heatsink) and thus results in a higher maximum and average substrate temperature. However, it is also important to note that although the maximum temperature is lower, the temperature gradient from the hot-spot to the edges of the chip is higher while employing the realistic package thermal model (e.g., $T_{\max}$ is 65.69 °C in Fig. 8(a) and 63.81 °C in Fig. 8(b); $T_{\max} - T_{\min}$ in Fig. 8(a) and (b) are about 8 °C and 11 °C, respectively). Due to the use of larger heat spreader and heatsink in the realistic package thermal model, better lateral heat spreading leads to a lower maximum temperature but to even lower temperatures at the edges of

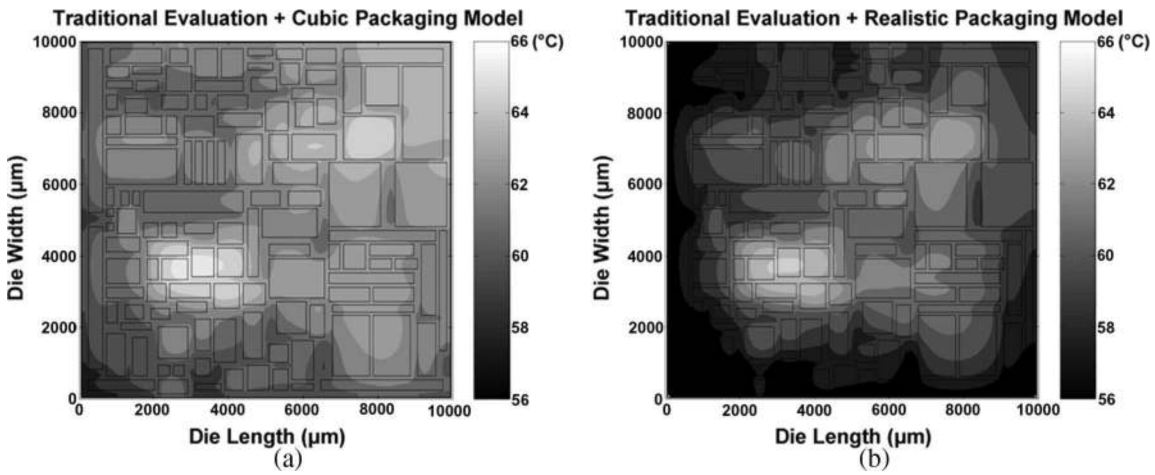


Fig. 7. Silicon-substrate temperature profile generated by traditional evaluation without considering electrothermal couplings. (a) A cubic package thermal model is employed. Only one hot spot can be observed. $T_{\max}$ is approximately 65.49 °C and located in the region with higher power density. $T_{\max} - T_{\min}$ is approximately 8 °C. (b) A realistic package thermal model is employed. Only one hot spot can be observed. $T_{\max}$ is approximately 64.23 °C and located in the region with higher power density. $T_{\max} - T_{\min}$ is approximately 11 °C.

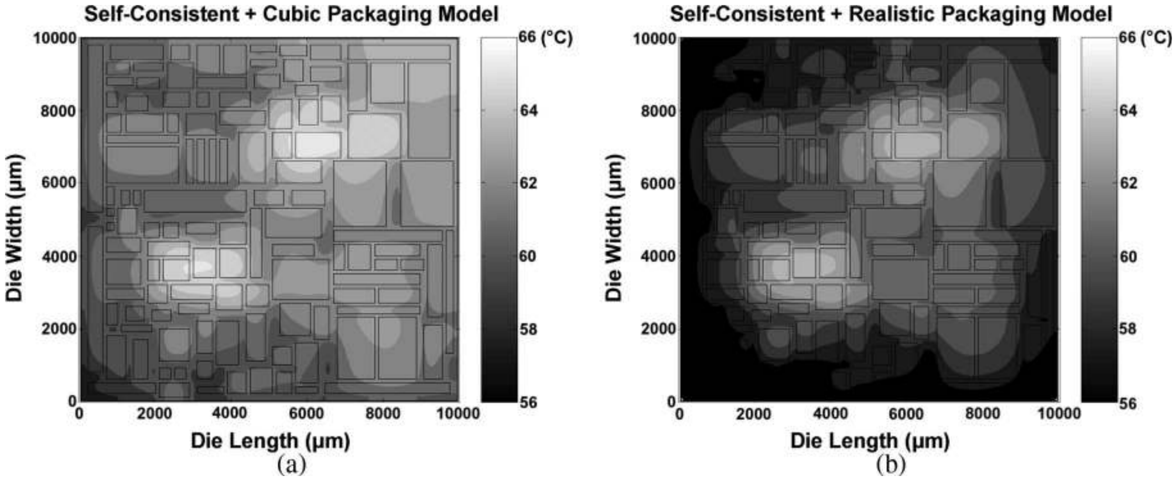


Fig. 8. Silicon-substrate temperature profile generated by the proposed self-consistent method. (a) A cubic package thermal model is employed. Two hot-spots can be observed. The highest temperature ($T_{\max}$) is approximately 65.69 °C and located in the region with higher percentage of leakage power. $T_{\max} - T_{\min}$ is approximately 8 °C. (b) A realistic package thermal model is employed. Two hot-spots can be observed. The highest temperature ($T_{\max}$) is approximately 63.81 °C. $T_{\max} - T_{\min}$ is approximately 11 °C.

the chip. This, in turn, is expected to impact physical design issues such as partitioning and placement schemes for high-performance microprocessors, including multicore designs.

Finally, for the chip used in this paper (Fig. 6), the power estimation, including active and leakage power, is shown in Fig. 10. The top horizontal bar represents the nominal value of power dissipation, and the rest three scenarios represent the conditions shown in Fig. 8(a) and (b) and Fig. 7(b), respectively. It can be observed that ignoring the electrothermal couplings in the traditional evaluation (even with realistic package thermal model) leads to erroneous leakage-power and total power estimation. On the other hand, the power distribution map (including switching and leakage power) can be self-consistently evaluated by the proposed methodology. In this particular case, the self-consistent method yields higher percentage of leakage power but lower total power due to the consideration of correlations between temperature and device drive/leakage currents within each functional block of the design.

IV. Conclusion

A leakage-aware self-consistent power and silicon-substrate temperature-profile-estimation methodology has been introduced in this paper for nanometer-scale CMOS technologies. The methodology takes various electrothermal couplings and realistic package thermal model into account. While traditional methodologies neglect electrothermal couplings and mislead hot-spot and thermal-gradient evaluation, it is demonstrated that the proposed methodology provides an accurate substrate temperature profile with an efficient numerical approach. In addition, the significance of incorporating electrothermal couplings is shown to increase with technology scaling. Moreover, it is shown that considering a realistic package thermal model not only improves the accuracy of estimating the heat distribution and power dissipation but also has significant implications for hot-spot and thermal-gradient management. For example, it is demonstrated that hot-spots can occur in regions with lower power densities if the percentage of leakage is higher

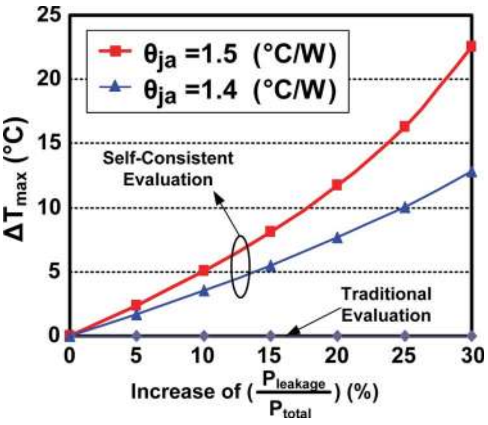


Fig. 9. Increase in maximum substrate temperature ($\Delta T_{\max}$) as a function of the leakage power dissipation increase. P_{leakage} and P_{total} denote the leakage power and total power consumption, respectively. The numbers in the x-axis represent the percentage increase of the ratio ($P_{\text{leakage}}/P_{\text{total}}$). $\Delta T_{\max}$ is defined as the temperature increase with respect to the value for nominal leakage-power dissipation. θ_{ja} is the effective thermal resistance between junction and ambient. Curves for traditional evaluation and different values of θ_{ja} are shown for comparison.

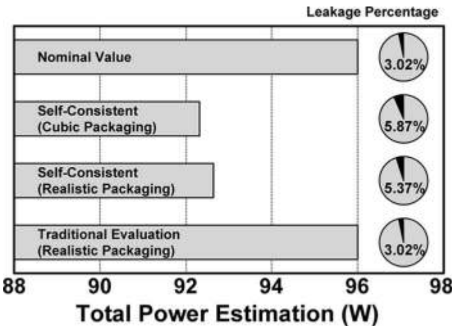


Fig. 10. Estimation of the total power dissipation under different scenarios. Nominal power dissipation is also shown for comparison. The horizontal bar (left-hand side) indicates the total power dissipation and the pie chart (right-hand side) shows the percentage of the leakage-power dissipation under each scenario.

in those regions. As power and thermal problems increasingly impact the scalability of CMOS devices and the architecture of high-performance IC products, including microprocessors, the proposed methodology will be invaluable for incorporating thermal-awareness in deep nanometer scale IC designs.

Appendix
Algorithm and Derivation of the Numerical Approach

In order to solve the parabolic PDE for the heat diffusion shown in [26, eq. (7)], the Crank–Nicholson implicit method [33] is employed for the second-order partial derivative, and the PDE can be rewritten as follows:

$$\frac{T^{n+1}-T^{n}}{\Delta t} = \left(\frac{k}{\rho C_p}\right)\left[\left(\frac{M_x}{2\Delta x^2}+\frac{M_y}{2\Delta y^2}+\frac{M_z}{2\Delta z^2}\right)\right.$$
$$\left.\times\left(T^{n+1}+T^{n}\right)\right]+\frac{p}{\rho C_p} \qquad \text{(A1)}$$

where Δt indicates the time interval between T^{n+1} and T^{n}. Note that T^{n} represents the temperature distribution $T(x,y,z,t)$ at $t=n$. The step sizes along the x-, y-, and z-directions are denoted as Δx, Δy, and Δz, respectively. M represents a linear operator and is defined as (A2) along the x-direction, where (i, j, k) denotes the grid point with coordinates $(i\Delta x, j\Delta y, k\Delta z)$.

$$\frac{\partial^2 T}{\partial x^2} \cong \frac{M_x T}{\Delta x^2} = \frac{T_{i+1,j,k}-2T_{i,j,k}+T_{i-1,j,k}}{\Delta x^2}. \qquad \text{(A2)}$$

A constant parameter is introduced in (A3) along the x-direction. Hence, (A1) can be rewritten as (A4).

$$\beta_x = \frac{k\Delta t}{2\Delta x^2 \rho C_p} \qquad \text{(A3)}$$

$$\begin{aligned}&(1-\beta_x M_x-\beta_y M_y-\beta_z M_z)(T^{n+1}-T^{n})\\ &\quad = 2(\beta_x M_x+\beta_y M_y+\beta_z M_z)T^{n}+\frac{\Delta t}{\rho C_p}p\end{aligned} \qquad \text{(A4)}$$

Equation (A4) can be further simplified into (A5) by introducing minor error terms [21]

$$\begin{aligned}&(1-\beta_x M_x)(1-\beta_y M_y)(1-\beta_z M_z)(T^{n+1}-T^{n})\\ &\quad = 2(\beta_x M_x+\beta_y M_y+\beta_z M_z)T^{n}+\frac{\Delta t}{\rho C_p}p.\end{aligned} \qquad \text{(A5)}$$

It can be observed from (A5) that the right-hand side is known at the present time step, and the unknown term $(T^{n+1}-T^{n})$ can be solved after transporting the terms $((1-\beta_x M_x)$, $(1-\beta_y M_y)$, and $(1-\beta_z M_z))$ to the right-hand side one at a time. For example, when the term $(1-\beta_x M_x)$ is transported from the left to the right-hand side, the solution is only updated for the x-direction (y- and z-directions are kept constant). Similarly, when transporting the term $(1-\beta_y M_y)$ from the left to the right-hand side, the solution is only updated for the y-direction (x- and z-directions are kept constant).

For instance, when transporting the term $(1-\beta_x M_x)$ from the left to the right-hand side, the set of linear equations can be represented as (A6), where U represents the remaining terms on the left-hand side of (A5), and V represents the terms in the right-hand side of (A5)

$$(1-\beta_x M_x)U_{x,y,z} = V_{x,y,z}. \qquad \text{(A6)}$$

After applying the linear operator M_x in (A6), the set of linear equations can be transformed into a tridiagonal-matrix form as (A7), where U and V are now $1 \times N$ matrices (N is the number of spatial grid points)

$$\begin{aligned}U_{x,y,z}-\beta_x M_x U_{x,y,z} &= V_{x,y,z}\\ \Rightarrow U_x-\beta_x[U_{x-1}-2U_x+U_{x+1}] &= V_x\\ \Rightarrow -\beta_x U_{x-1}+(1+2\beta_x)U_x-\beta_x U_{x+1} &= V_x\end{aligned}$$

$$\begin{bmatrix} b & a & 0 & 0 & \cdots & 0\\ a & b & a & & & \vdots\\ 0 & a & \ddots & & & 0\\ 0 & & & \ddots & a & 0\\ \vdots & & & a & b & a\\ 0 & \cdots & 0 & 0 & a & b \end{bmatrix}\begin{bmatrix} u_1\\ u_2\\ \vdots\\ u_{N-1}\\ u_N\end{bmatrix} = \begin{bmatrix} \nu_1\\ \nu_2\\ \vdots\\ \nu_{N-1}\\ \nu_N\end{bmatrix}. \qquad \text{(A7)}$$

Since the y- and z-directions are set to be constant in this step, U and V are only dependent on x. Note that the parameter a represents the coefficient $-\beta_x$ and that b represents $1 + 2\beta_x$ in the matrix. However, for physical case, the space of interest is finite and bounded. As in (A7), the first $(x = 1)$ and the last $(x = N)$ linear equations involve u_0 and u_{N+1}, respectively. These two equations are out of the simulation boundary $(u_1, u_2, \ldots, u_N)$. For instance, the first row of the tridiagonal matrix should be (A8), but u_0 is out of the boundary

$$au_0 + bu_1 + au_2 = \nu_1. \tag{A8}$$

From [26, eq. (8)], u_0 can be approximated and derived as a function of u_1 as follows:

$$u_0 = \frac{u_1}{1 + (h\Delta x/k)}. \tag{A9}$$

Similarly, in the last linear equation, u_{N+1} can be derived in terms of u_N. The coefficients of the first and the last rows of the tridiagonal matrix in (A7) need to be modified. Gauss-elimination method can be applied to solve the set of linear equations. Finally, T^{n+1} (temperature distribution for the next time step) can be obtained after transporting all the three terms $[(1-\beta_x M_x)$, $(1-\beta_y M_y)$, and $(1-\beta_z M_z)$ in (A5)]. Thus, the methodology essentially transfers the multiple dimensional parabolic partial differential equations into a succession of 1-D problems (such as (A7) for the x-direction and similar equations for the y- and z-directions). According to the change in temperature, all corresponding temperature-dependent parameters are evaluated and updated for the next time step until the temperature profile converges to steady-state.

REFERENCES

[1] K. Banerjee, “Thermal effects in deep sub-micron VLSI interconnects and implications for reliabilty and performance,” Ph.D. thesis, Univ. California Berkeley, Berkeley, CA, 1999.

[2] T. Karnik, S. Borkar, and V. De, “Sub 90-nm technologies-challenges and opportunities for CAD,” in *Proc. Int. Conf. Comput.-Aided Des.*, 2002, pp. 203–206.

[3] H. F. Hamann, A. Weger, J. A. Lacey, E. Cohen, and C. Atherton, “Power distribution measurements of the dual core PowerPC 970MP microprocessor,” in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2006, pp. 2172–2179.

[4] H. F. Hamann, A. Weger, J. A. Lacey, Z. Hu, P. Bose, E. Cohen, and J. Wakil, “Hotspot-limited microprocessors: Direct temperature and power distribution measurements,” *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 56–65, Jan. 2007.

[5] R. McGowen, “Adaptive designs for power and thermal optimization,” in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 118–121.

[6] C. Poirier, R. McGowen, C. Bostak, and S. Naffziger, “Power and temperature control on a 90 nm titanium-family processor,” in *Proc. Int. Solid-State Circuits Conf.*, 2005, pp. 304–305.

[7] A. H. Ajami, M. Pedram, and K. Banerjee, “Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs,” in *Proc. Custom Integr. Circuits Conf.*, 2001, pp. 233–236.

[8] A. H. Ajami, K. Banerjee, and M. Pedram, “Modeling and analysis of non-uniform substrate temperature effects on global ULSI interconnects,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 6, pp. 849–861, Jun. 2005.

[9] A. H. Ajami, K. Banerjee, and M. Pedram, “Analysis of substrate thermal gradient effects on optimal buffer insertion,” in *Proc. Int. Conf. Comput.-Aided Des.*, 2001, pp. 44–48.

[10] J. Lee, “Thermal placement algorithm based on heat conduction analogy,” *IEEE Trans. Compon. Packag. Technol.*, vol. 26, no. 2, pp. 473–482, Jun. 2003.

[11] B. Goplen and S. Sapatnekar, “Efficient thermal placement of standard cells in 3D ICs using a force directed approach,” in *Proc. Int. Conf. Comput.-Aided Des.*, 2003, pp. 86–89.

[12] C. N. Chu and D. F. Wong, “A matrix synthesis approach to thermal placement,” in *Proc. Int. Symp. Phys. Des.*, 1997, pp. 163–168.

[13] A. H. Ajami, K. Banerjee, and M. Pedram, “Scaling analysis of on-chip power grid voltage variations in nanometer scale ULSI,” *Analog Integr. Circuits Signal Process.*, vol. 42, no. 3, pp. 277–290, Mar. 2005.

[14] R. S. Prasher, J.-Y. Chang, I. Sauciuc, S. Narasimhan, D. Chau, G. Chrysler, A. Myers, S. Prstic, and C. Hu, “Nano and micro technology-based next-generation package-level cooling solutions,” *Intel Technol. J.*, vol. 9, no. 4, pp. 285–296, Nov. 2005.

[15] S.-C. Lin, R. Mahajan, V. De, and K. Banerjee, “Analysis and implications of IC cooling for deep nanometer scale CMOS technologies,” in *IEDM Tech. Dig.*, 2005, pp. 1041–1044.

[16] Y.-K. Cheng, C.-C. Teng, A. Dharchoudhury, E. Rosenbaum, and S.-M. Kang, “A chip-level electrothermal simulator for temperature profile estimation of CMOS VLSI chips,” in *Proc. IEEE Int. Symp. Circuits Syst.*, 1996, pp. 580–583.

[17] 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 Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 17, no. 8, pp. 668–681, Aug. 1998.

[18] T.-Y. Wang and C. C.-P. Chen, “3-D thermal-ADI: A linear-time chip level transient thermal simulator,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 12, pp. 1434–1445, Dec. 2002.

[19] 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,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 4, pp. 691–700, Aug. 2003.

[20] D. W. Peaceman and H. H. Rachford, “The numerical solution of parabolic and elliptic differential equations,” *J. Soc. Ind. Appl. Math.*, vol. 3, no. 1, pp. 28–41, Mar. 1955.

[21] J. Douglas and H. H. Rachford, “On the numerical solution of heat conduction problems in two or three space variables,” *Trans. Amer. Math. Soc.*, vol. 82, no. 2, pp. 421–439, Jul. 1956.

[22] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, “Efficient full-chip thermal modeling and analysis,” in *Proc. Int. Conf. Comput.-Aided Des.*, 2004, pp. 319–326.

[23] Y. Zhan and S. Sapatnekar, “A high efficiency full-chip thermal simulation algorithm,” in *Proc. Int. Conf. Comput.-Aided Des.*, 2004, pp. 634-637.

[24] Z. Yu, D. Yergeau, R. W. Dutton, S. Nakagawa, J. Deeney, N. Chang, S. Lin, and W. Xie, “Fast placement-dependent full chip thermal simulation,” in *Proc. Int. Symp. VLSI Technol., Syst., Appl.*, 2001, pp. 249–252.

[25] W. Huang, M. R. Stan, K. Skadron, K. Sankaranarayanan, S. Ghosh, and S. Velusamy, “Compact thermal modeling for temperature-aware design,” in *Proc. Des. Autom. Conf.*, 2004, pp. 878–883.

[26] S.-C. Lin, G. Chrysler, R. Mahajan, V. De, and K. Banerjee, “A self-consistent substrate thermal profile estimation technique for nanoscale ICs—Part I: Electrothermal couplings and full-chip package thermal model,” *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3342–3350, Dec. 2007.

[27] Icepak. [Online]. Available: http://www.icepak.com/

[28] M. N. Özişik, *Boundary Value Problems of Heat Conduction*. New York: Dover, 2002.

[29] R. Haberman, *Elementary Applied Partial Differential Equations With Fourier Series and Boundary Value Problems*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[30] K. Banerjee, S.-C. Lin, A. Keshavarzi, S. Narendra, and V. De, “A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management,” in *IEDM Tech. Dig.*, 2003, pp. 887–890.

[31] S.-C. Lin and K. Banerjee, “An electrothermally-aware full-chip substrate temperature gradient evaluation methodology for leakage dominant technologies with implications for power estimation and hot-spot management,” in *Proc. IEEE Int. Conf. Comput.-Aided Des.*, 2006, pp. 568–574.

[32] B. A. Zahn, “Evaluating thermal characterization accuracy using CFD codes—A package level benchmark study of IcePak and Flotherm,” in *Proc. Intersoc. Conf. Therm. Phenom.*, 1998, pp. 322–329.

[33] J. Crank and P. Nicholson, “A practical method for numerical evaluation solutions of partial differential equations of the heat conduction type,” in *Proc. Cambridge Philosoph. Soc.*, 1947, pp. 50–67.

Sheng-Chih Lin (S'03) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1996. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of California, Santa Barbara.

From 1998 to 2002, he was with the Phoenixtec Electronics Company, Ltd., and the CHROMA ATE Inc., respectively, in Taiwan. He joined Prof. Banerjee's research group at the University of California, Santa Barbara in Winter 2003. During the summer of 2005 and 2006, he worked as an intern in the Assembly and Test Technology Development of Intel in Chandler, Arizona. His research interests include electrothermal modeling and analysis of integrated circuits, variation-aware circuit design and optimization, and power/thermal management for nanoscale CMOS ICs. He has authored or coauthored over a dozen papers in journals and refereed international conferences.

Mr. Lin is a corecipient of the 2007 IEEE Micro Award.

Greg Chrysler (M'07) received the Ph.D. degree in mechanical engineering, specializing in thermal sciences, from the University of Minnesota, Minneapolis, in 1984.

He is currently a Principal Engineer with the Pathfinding Group, Assembly and Test Technology Development, Intel Corporation, Chandler, AZ. His major activities include the identification of new thermal-management and packaging technologies. He has authored several technical papers and is the holder of over 70 patents in packaging and cooling of electronics.

Dr. Chrysler is a member of the American Society of Mechanical Engineers (ASME). He was an Associate Editor of the *ASME Journal of Heat Transfer*.

Ravi Mahajan (M'01–SM'02) received the B.S. degree in mechanical engineering from the University of Bombay, Mumbai, India, in 1985, the M.S. degree in mechanical engineering from the University of Houston, Houston, TX, in 1987, and the Ph.D. degree in mechanical engineering from Lehigh University, Bethlehem, PA, in 1992. He specialized in fracture mechanics during his work toward the M.S. and Ph.D. degrees.

He is currently a Senior Principal Engineer with the Pathfinding Group, Assembly and Test Technology Development, Intel Corporation, Chandler, AZ. In this capacity, he is responsible for setting technology directions to enable packaging and assembly process for silicon at future nodes. He is also responsible for the technical direction of an Intel- and consortia-funded research in assembly and packaging. He has authored several technical papers in the areas of experimental and analytical stress analysis and thermal management. He is the holder of several patents in the area of microelectronic packaging.

Dr. Mahajan is a Fellow of the American Society of Mechanical Engineers and currently serves as an Associate Editor of the IEEE TRANSACTIONS ON ADVANCED PACKAGING. He has been the Editor and one of the Founding Members for the Section on Micro-Electronics for the Society of Experimental Mechanics. He is also one of the Founding Editors for the *Intel Assembly and Test Technology Journal*—an Intel internal journal that documents challenges and current progress in the area of assembly and packaging.

Vivek K. De (S'89–M'89–SM'07) received the B.S. degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 1985, the M.S. degree in electrical engineering from Duke University, Durham, NC, in 1986, and the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, in 1992.

He is an Intel Fellow and Director of Circuit Technology Research in the Circuits Research Lab (CRL) of Corporate Technology Group in Hillsboro, Oregon. In his current role, he provides strategic direction for future circuit technologies and is responsible for aligning Intel's circuit research with technology scaling challenges. He has published 152 technical papers in refereed conferences and journals, and 6 book chapters on low power circuits. He holds 136 patents, with 57 more patents filed (pending).

Dr. De received an Intel Achievement Award for his contributions to a novel integrated voltage-regulator technology.

Kaustav Banerjee (S'92–M'99–SM'03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

From 1999 to 2001, he was with Stanford University, Stanford, CA, as a Research Associate in the Center for Integrated Systems. From February to August 2002, he was a Visiting Faculty with the Circuit Research Laboratories, Intel, Hillsboro, OR. Since July 2002, he has been a member of the faculty in the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. He also held summer/visiting positions with Texas Instruments Incorporated, Dallas, from 1993 to 1997, and the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2001. His research has been chronicled in over 140 journals and refereed international conference papers and in a book chapter. He has also coedited a book entitled *Emerging Nanoelectronics: Life with and after CMOS* (Springer, 2004). His current research interests focus on nanometer-scale issues in high-performance/low-power VLSI as well as on circuit and system issues in emerging nanoelectronics.

Dr. Banerjee has served on the technical program committees of several leading IEEE and Association for Computing Machinery (ACM) conferences, including International Electron Devices Meeting (IEDM), Design Automation Conference (DAC), International Conference on Computer-Aided Design (ICCAD), and International Reliability Physics Symposium (IRPS). He has also served on the organizing committee of International Symposium on Quality Electronic Design at various positions including Technical Program Chair in 2002 and General Chair in 2005. Currently, he serves as a member of the Nanotechnology Committee of the IEEE Electron Devices Society. He has received a number of awards in recognition of his work, including the ACM Special Interest Group on Design Automation Outstanding New Faculty Award in 2004, a Research Award from the Electrostatic Discharge Association in 2005, a Best Paper Award at the Design Automation Conference in 2001, an Outstanding Student Paper Award at the VLSI/ULSI Multilevel Interconnection Conference in 2005, and an IEEE Micro Award in 2007.