A Self-Consistent Substrate Thermal Profile Estimation Technique for Nanoscale ICs—Part I: Electrothermal Couplings and Full-Chip Package Thermal Model

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 CMOS technology scales to nanometer regime, power dissipation issues and associated thermal problems have emerged as critical design concerns in most high-performance integrated circuits (ICs) including microprocessors. In this scenario, accurate estimation of the silicon junction (substrate or die) temperature is crucial for various performance analyses and chip-level thermal management. This paper introduces the notion of self-consistency in the junction temperature estimation by taking into account various electrothermal couplings between chip power, average junction temperature, operating frequency, and supply voltage. The self-consistent solutions of the average junction temperature are shown to have significant implications for various chip-level power, performance, reliability, and cooling-cost tradeoffs. Moreover, a realistic package thermal model is introduced that comprehends different packaging layers and non-cubic structure of the package, which are not accounted for in traditional analyses. The model is subsequently incorporated in the self-consistent substrate thermal profile estimation, which is discussed in Part II with implications for power estimation and thermal management in nanometer-scale CMOS technologies.**

***Index Terms*—Electrothermal couplings, integrated circuits (ICs), leakage, package, performance, power, thermal management.**

I. Introduction

DURING the past few decades, the key driver for the IC industry has been Moore's law, which states that the number of transistors per chip doubles approximately every 12 months [1] (later extended to 24 months [2]). The planar bulk MOSFET, which forms the building block of ICs and the enabler of Moore's law, has been continuously scaled in all physical dimensions [3] and in the power supply voltage applied to it [4]. While continued scaling of CMOS technologies provides substantial benefits in the form of higher transistor packing density, higher circuit performance, and lower cost of ICs, power consumption (in watts) and power densities (in watts per unit chip area) have been increasing steadily [4], [5]. Moreover, as CMOS has scaled from generation to generation, power dissipation has historically increased proportionately to increasing transistor density and switching speeds. However, with the minimum feature size of the transistor entering the nanometer regime (< 100 nm), leakage power has become a significant fraction of the overall chip power [6]. In addition, most leakage mechanisms are strongly temperature-dependent. This strong coupling between temperature and leakage can cause further increase in the total power dissipation. In fact, the International Technology Roadmap for Semiconductors (ITRS) [7] forecasts that high-performance microprocessors will dissipate around 200 W within a few years. Since the power consumed by the ICs is converted into heat, the corresponding heat densities also rise with increasing power consumption and power densities, resulting in elevated and nonuniform substrate temperatures. These electrothermal effects within the chip are leading to issues and challenges in the design and analysis of high-performance ICs that previous generations did not exhibit [8].

Elevated substrate temperature is widely known to have a strong impact on the performance and lifetime of devices and interconnects under "field," "accelerated testing," and "burn-in" conditions. Higher temperature increases the risk of damaging the devices and interconnects (since major back-end and front-end reliability issues including electromigration (EM), time-dependent dielectric breakdown (TDDB), and negative-bias temperature instability (NBTI) have strong dependence on temperature) even with advanced thermal management technologies [9]–[11]. Moreover, due to the increase in the number of interconnect levels and the introduction of low-κ dielectric materials with poor thermal conductivity, chip-level thermal problems have become even worse [12]–[14].

Due to technology scaling and parameter variations, including nonuniform dopant distribution in the channel region of the transistors [15]–[17], leakage power dissipation, which is dominated by subthreshold leakage for high-performance ICs, becomes a significant component of total chip power

Manuscript received October 5, 2006; revised August 23, 2007. 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 Pathfinding Group, Assembly and Test Technology Development, Intel Corporation, Chandler, AZ 85226 USA.

V. K. De is with the Circuit Research Lab, Intel Corporation, Hillsboro, OR 97124-6463 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.2007.909039

0018-9383/$25.00 © 2007 IEEE

consumption [18]. The subthreshold leakage is exponentially dependent on temperature [19] and exacerbates with technology scaling [6], [17]. Therefore, the increase in total chip power consumption that causes higher substrate temperature further increases the subthreshold leakage, thereby creating a strong feedback loop leading to various electrothermal couplings between power, temperature, operating frequency, and supply voltage [20]. Hence, a self-consistent electrothermal analysis method is highly desirable to accurately estimate the average junction temperature for leakage-dominant sub-100-nm technologies.

The rest of the paper is organized as follows. Factors leading to electrothermal couplings at the 100-nm technology node and beyond are discussed in Section II. Moreover, the details and significance of electrothermal couplings between various design metrics of high-performance ICs and temperature are described. In Section III, first-order electrothermally-aware analyses with implications for performance, reliability, and thermal management are demonstrated by using a self-consistent average junction temperature. In Section IV, based on a typical thermal (packaging) solution for high-performance microprocessors, a realistic package thermal model is formulated to comprehend the impact of packaging layers and to embed the electrothermal couplings into the traditional heat diffusion partial differential equations. The model is subsequently used to comprehend the impact of packaging layers in the companion paper [21].

II. Origin and Significance of Electrothermal Couplings

Typically, chip power dissipation has two major components: switching and leakage power dissipation. The short-circuit component is relatively small and temperature-independent and can be considered as a constant factor of total power [22], [23]. Hence, in this analysis, the short-circuit power has been neglected.

The switching power results from the charging and discharging of circuit capacitances between different voltage levels and increases with the chip frequency and supply voltage. The leakage power, particularly subthreshold leakage, used to be negligible but is rapidly becoming the dominant contributor to the total chip power because it is highly temperature-sensitive (being thermionic-emission-based) [Fig. 1(a)] and exacerbates with technology scaling [Fig. 1(b)]. Note that gate leakage (tunneling-based) is temperature-independent and can be mitigated by gate engineering [24]. In addition, the junction (diode) leakage is relatively small as compared with the subthreshold leakage [18].

The subthreshold leakage increases significantly due to the fact that supply voltage $(V_{\rm dd})$ scaling necessitates threshold voltage $(V_{\rm th})$ scaling to maintain a required performance according to the ITRS prediction [Fig. 2(a)]. In addition, elevated temperature lowers the threshold voltage of the transistor and thus increases the leakage further [19]. Moreover, since the gap between the wavelength of light for optical lithography and the polysilicon gate length is increasing [Fig. 2(b)] [17], channel length exhibits a significant amount of within-die variations

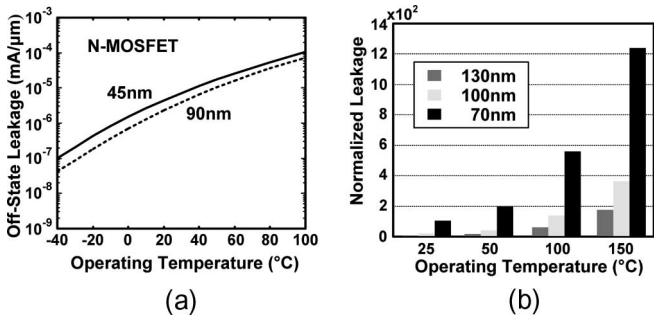


Fig. 1. (a) Transistor OFF-state leakage current for N-MOSFET (45-nm and 90-nm effective channel lengths) based on the BSIM3 models as a function of operating temperature. (b) Leakage power dissipation of an NMOS device for different technology nodes based on the BSIM3 models showing the impact of temperature. The leakage power dissipation is normalized w.r.t the value at 130-nm node at 25 °C.

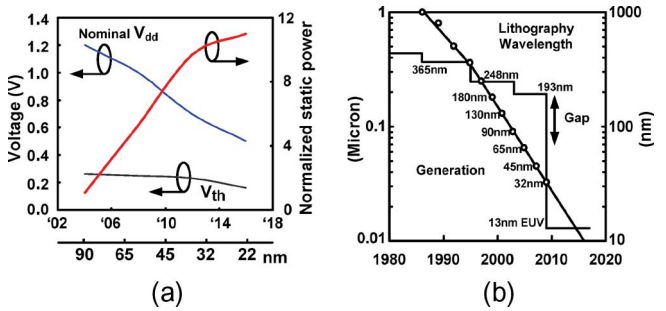


Fig. 2. (a) Nominal supply voltage, threshold voltage, and static power, as technology scales, based on the ITRS'04. (b) Increasing gap between polysilicon gate length and lithographic wavelength for different technology nodes [17].

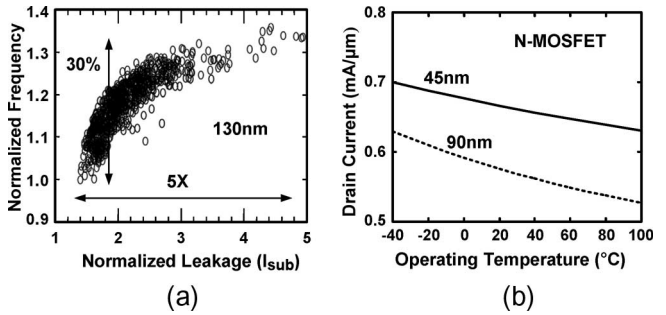


Fig. 3. (a) Distributions of frequency and standby leakage current for different microprocessors on a single wafer (courtesy of S. Borkar, Intel). (b) Transistor drive (drain) current for N-MOSFET (45- and 90-nm effective channel lengths) based on the BSIM3 models as a function of operating temperature.

[15], which in turn, leads to a significant impact on the distribution of leakage, as shown in Fig. 3(a).

The performance itself depends on temperature due to the dependence of the transistor on-current on operating temperature. Although the threshold voltage decreases at higher operating temperature and partially offsets the performance degradation resulting from the lower carrier mobility, the transistor on-current still decreases at higher operating temperatures [Fig. 3(b)].

The increase in total chip power consumption causes higher die temperature, which further increases the subthreshold leakage. Therefore, a strong feedback loop builds up, leading to various electrothermal couplings, which had been inconspicuous in

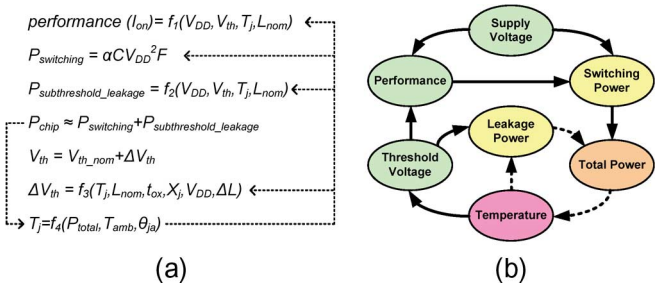


Fig. 4. (a) Models for various metrics are expressed in functional format. Couplings are indicated using broken lines. $L_{\rm nom}$ is the nominal gate length, α is the switching activity, C is the total load capacitance, F is the operating frequency, $t_{\rm ox}$ is the gate-oxide thickness, and X_j is the junction depth. (b) Schematic view of electrothermal couplings between different design metrics. As technology scales, the couplings between the total power, leakage, and temperature (shown by dotted arrows) become increasingly prominent.

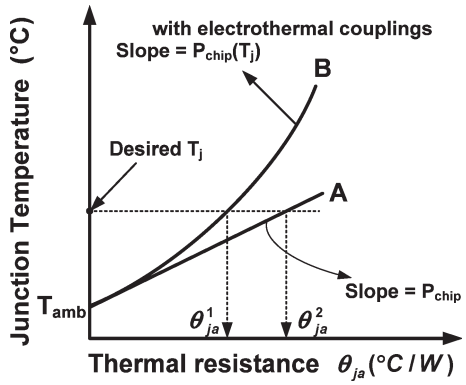


Fig. 5. Schematic diagram illustrates the dependence of junction temperature on thermal resistance for the conventional method (curve A) and for the method considering electrothermal couplings (curve B).

earlier generation of ICs. Fig. 4 shows such electrothermal couplings between performance, power dissipation, supply voltage, threshold voltage, and die temperature (see Appendix for more details).

Increasing subthreshold leakage power with scaling has a profound impact on chip-level thermal management strategies and can be understood by examining the following well-known expression for the junction-to-ambient thermal resistance:

$$\theta_{\rm ja} = \frac{T_j - T_{\rm amb}}{P_{\rm chip}}$$

where $\theta_{\rm ja}$ represents a lumped value of the thermal resistance between the silicon junction (substrate) and ambient, T_j is the average junction temperature, $T_{\rm amb}$ is the ambient temperature, and $P_{\rm chip}$ is the total chip power consumption. From the aforementioned equation, it can be observed that a larger junction-to-ambient temperature difference allows a larger $\theta_{\rm ja}$, which translates to smaller heatsink and air flow rate (i.e., smaller cooling cost) for dissipating the same power. Reduction in $\theta_{\rm ja}$ increases the packaging and cooling cost rapidly. This has been the primary reason why package designers in the recent past have allowed T_j to increase with increasing $P_{\rm chip}$. Maintaining larger T_j relaxes $\theta_{\rm ja}$ requirements in an active-power-dominated technology. However, for technologies that are leakage-dominant, a larger T_j will impact leakage power $(P_{\rm leakage})$ (Fig. 4), and hence $P_{\rm chip}$, and thereby influence $\theta_{\rm ja}$ and the cooling cost.

The importance of incorporating electrothermal couplings for evaluating system-level thermal solutions in leakage-dominant technologies is shown in Fig. 5 by sketching the dependence of T_j on $\theta_{\rm ja}$ [20]. Typically, T_j varies approximately linearly with $\theta_{\rm ja}$ when leakage is negligible (curve A in Fig. 5) as per the expression for the aforementioned $\theta_{\rm ja}$. However, T_j increases nonlinearly with $\theta_{\rm ja}$ due to the electrothermal couplings, arising primarily from the strong dependence of leakage on temperature (curve B in Fig. 5) in leakage-dominant technologies. Hence, in order to maintain a desired value of T_j to meet the reliability requirements, a lower value of $\theta_{\rm ja}(\theta_{\rm ja}^1 < \theta_{\rm ja}^2)$ will be needed, leading to an increase in the packaging/cooling cost. Therefore, for leakage-dominant technologies, it is important to comprehend electrothermal couplings in a self-consistent manner for accurate estimation of the silicon junction temperature for power and performance analysis/optimization and also for full-chip thermal management.

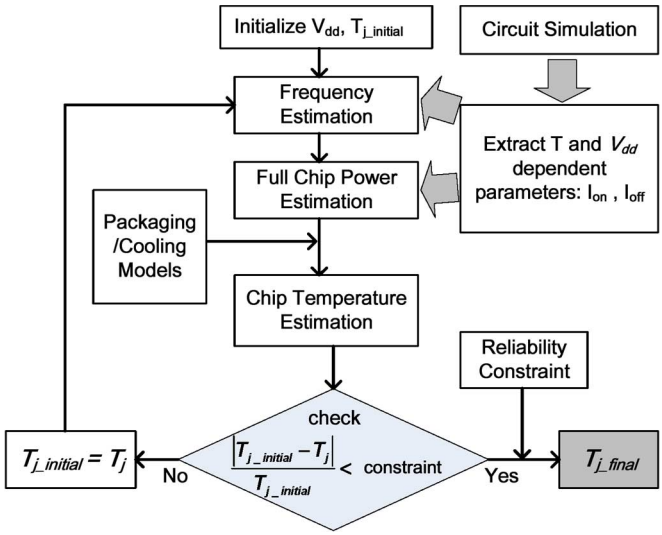


Fig. 6. Overview of the self-consistent average-junction-temperature-estimation technique [20].

III. Self-Consistent Average Junction Temperature: Analyses and Implications

First-order power and performance analyses incorporating electrothermal couplings can be performed by estimating an average silicon junction temperature using the self-consistent methodology (Fig. 6) introduced in [20]. In Fig. 6, the supply voltage and temperature dependent parameters ($I_{\rm ON}$ and $I_{\rm OFF}$) are first generated and tabulated using circuit simulation for representative transistors classified by different threshold voltages (devices with lower threshold voltages are only employed in the critical paths). For a given $V_{\rm dd}$ and initial T_j, the operating frequency of the chip is estimated using the extracted $V_{\rm dd}$ and temperature dependent $I_{\rm ON}$ with considerations of the logic depth in the critical path and associated capacitances (from the driver and load). The estimated frequency is then used in the calculation of the switching power. The leakage power is estimated by $I_{\rm OFF}$ (also $V_{\rm dd}$ and temperature dependent) as well as

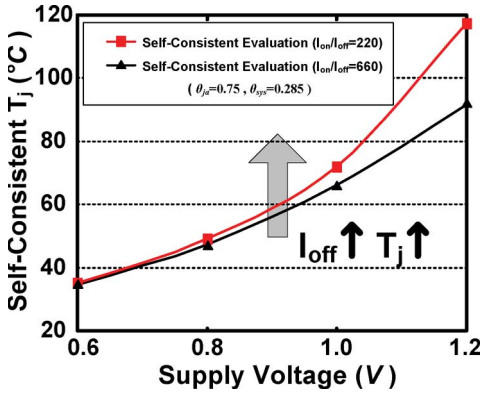


Fig. 7. Average junction temperature versus supply voltage for different $I_{\rm ON}/I_{\rm OFF}$ ratios. As the leakage increases (lowering the $I_{\rm ON}/I_{\rm OFF}$ ratio), the average junction temperature increases superlinearly with the supply voltage. Note that $\theta_{\rm sys}$ (°C/watt) represents the thermal resistance that relates the system power to the temperature difference between the ambient (45 °C) and outside (room) temperature (25 °C).

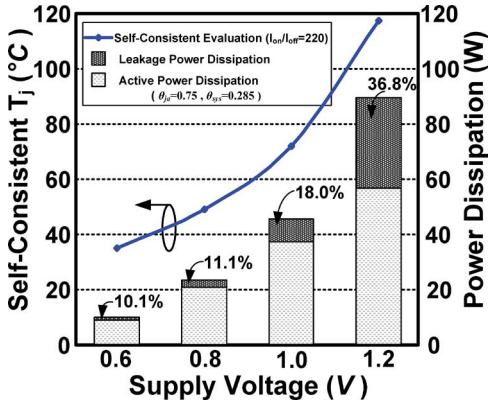


Fig. 8. Self-consistent average junction temperature and the total chip power dissipation evaluated by applying the self-consistent methodology at different supply voltages. An increase in the supply voltage will increase the active power dissipation and the junction temperature. However, the leakage power dissipation increases significantly due to the exponential dependence of the subthreshold leakage on temperature.

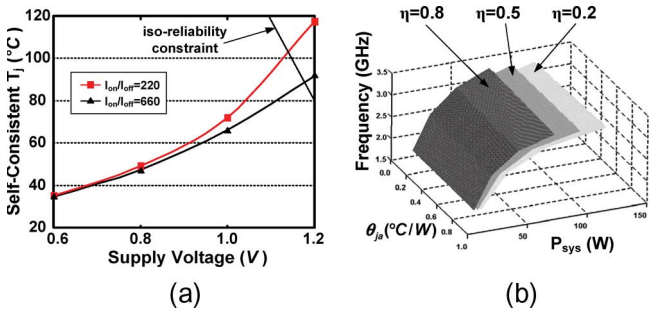


Fig. 9. (a) Implications of self-consistent approach on satisfying chip-level reliability constraints. (b) Design guidelines for integrated packaging and cooling solutions [20]. A higher chip performance can be achieved by employing more efficient cooling solutions (higher η) for a given constant system-level power dissipation and silicon junction-to-ambient thermal resistance (also known as constant packaging).

the gate leakage, while considering the transistor channel length variation. Compact packaging and cooling models [20] are employed to calculate the new T_j from the estimated total chip power. The estimated T_j is then compared with the initial T_j to check for convergence. The process continues till a convergence in the value of T_j is achieved and validated by a chip-level reliability check (that determines the maximum allowable T_j) for the particular value of $V_{\rm dd}$ in the analysis. The methodology was calibrated against the measured power (both active and leakage), frequency, and junction temperature data for a 32-bit microprocessor to extract and/or tune certain parameters used in the active- and leakage-power-estimation models.

Fig. 7 shows the impact of increasing subthreshold leakage (decreasing $I_{\rm ON}/I_{\rm OFF}$ ratio) on the average self-consistent junction temperature estimation. T_j is plotted as a function of supply voltage $(V_{\rm dd})$ for increasing values of the subthreshold leakage $(I_{\rm OFF})$. It can be observed that as $V_{\rm dd}$ increases, T_j increases more rapidly with increasing $I_{\rm OFF}$. This is due to the fact that at higher $V_{\rm dd}$, the chip temperature begins to increase and couples more strongly with the subthreshold-leakage current resulting in higher $P_{\rm leakage}$ and higher T_j.

Fig. 8 shows the self-consistent T_j (considering electrothermal couplings) and the percentage of leakage power in total chip power. It can be observed that leakage power dissipation becomes a major contributor of total power dissipation when T_j is high (higher $V_{\rm dd}$). This figure also implies that for leakage-dominant technologies, lowering of $V_{\rm dd}$ may not always lead to performance degradation as per traditional wisdom, but may offset the degradation or may even improve performance. This is due to the significant reduction of the junction temperature with lower supply voltage. In Fig. 9(a), a chip-level isoreliability constraint $(V_{\rm dd} \leq V_{\rm max} = T_j \cdot \beta + c)$ is superimposed on Fig. 7. The parameter β represents a chip-level reliability factor with a typical value of -3 mV/°C, and c is a constant depending on design requirements (e.g., $c = 1.44$ V in this case). It can be observed that for the leakier technology $(I_{\rm ON}/I_{\rm OFF} = 220)$, a lower value of $V_{\rm dd}$ is needed to meet the reliability constraint due to a higher value of T_j (determined by the intersection of the curves with the isoreliability line). Fig. 9(b) illustrates the impact of various packaging and cooling solutions corresponding to different values of the junction-to-ambient thermal resistance $(\theta_{\rm ja})$, cooling efficiency (η), and system power dissipation $(P_{\rm sys})$ [20] on the chip operating frequency. Note that $P_{\rm sys}$ consists of $P_{\rm chip}$ and cooling power dissipation $(P_{\rm cooling})$, where $P_{\rm cooling}$ is represented by $(1-\eta) \cdot P_{\rm chip}$ [20].

This plot provides design guidelines for efficient thermal-management of high-performance ICs. For example, in case of the processor considered in this study, a maximum frequency of ~3.4 GHz will be obtained with $\theta_{\rm ja} = 0.2$ °C/W and $\eta = 0.8$ along with $P_{\rm sys} \approx 75$ W. On the other hand, if the same performance is to be maintained with lower cooling efficiency, such as $\eta = 0.5$ and $\eta = 0.2$, $P_{\rm sys}$ will increase (due to increased $P_{\rm cooling}$) to around 95 W and 113 W, respectively.

Due to the correlation between power dissipation and temperature, it is important to note that if total power dissipation is given to be constant (e.g., measured power always includes electrothermal couplings automatically), employing the self-consistent methodology (considering electrothermal couplings) will always yield a lower estimation of average junction temperature than that obtained by using a non-self-consistent methodology (neglecting electrothermal couplings). This is because the self-consistent methodology takes into account the coupling

between temperature and power (primarily due to the strong dependence of P_{leakage} on T_j), whereas the latter will be limited by P_{leakage}, which cannot increase unbounded due to the constant power constraint. Hence, applying self-consistent methodology for average junction temperature estimation will not only provide more accurate temperature estimation but will also allow designers to avoid employing overly conservative design rules and thereby improve performance.

The benefit of applying self-consistent methodology for interconnect reliability lifetime due to EM has been shown in [25]. Not only the interconnect mean-time-to-failure (MTTF) can be evaluated accurately but also a higher peak current density of interconnect during the design phase is allowed due to a lower estimation of T_j at a constant system power dissipation [25]. Moreover, at the circuit level, electrothermal-energy-delay-product (EEDP) based optimization methodology and electrothermally-aware design-specific optimization metrics were shown to have significant implications for simultaneous optimization of power and performance to achieve design-specific targets [26], [27].

First-order tradeoffs between power, performance, reliability, and cooling-cost of high-performance ICs can be efficiently analyzed based on an average silicon junction temperature by the self-consistent methodology outlined earlier. However, a detailed profile of silicon junction temperature is imperative for investigating the impact of on-chip thermal gradients (hotspots) on power, performance, and reliability analyses. In order to generate an accurate substrate temperature profile of high-performance ICs, (as shown in the companion paper [21]) specifically for microprocessors, a more detailed consideration of thermal (packaging) solutions must be taken into account. This, however, requires a full-chip package thermal model as described in the next section.

IV. Full-Chip Package Structure and Thermal Model

A. Typical Chip Packaging Structure

Due to the increase in silicon junction temperature for nanometer-scale technologies, packaging has been transformed from playing the traditional role of a protective mechanical enclosure to a sophisticated thermal management platform [28], [29]. Fig. 10 shows a cross-sectional view of a typical microprocessor package structure containing a Flip-Chip Land Grid Array (FC-LGA) package and also a socket that interfaces with the printed circuit board (PCB). The microprocessor die is mounted on a package substrate (carrier).

Along the main heat-transfer path, as shown in Fig. 10, the microprocessor die and the package substrate are attached to an integrated heat spreader (IHS). The IHS, with a relatively larger area than that of the die, spreads the nonuniform heat from the die region to the top of the IHS. This improves the heat flux from a smaller die area to a larger surface that serves as the mating surface for the heatsink. Since the surface of these three major components (die, IHS, and heatsink) are never smooth enough to have a perfect contact, they are bonded together with a thermal interface material (TIM) applied between them. The TIM improves the poor thermal conductivity caused by surface roughness (conductivity of TIM is much larger than that of air) and thus enhances the overall thermal performance of the packaging stack-up and cooling mechanisms.

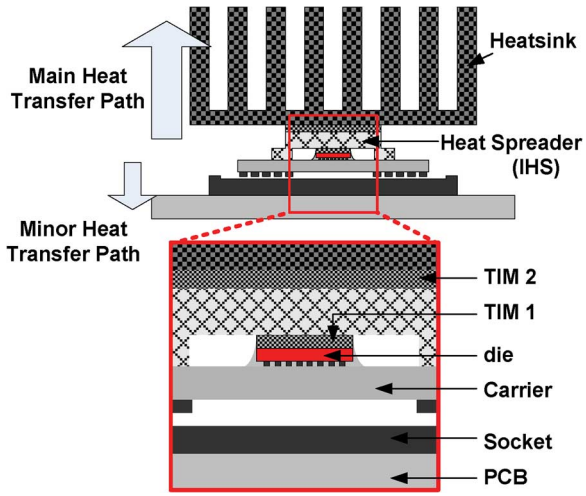


Fig. 10. Sketch of a microprocessor package assembly (drawing not to scale).

There is a second heat-transfer path from the die to the PCB, through the interconnect and dielectric layers, I/O pads, and carrier, as shown in Fig. 10. The thermal resistance of this path (from junction to the PCB) is normally several orders of magnitude higher than that of the major heat-transfer path [30]. Therefore, this path has been neglected throughout this paper because of the small fraction of heat it can transfer.

B. Heat Transfer Mechanisms in Packaged Chips

Heat is a form of energy that can be transferred as a result of temperature difference by three different modes: 1) conduction, in which heat passes through the matter itself; 2) convection, in which heat is transferred by the relative motion of portions of the heated body; and 3) radiation, in which heat is directly transferred between the distant portions of the body by electromagnetic radiation. In this paper, the effect of radiative heat losses is neglected for simplicity (effects of heat conduction and convection are considered) since its influence is negligible when forced convection is employed in most high-performance ICs [31]. The silicon die is the main source of heat generation. Heat can be exchanged and transferred by conduction within the entire packaging stack-up and by convection at the surface of the heatsink.

In a 3-D system, heat conduction can be quantified by Fourier's law as shown in the following:

$$\text{Rate of heat conduction (watt)} = -kA\frac{\partial T}{\partial n} \tag{1}$$

where the negative sign indicates that the heat transfer will be a positive quantity in the direction of a decreasing temperature (i.e., temperature gradient $\partial T/\partial n$ is negative) based on the second law of thermodynamics. The surface area normal to the direction of heat transfer is represented by A. The outward direction normal to the surface A is represented by n. The quantity T is the temperature distribution of the material. The thermal conductivity of the material is denoted by k and is a

measure of the ability of the material to conduct heat. Although the thermal conductivity varies with temperature, the variance is relatively small within the range of operation [32]. Hence, a constant thermal conductivity is employed for each material in the packaging structure at the nominal temperature in the analysis. Furthermore, for each layer, the thermal conductivity is identical in all directions (i.e., the material of each packaging layer is considered to be isotropic and homogeneous).

Next, heat convection, contributed mainly by the surface of the heatsink, can be expressed by Newton's law of cooling as shown in the following:

$$\text{Rate of heat convection (watt)} = hA(T_{\text{surface}} - T_{\infty}) \quad (2)$$

where h denotes the convection heat-transfer coefficient, T_{surface} is the surface temperature of heatsink, and T_{∞} is the environmental (coolant) temperature sufficiently far from the surface.

It is instructive to note that the thermal resistance mentioned in the previous section can be derived as (3) and (4) under different scenarios (conduction and convection) by the duality of electrical and thermal quantities, heat flow (Q), temperature (T), and thermal resistance (θ), which are analogous to current flow (I), voltage (V), and electrical resistance (R), respectively.

$$\theta_{\text{conduction}} = \frac{L}{kA}(^{\circ}\text{C/watt}) \quad (3)$$

$$\theta_{\text{convection}} = \frac{1}{hA}(^{\circ}\text{C/watt}). \quad (4)$$

Hence, one may establish a thermal resistance network to represent the entire packaging structure consisting of layers, and solve the voltages of the thermal resistance network to obtain the steady-state temperature distribution of all the layers. However, for large-scale problems, this approach becomes complicated when both computational efficiency and profile resolution are of importance. The problem exacerbates when a realistic packaging structure is considered [Fig. 11(b)].

Realistic packaging structures typically employ heat spreaders and heatsinks with larger dimensions (compared to the die) to improve the thermal performance of the main heat-transfer path (shown in Fig. 10). In practice, the area of the heat spreader and heatsink are at least 9 and 30 times larger than the area of the die, respectively. Although employing a cubic package thermal model for simplicity, as shown in Fig. 11(a), can improve the computational efficiency, this unrealistic package thermal model underestimates the lateral heat spreading due to the large packaging layers. Fig. 11(b) shows the relative dimensions of the realistic packaging layers, which are considered in the proposed methodology. Note that not only does the packaging structure involve different materials with different thermal properties but also their dimensions with respect to the silicon die are different, which will significantly influence the heat transfer as well as the substrate thermal profile.

C. Full-Chip Package Thermal Model

The fundamental physics of heat transfer in a chip substrate is governed by the following 3-D heat-conduction

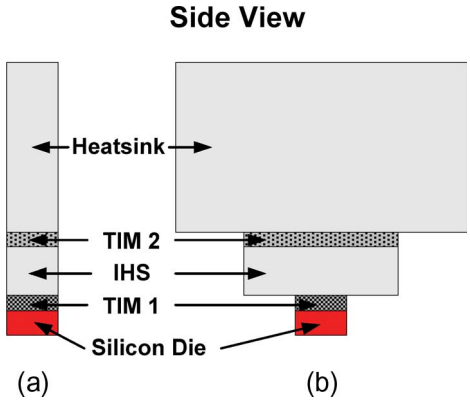


Fig. 11. Side view of (a) cubic package thermal model and (b) realistic package thermal model indicating different dimensions for each layer. The layout, power-density distribution, and dimension of the die are identical for both packaging cases. The thickness of different layers and the dimension of the layers are not drawn to scale.

equation and is subject to heat convection as the boundary condition [32]:

$$\rho C_p \frac{\partial}{\partial t} T(x,y,z,t) = \nabla \cdot [k(x,y,z,t)\nabla T(x,y,x,t)] + g(x,y,z,t) \quad (5)$$

$$k(x,y,z,t)\frac{\partial}{\partial n_i} T(x,y,z,t) = h\,[T(x,y,z,t) - T_{\text{amb}}] \quad (6)$$

where ρ is the density of the material (kg/m^3), C_p is the specific heat of the material $(\text{J/kg }^{\circ}\text{C})$, T is the temperature $(^{\circ}\text{C})$, k is the thermal conductivity of the material $(\text{W/m }^{\circ}\text{C})$, g is the internal heat generation (W/m^3), n_i is the outward direction normal to the boundary surface, h is the convective heat-transfer coefficient $(\text{W/m}^2\ ^{\circ}\text{C})$, and T_{amb} is the temperature of the ambient air surrounding the package measured at a specified distance sufficiently far away from the surface of the entire package.

As mentioned earlier, each discretized layer is considered to be isotropic and homogeneous. Therefore, a constant thermal conductivity is employed within one layer, and the temperature of the entire structure is modeled by rewriting the partial differential equations and boundary conditions as (7) and (8), where the temperature (T) is a function of the position (x, y, z) and the time (t).

$$\frac{\partial T}{\partial t} = \left(\frac{k}{\rho C_p}\right)\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right) + \frac{p}{\rho C_p} \quad (7)$$

$$\frac{\partial T}{\partial n_i} = \frac{h}{k}[T - T_{\text{amb}}]. \quad (8)$$

As discussed in the previous section, various electrothermal couplings need to be considered and incorporated into the thermal model and analysis. The parameter p in (7) is a function of temperature, time, and position within the die. Unlike the constant quantity g in (5), the parameter p represents the heat generation including the electrothermal couplings, and it needs to be recalculated at each evaluation step in a self-consistent manner (Fig. 6).

The entire thermal packaging stack-up (packaging material layers) is discretized based on a typical microprocessor package structure according to its physical dimensions. The

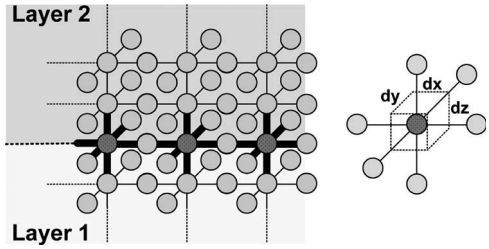


Fig. 12. Sketch of the discretization of the thermal packaging stack-up. Each node (circle) represents a discretized cell with a temperature value (T). Each discretized cell has six adjacent cells connected by edges (lines). Relationships between two adjacent cells are governed by (7) or (8) depending on heat-transfer mechanisms. The effective thermal conductivity of cells between two adjacent layers (darker nodes) can be determined by (9) since the dimensions of a discretized cell are equal (i.e., $dx = dy = dz$).

relationships between the discretized cells are governed by the heat partial differential equations and boundary conditions shown in (7) and (8). Physical thermal parameters, such as thermal conductivity, density, and specific heat of different layers, depend on material properties. Note that the dimensions of a discretized cell are chosen to be equal (i.e., $dx = dy = dz$). Thus, the effective thermal conductivity (k_{eff}) of cells between two adjacent layers, as represented by the darker nodes in Fig. 12 between layers 1 and 2, can be simply determined by

$$\frac{2}{K_{\text{eff}}} = \left(\frac{1}{K_1} + \frac{1}{K_2}\right) \tag{9}$$

where k_1 and k_2 represent the thermal conductivity of the material in layers 1 and 2, respectively. Since TIM is applied between two different layers to reduce the thermal contact resistance caused by surface roughness, a perfect thermal contact between the TIM layer and the adjacent materials is considered in the analysis.

Due to the presence of complex geometry and complicated boundary conditions, the silicon junction temperature profile cannot be solved analytically. However, a numerical solution can be found by finite-difference approaches and approximation schemes. A companion paper [21] investigates the implications of employing different package thermal models for substrate thermal profile estimation.

V. Conclusion

Electrothermal effects and couplings between chip power, subthreshold leakage, and operating temperature become increasingly prominent as CMOS technology scales below 90-nm. In this paper, the significance of considering electrothermal couplings for self-consistent average junction temperature estimation is highlighted and discussed. Implications for trading-off chip-level power, performance, reliability, and cooling-cost are also demonstrated by the self-consistent methodology. Moreover, based on a typical thermal solution of a high-performance microprocessor, a realistic package thermal model is introduced, which is incorporated in the self-consistent substrate thermal profile estimation in the companion paper.

Appendix

Temperature-dependent quantities in Section II are described using analytical models. As mentioned in Section I, the subthreshold-leakage current is the main source of MOSFET leakage and is dominated by source–drain diffusion current that is highly temperature-sensitive. The subthreshold current can be approximated by the following expression [19]:

$$l_{\text{ds}} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L}(m-1)(V_T)^2 \left[\exp\left(\frac{V_{\text{gs}} - V_{\text{th}}}{mV_T}\right)\right] \times \left[1 - \exp\left(\frac{-V_{\text{ds}}}{V_T}\right)\right] \tag{A1}$$

where μ_{eff} is the effective carrier mobility, C_{ox} is the oxide capacitance per unit area, W is the channel width, L is the channel length, m is the body-effect coefficient that typically lies between 1.1 and 1.4, V_T is the thermal voltage ($V_T = k_B T/q$), and V_{ds} represents the drain-to-source voltage, k_B is the Boltzmann constant, T is the absolute temperature, q is the electronic charge, V_{gs} represents the gate-to-source voltage, and V_{th} is the threshold voltage of a MOSFET. In (A1), the effective carrier mobility (μ_{eff}) and the threshold voltage (V_{th}) depend on temperature and can be modeled by the following relations:

$$\mu_{\text{eff}}(T) = \mu_{\text{eff}}(T_0)\left(\frac{T}{T_0}\right)^{\text{UTE}} \tag{A2}$$

$$V_{\text{th}}(T) = V_{\text{th}}(T_0) - \Delta V_{\text{th}}(T - T_0) \tag{A3}$$

where T is the junction temperature, and T_0 is the nominal (room) temperature. UTE is the mobility-temperature-exponent and is about -1.5 [33]. ΔV_{th} represents the temperature-dependent threshold voltage deviation ($\sim$0.7–1 mV/K) [19]. Note that μ_{eff} changes for different carriers.

The drive current of a short-channel MOS transistor is usually considered as the drain current under velocity saturation and can be modeled as follows:

$$l_{\text{ds}} = \upsilon_{\text{sat}} C_{\text{ox}} W \left(V_{\text{gs}} - V_{\text{th}} - \frac{1}{2}V_{\text{ds}}\right) \tag{A4}$$

where υ_{sat} is the saturation velocity of carriers, and V_{ds} is the drain-to-source voltage at saturation. Carrier saturation velocity (υ_{sat}) decreases slightly as the temperature increases [19] and can be modeled by the following relation:

$$\upsilon_{\text{sat}}(T) = \upsilon_{\text{sat}}(T_0) - \text{AT}\left(\frac{T}{T_0} - 1\right) \tag{A5}$$

where AT represents the temperature coefficient for saturation velocity and is around 3.3×10^4 m/s [33].

References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.
[2] G. E. Moore, "Progress in digital integrated electronics," in *IEDM Tech. Dig.*, 1975, pp. 11–13.
[3] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SSC-9, no. 5, pp. 256–268, Oct. 1974.
[4] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul./Aug. 1999.

[5] P. P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities, and new frontiers," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2001, pp. 22–25.
[6] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. IEEE Int. Symp. Low Power Electron. Des.*, 1999, pp. 163–168.
[7] *International Technology Roadmap for Semiconductors (ITRS)*. [Online]. Available: http://www.itrs.net
[8] K. Banerjee, S.-C. Lin, and N. Srivastava, "Electrothermal engineering in the nanometer era: From devices and interconnects to circuits and systems," in *Proc. 11th Asia South Pac. Des. Autom. Conf.*, 2006, pp. 223–230.
[9] P. Tadayon, "Thermal challenges during microprocessor testing," *Intel Technol. J.*, vol. 4, no. 3, pp. 1–8, 2000. 3rd quarter.
[10] R. Viswanath, V. Wakharkar, A. Watwe, and V. Lebonheur, "Thermal performance challenges from silicon to system," *Intel Technol. J.*, vol. 4, no. 3, pp. 1–16, 2000. 3rd quarter.
[11] 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, 2005. 4th quarter.
[12] K. Banerjee, A. Amerasekera, G. Dixit, and C. Hu, "The effect of interconnect scaling and low-k dielectric on the thermal characteristics of the IC metal," in *IEDM Tech. Dig.*, 1996, pp. 65–68.
[13] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," in *IEDM Tech. Dig.*, 2000, pp. 727–730.
[14] K. Banerjee and A. Mehrotra, "Global (interconnect) warming," *IEEE Circuits Devices Mag.*, vol. 17, no. 5, pp. 16–32, Sep. 2001.
[15] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. Des. Autom. Conf.*, 2003, pp. 338–342.
[16] S. Narendra, V. De, S. Borkar, D. Antoniadis, and A. Chandrakasan, "Full-chip sub-threshold leakage power prediction model for sub-0.18 μm CMOS," in *Proc. Int. Symp. Low Power Electron. Des.*, 2002, pp. 19–23.
[17] P. Gelsinger, in *Proc. 41st Des. Autom. Conf. Keynote*, 2004.
[18] Y.-S. Lin, C.-C. Wu, C.-S. Chang, R.-P. Yang, W.-M. Chen, J.-J. Liaw, and C. H. Diaz, "Leakage scaling in deep submicron CMOS for SoC," *IEEE Trans. Electron Devices*, vol. 49, no. 6, pp. 1034–1041, Jun. 2002.
[19] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[20] 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. 36.7.1–36.7.4.
[21] S.-C. Lin, G. Chrysler, R. Mahajan, V. K. De, and K. Banerjee, "A self-consistent substrate thermal profile estimation technique for nanoscale ICs—Part II: Implementation and implications for power estimation and thermal management," *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3351–3360, Dec. 2007.
[22] A. Chatterjee, M. Nandakumar, and I. C. Chen, "An investigation of the impact of technology scaling on power wasted as short-circuit current in low voltage static CMOS circuits," in *Proc. IEEE Int. Symp. Low Power Electron. Des.*, 1996, pp. 145–150.
[23] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. Electron Devices*, vol. 49, no. 11, pp. 2001–2007, Nov. 2002.
[24] P. M. Zeitzoff, "MOSFET scaling trends and challenges through the end of the roadmap," in *Proc. Custom Integr. Circuits Conf.*, 2004, pp. 233–240.
[25] S.-C. Lin, A. Basu, A. Keshavarzi, V. De, A. Mehrotra, and K. Banerjee, "Impact of off-state leakage current on electromigration design rules for nanometer scale CMOS technologies," in *Proc. IEEE Int. Reliab. Phys. Symp.*, 2004, pp. 74–78.
[26] A. Basu, S.-C. Lin, V. Wason, A. Mehrotra, and K. Banerjee, "Simultaneous optimization of supply and threshold voltages for low-power and high-performance circuits in the leakage dominant era," in *Proc. ACM Des. Autom. Conf.*, 2004, pp. 884–887.
[27] S.-C. Lin, N. Srivastava, and K. Banerjee, "A thermally-aware methodology for design-specific optimization of supply and threshold voltages in nanometer scale ICs," in *Proc. Int. Conf. Comput. Des.*, 2005, pp. 411–416.
[28] J. Adam, C.-S. Chang, J. J. Stankus, M. K. Iyer, and W. T. Chen, "Addressing packaging challenges," *IEEE Circuits Devices Mag.*, vol. 18, no. 4, pp. 40–49, Jul. 2002.
[29] R. Mahajan, R. Nair, V. Wakharkar, J. Swan, J. Tang, and G. Vandentop, "Emerging directions for packaging technologies," *Intel Technol. J.*, vol. 6, no. 2, pp. 62–75, 2002. 2nd quarter.
[30] S. Im, N. Srivastava, K. Banerjee, and K. E. Goodson, "Scaling analysis of multilevel interconnect temperatures for high-performance ICs," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2710–2719, Dec. 2005.
[31] R. D. Cess, "The effect of radiation upon forced-convection heat transfer," *Appl. Sci. Res.*, vol. 10, no. 1, pp. 430–438, Jan. 1961.
[32] M. N. Özişik, *Boundary Value Problems of Heat Conduction*. New York: Dover, 2002.
[33] *BSIM4 MOSFET Model by the BSIM Research Group in the Department of EECS, UC-Berkeley*. [Online]. Available: http://www.device.eecs.berkeley.edu/~bsim3/bsim4_intro.html

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 the 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 a Principal Engineer with the Pathfinding Group, Assembly and Test Technology Development, Intel Corporation, Chandler, AZ. His major activities include identification of new thermal-management and packaging technologies. He has authored several technical papers, was an Associate Editor of the *American Society of Mechanical Engineers (ASME) Journal of Heat Transfer*, and is the holder of over 70 patents in packaging and cooling of electronics.

Dr. Chrysler is a member of ASME.

Ravi Mahajan (SM'02) received the B.S. degree from the University of Bombay, Mumbai, Maharashtra, in 1985, the M.S. degree from 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.

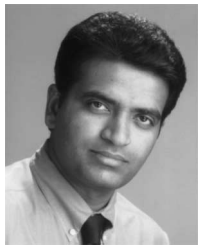
He is currently a Senior Principal Engineer with the Pathfinding Group, Assembly and Test Technology Development, Intel Corporation, Chandler, AZ. In this capacity, he is responsible for setting the technology directions to enable packaging and assembly process for silicon at future nodes. He is also responsible for the technical direction for 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 has been an Editor and one of the Founding Members of the Section on Microelectronics 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. 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.

Vivek K. De (S'89–M'91–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 the 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 the Center for Integrated Systems, Stanford University, Stanford, CA, as a Research Associate. From February to August 2002, he was a Visiting Faculty with the Circuit Research Lab, Intel Corporation, Hillsboro, OR. Since July 2002, he has been on the faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. He has also held summer/visiting positions at Texas Instruments, Inc., Dallas, TX, from 1993 to 1997, and the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2001. His research has been chronicled in over 140 journal 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 very large scale integrated circuits (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 the International Electron Devices Meeting, the Design Automation Conference, the International Conference on Computer-Aided Design, and the International Reliability Physics Symposium. He has also served on the organizing committee of the International Symposium on Quality Electronic Design at various positions including the Technical Program Chair in 2002 and the 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 works, including the ACM SIGDA Outstanding New Faculty Award in 2004, a Research Award from the Electrostatic Discharge Association in 2005, the Best Paper Award at the Design Automation Conference in 2001, the Outstanding Student Paper Award at the VLSI/ULSI Multilevel Interconnection Conference in 2005, and the IEEE Micro Award in 2007.