An Electrothermally-Aware Full-Chip Substrate Temperature Gradient Evaluation Methodology for Leakage Dominant Technologies with Implications for Power Estimation and Hot-Spot Management

Sheng-Chih Lin and Kaustav Banerjee
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106
{sclin, kaustav}@ece.ucsb.edu

ABSTRACT

As CMOS technology scales into the nanometer regime, power dissipation and associated thermal concerns in high-performance ICs due to on-chip hot-spots and thermal gradients are beginning to impact VLSI design. Moreover, elevated substrate (junction or die) temperature strongly influences IC performance, reliability, and packaging/cooling cost. Hence, accurate estimation of substrate thermal profiles is critical. This paper presents an accurate chip-level electrothermally-aware methodology for spatial silicon substrate temperature estimation. The methodology self-consistently incorporates various electrothermal couplings arising mainly due to the strong dependence of subthreshold leakage on temperature and also employs an accurate package thermal model, to account for inhomogeneous layers and non-cubic structure, which are not considered in traditional methods. The proposed methodology becomes increasingly effective as technology scales due to increasing leakage. Furthermore, it is shown that considering realistic package thermal models not only improves the accuracy of estimating temperature distribution but also has significant implications for power estimation and hot-spot management.

1. INTRODUCTION

While continued scaling of CMOS technologies provides substantial benefits in transistor density and circuit performance, increasing chip power consumption and power densities are rapidly leading to thermal management concerns. Moreover, highly integrated circuits including System-on-chips (SoCs) with different functional blocks, blocks with different activity rates (for example, logic vs. memory) and clock/power gating techniques essentially create non-uniform temperature distributions across the chip substrate [1]. The regions with higher temperature are commonly referred to as hot-spots. Hot-spots simultaneously lead to temperature gradients that affect performance [2] (including delay and timing) and reliability among a host of other issues, and also result in a general over-design in high-performance microprocessor packaging and cooling solutions. These thermal problems have now been identified and projected as major challenges for future IC design by leading semiconductor manufacturers and by the International Technology Roadmap for Semiconductors (ITRS) [3].

ICCAD'06, November 5-9, 2006, San Jose, CA

1.1 Implications of Substrate Temperature Rise and Non-Uniform Thermal Profile

Elevated substrate temperature is widely known to have a strong impact on the lifetime of devices and back-end reliability, including interconnects under "field," "accelerated testing" and "burn-in" conditions. Due to non-uniform power distribution, temperature in the local hot-spot regions can easily exceed the maximum reliability limit and increase the risk of damaging the device and interconnect (since major reliability mechanisms, including electromigration (EM), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI) have a strong dependence on temperature) even with advanced thermal management technologies [4]. Moreover, due to the increase in the number of interconnect levels and introducing low-k dielectric materials with poor thermal conductivity, thermal problems have become even worse [5-7].

High temperature not only leads to the onset and acceleration of reliability problems at the device and interconnect level but also impacts on circuit- and system-level metrics. Increased temperature deteriorates circuit performance by degrading device carrier mobility and increasing interconnect metal resistivity. This, in turn, will impact physical design issues including Power/Ground integrity [8] and placement and routing schemes [9-11]. At the system-level, thermal management (packaging and cooling) solutions are also affected by substrate temperature because they have to meet the maximum heat-flux requirements at the silicon-package interface [12].

1.2 Measurement and Modeling of Substrate Thermal Profile: Prior Work

Although thermal infrared imaging system is used in the industry for acquiring chip thermal profiles, it is not a useful tool during the design process, but is merely a verification technique after chip fabrication and packaging. Also, for typical high-performance microprocessors, the thermal profile obtained by the infrared imaging system does not accurately reflect the realistic substrate thermal profile of an operating microprocessor with a sophisticated packaging structure. 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 integrated thermal 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 gradient as well as the temperature profile of high-performance ICs, especially microprocessors, several methodologies have been developed to perform a full-chip thermal analysis. In [13, 14], a chip-level temperature profile is generated by a numerical finite difference approach incorporating temperature dependent device models and lumped R-C network models. This approach solves the partial differential equations (PDE) of heat transfer by direct matrix factorization, which becomes complicated for large scale problems. Different thermal simulation algorithms have been proposed for

improving computation efficiency. A chip-level 2-D and 3-D thermal simulator is presented in [15, 16]. Instead of direct matrix solving, the simulator solves the similar heat diffusion PDE by performing the Alternating Direction Implicit method with higher efficiency. In [17], multigrid (MG) method, along with coarsening grid process, is presented to reduce the memory usage for computation. In [18], a combination of Green's function method and transformation is proposed for highly efficient steady-state thermal analysis. A full-chip thermal simulation methodology using pre-calculated constant power dissipation at the functional block level is proposed in [19]. In [20], analysis for a full-chip and cooling system thermal model is presented. However, all these analysis are mainly focused on either improving algorithms for solving heat transfer equations or accelerating the computational efficiency for temperature estimations (improving the simulation runtime within a range of minute). Also, these analyses are all based on a cubic (unrealistic) thermal model for the entire chip and the 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, due to technology scaling and parameter variations [21, 22], including non-uniform dopant distribution in the channel region of the transistors [23], leakage power dissipation, which is dominated by subthreshold leakage for high-performance ICs, becomes a significant component of total chip power consumption. The subthreshold leakage is exponentially dependent on temperature [24] and exacerbates with technology scaling [25]. Also, the increase in total chip power consumption causes higher substrate temperature, which further increases the subthreshold leakage, thereby creating a strong feedback loop leading to various electrothermal couplings between power, temperature, operating frequency and supply voltage [26]. All prior substrate temperature profile simulation methods not only employ an overly simplistic package thermal model [13-20], but also ignore these electrothermal couplings that are an inseparable aspect of nanometer scale chip operation. Hence, unlike previous works that target only computational efficiency, we propose a full-chip thermal analysis methodology that incorporates these electrothermal couplings, as well as a realistic package thermal model to improve the accuracy of the substrate thermal profile estimation and the methodology is implemented via one of the widely-used efficient algorithms for solving the heat transfer diffusion equations.

The rest of the paper is organized as follows. Details of various electrothermal couplings between power dissipation, operating frequency and substrate temperature are described in Section 2. In Section 3, we formulate a realistic package thermal model and incorporate electrothermal couplings into heat partial differential equations. In Section 4, the implementation and discussion of proposed methodology are outlined. Impact of the realistic package thermal model formulation and implications of the electrothermally-aware methodology for power estimation and thermal management are discussed in Section 5. Finally, concluding remarks are made in Section 6.

2. ELECTROTHERMAL COUPLINGS

Chip power dissipation at the nanometer scale has two major components: switching power and leakage power dissipation (The short-circuit component is relatively small and therefore we neglect it throughout this paper). The switching power consumption increases with the chip frequency and supply voltage. Moreover, the performance itself is dependent on temperature. Increase in substrate temperature will decrease the transistor drive current due to the reduction in carrier mobility (although the threshold voltage decreases at higher operating temperature and partially offsets the performance degradation resulting from the lower carrier mobility, the transistor drive current still decreases at higher operating temperatures) as shown in **Fig. 1(a)**.

As mentioned earlier, subthreshold leakage, the main leakage contributor, is highly temperature sensitive (**Fig. 1(b)**). Moreover, due to technology scaling and parameter variations, leakage power dissipation becomes a major contributor to total chip power consumption [25]. The increase in total chip power consumption causes higher on-chip temperature, which further increases the subthreshold leakage. Therefore, a strong feedback loop leading to various electrothermal couplings occurs [26]. **Fig. 2** illustrates various electrothermal couplings between performance (frequency), power dissipation, supply voltage, threshold voltage, and substrate temperature. Hence, in order to accurately estimate power dissipation and resulting substrate temperature profile, various electrothermal couplings must be embedded into full-chip thermal modeling and analysis.

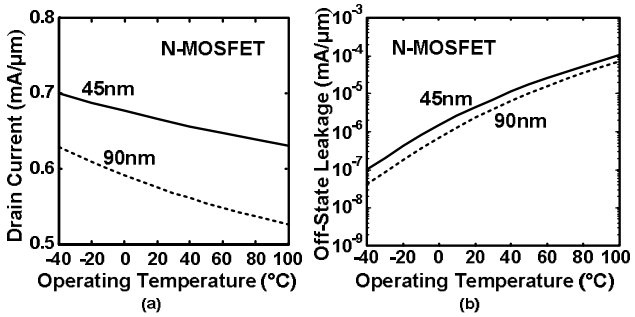

Figure 1. (a) Drive (drain) current **(b)** Off-state leakage current for NMOSFET (45 nm and 90nm technology nodes) as a function of operating temperature.

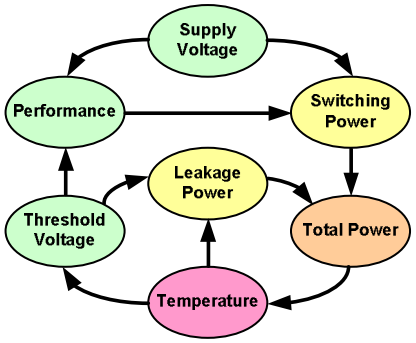

Figure 2. Schematic view of electrothermal couplings between different design parameters.

3. FULL-CHIP REALISTIC PACKAGE THERMAL MODEL

We consider a realistic microprocessor package structure, Flip-Chip Land Grid Array (FC-LGA) and a socket that interfaces with the printed-circuit board (PCB) as shown in **Fig. 3**.

The microprocessor die is mounted on a package substrate. An integrated heat spreader (IHS) is attached to the package substrate and microprocessor die. The IHS spreads the non-uniform heat from the die to the top of the IHS, and it improves the heat flux from a smaller die area to a larger surface and serves as the mating surface for a heatsink. Since the surface of these three major components (die, IHS and heatsink) are never smooth enough to have perfect contact, they are bonded together with a thermal interface material (TIM) applied between them. The thermal interface material improves the poor thermal conductivity caused by surface roughness (conductivity of

TIM is much larger than air) and thus, enhances the overall thermal performance of the stack-up packaging and cooling mechanisms. Moreover, a minor heat transfer path exists from the die to the printed-circuit board (PCB) mainly corresponding to the interconnect/dielectric layers and I/O pads. The thermal conductivity of this path (from substrate to the printed-circuit board) is normally several orders of magnitude smaller than that of the major heat transfer path [27]. Therefore, we neglect this path throughout this paper because of the small fraction of heat it can transfer.

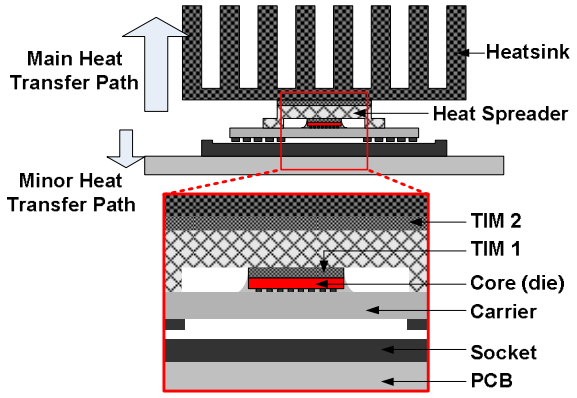

Figure 3. Sketch of microprocessor package assembly. (drawing not to scale)

Heat is considered to be a form of energy that can be transferred as a result of temperature difference by three different modes: conduction, convection, and radiation. However, since radiative heat losses are negligible for a packaged chip, we only consider conduction and convection in this paper.

In a three-dimensional system, heat conduction and convection can be quantified by Fourier's law and Newton's law of cooling as shown in (1) and (2) respectively.

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

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

where the negative sign indicates that the heat transfer will be a positive quantity in the direction of decreasing temperature (i.e., temperature gradient, ∂T, is negative), based on the second law of thermodynamics. The surface area normal to the direction of heat conduction 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. The thermal conductivity varies with temperature and depends on material characteristics [28]. Here, we use the thermal conductivity at the average temperature and treat it as a constant in all calculations. Also, we consider the thermal conductivity of each packaging layer to be isotropic. h denotes the convection heat transfer coefficient. $T_{surface}$ is the surface temperature, and T_{∞} is the environmental temperature sufficiently far from the surface.

Equivalent thermal resistances can be derived from (3) and (4) below under different scenarios (conduction and convection). By the duality of electrical and thermal quantities, power dissipation (heat flow), temperature, and thermal resistance are analogous to current flow, voltage, and electrical resistance, respectively. The equivalent thermal model can be established by a thermal resistance network to represent these inhomogeneous packaging material layers, and solving the voltages of the network gives the 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.

$$\theta_{conduction} = \frac{L}{kA} \quad (3)$$

$$\theta_{convection} = \frac{1}{hA} \quad (4)$$

In order to improve the thermal performance of the major heat transfer path, typically, a larger dimension of heat spreader and heatsink is used. In practice, the area of heat spreader and heatsink is at least *9X* and *30X* larger than the area of a die, respectively. Traditional chip-level thermal analyses employ a cubic thermal model for simplicity as shown in **Fig. 4(a)**. Although the resolution of the die region and the computation efficiency could be improved, this unrealistic package thermal model underestimates the lateral heat spreading due to large packaging layers. **Fig. 4(b)** illustrates the relative dimensions of realistic packaging layers, which we considered in the proposed methodology. Note that not only are they different materials with different thermal properties but also their dimensions with respect to the silicon die will significantly influence the heat transfer as well as the substrate thermal profile.

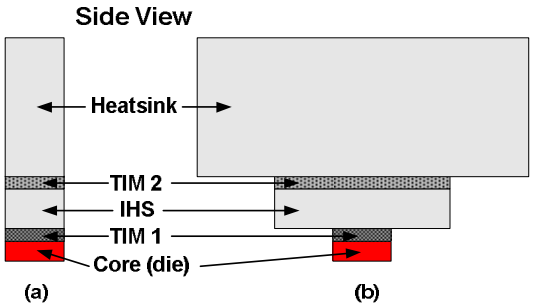

Figure 4. Side view of thermal packaging stack-up. The layout, power density distribution, and dimension of the die are identical for both packaging cases. The thickness of different layers and the dimension of the layers are not drawn to scale. **(a)** Cubic packaging model **(b)** Realistic packaging model indicating different dimensions for each layer.

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 convection at the surface of the heatsink. The fundamental physics of heat transfer in a chip substrate is governed by the following three-dimensional heat conduction equation and subject to heat convection as the boundary condition [28]:

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

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

where ρ is the density of the material (kg/m^3), C_p is the specific heat of material (J/kg°C), T is the temperature (°C), k is the thermal conductivity of the material (W/m°C), g is the internal heat generation (W/m^3), n_i is the outward direction normal to the boundary surface, h is the heat transfer coefficient (W/m^2°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, we consider each discretized layer to be isotropic and homogeneous. Therefore, we use a constant thermal conductivity within one layer, and the temperature of the entire structure will be modeled by rewriting the partial differential equation

and boundary condition as (7) and (8) where temperature (T) is a function of the position (x, y ,z) and time (t).

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

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

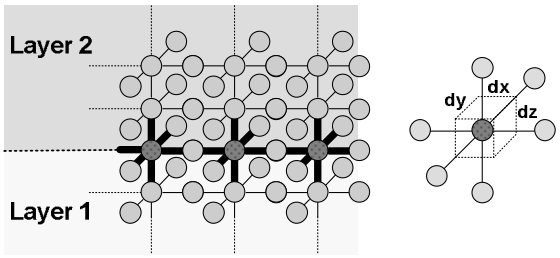

Figure 5. Sketch of the discretization of the thermal packaging stack-up. Each node (circle) in the figure represents a discretized cell with a temperature (T). Each discretized cell has six adjacent cells connected by edges (lines). Relationships between two adjacent cells are governed by (7) or (8) depending on heat transfer mechanisms. Effective thermal conductivity of cells between two adjacent layers (darker nodes) can be determined by (9) since the dimensions of each discretized cell are equal (i.e., $dx = dy = dz$).

As discussed in the previous section, various electrothermal couplings need to be considered and incorporated into the thermal model and analysis. Therefore, the parameter p in equation (7) is a function of temperature, time and the position within the die. It represents a modified parameter of the constant quantity g in (5) for the internal heat generation. Since there are electrothermal couplings between power dissipation, operating frequency and die temperature, the quantity p is not a constant value like g and will be evaluated in a self-consistent manner at each iteration step.

We discretize the entire thermal packaging stack-up (inhomogeneous packaging material layers) based on a typical microprocessor package structure according to its physical dimensions. Relationships between 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 effective thermal conductivity (k_{eff}) of cells between two adjacent layers, as represented by a darker node in **Fig. 5** between layer 1 and layer 2, can be simply determined by (9).

$$\frac{2}{k_{eff}} = \left(\frac{1}{k_1} + \frac{1}{k_2}\right) \quad (9)$$

where k_1, k_2 represent the thermal conductivity of material 1 and 2, respectively. k_{eff} is the effective thermal conductivity between these two layers. Since thermal interface material (TIM) is applied between two different layers (**Fig. 3**) to reduce the contact resistance caused by surface roughness, we assume there is a perfect thermal contact between the TIM layer and the other material.

Due to the presence of complex geometry and the complicated boundary condition, the silicon junction temperature profile can not be solved analytically. However, a numerical solution can be found by finite difference approaches and approximation schemes. In the next section, the self-consistent electrothermally-aware methodology for estimating substrate temperature profile is proposed and implemented.

4. IMPLEMENTATION OF THE ELECTROTHERMALLY-AWARE METHODOLOGY

Several numerical approaches exist in the literature for solving partial differential equations [28, 29]. Here we use the Alternating-Direction-Implicit method [30, 31] because it is a widely used method for the efficient numerical solution of parabolic partial differential equations in multiple spatial variables. The advantage of applying this method is that we are able to transfer a multiple dimensional parabolic partial differential equation into a succession of one-dimensional problems. Therefore, no large scale matrix has to be computed and it is easy to be implemented. Thus, we use the alternating direction implicit method as the core algorithm to solve the heat partial differential equations for achieving higher computation efficiency. It is important to note that although other computationally efficient methods exist, but choosing any one of them over the others does not affect the core results of our proposed methodology. The key aspect of our self-consistent approach is that it inherently generates a more accurate power profile (**Fig. 6**), which can then be used to generate a temperature profile using any computationally efficient PDE solvers.

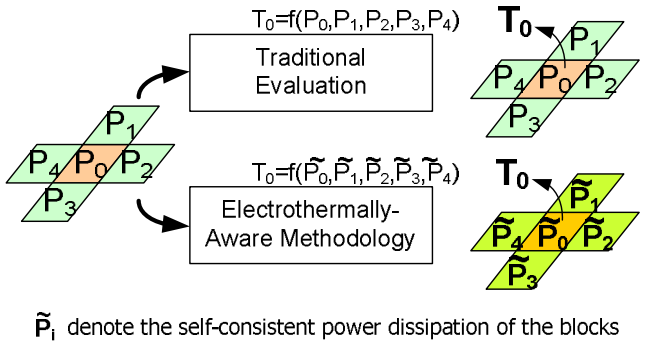

Figure 6. Key aspect of the electrothermally-aware methodology. Due to the strong interdependence of temperature and leakage power, temperature at the center block (T_0) is not simply a function of the power dissipation within and adjacent to the center block as per traditional analysis. Nominal power distribution map should be updated self-consistently with the temperature evaluation.

The overview of the proposed electrothermally-aware substrate temperature gradient evaluation methodology is illustrated in **Fig. 7**. The chip (target simulation domain) is partitioned into a mesh according to the information provided by the layout geometry (position) and power distribution map. Nominal power distribution (including switching and leakage power dissipation) for each functional block according to its activity factor is used as initial values depending on the circuit implementation and technology nodes. Physical parameters such as specific heat, thermal conductivity and heat transfer coefficient depend on given packaging material properties and applied cooling techniques. The full-chip realistic packaging model is incorporated and comprehends both vertical and lateral heat transfer paths. Boundary conditions are determined by the operating environment. The simulator uses the layout geometry, power distribution, boundary conditions, and physical thermal parameters as initial values to formulate parabolic partial differential energy equations and then solves these equations in a self-consistent manner using the Alternating-Direction-Implicit method for every mesh element. The algorithm converts a multiple dimensional parabolic partial differential equation into a succession of one-dimensional linear problems. The electrothermal couplings are also embedded in the core of the simulator that simultaneously estimates temperature dependent quantities for each simulation step. Once the difference of the temperature evaluation between two steps is within a

certain range (e.g. *0.01°C*), 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/cooling solutions or high power dissipation, the evaluation stops and thermal runaway will be reported.

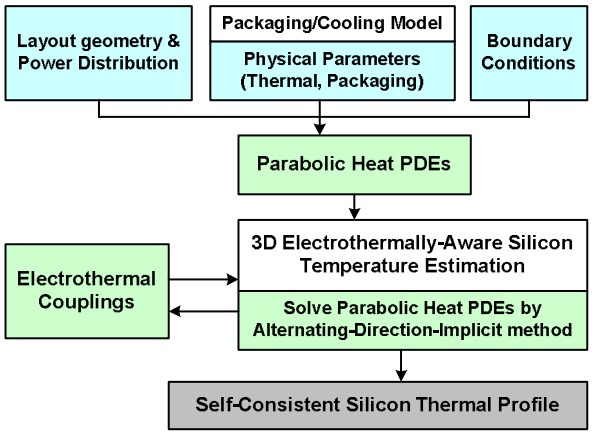

Figure 7. Overview of the electrothermally-aware silicon temperature profile simulator.

We further implemented the proposed simulator on a PC (3.06 GHz Pentium 4 processor, 1 GB memory) using C++ language. A microprocessor design with die size of 10 mm × 10 mm (discretized into 100×100 grids) and with power densities per functional block is shown in **Fig. 8**. Note that we consider the power dissipation of each block calculated according to each block's nominal switching power, leakage power, and activity factor. The nominal total power consumption of the chip at nominal temperature (*25°C*) is *96 W* (nominal active power = *93.1 W*, leakage power = *2.9 W*). The short-circuit component is relatively small; therefore we neglect it for simplicity. The physical and thermal properties of the packaging layers are similar to a practical package of a high-performance microprocessor.

The proposed methodology has significant implications for various temperature-dependent effects because of the accurate substrate temperature profile it can generate. For instance, temperature increase in the local areas (hot-spots) tends to accelerate device and interconnect failure mechanisms that are strongly temperature dependent. Also, signal integrity analysis under non-uniform substrate temperature [32] requires an accurate substrate temperature estimation. At the circuit level, buffer insertion and gate sizing (physical design process) can be made thermally aware because interconnect and gate delays are strongly dependent on temperature [33]. Besides synthesis, placement, and routing algorithms, power-grid analysis can be made thermally aware to ensure acceptable voltage-drop levels in the presence of significant chip substrate temperature gradients [8, 10].

In the next section, a comparison between different scenarios is shown to highlight the importance and the impact of employing electrothermal couplings and realistic package thermal models for substrate thermal gradient estimation. The proposed methodology inherently provides an accurate estimation of power dissipation in leakage dominant CMOS technologies. Moreover, implications for IC-cooling and thermal (hot-spot) management for nanometer scale ICs are also presented and discussed.

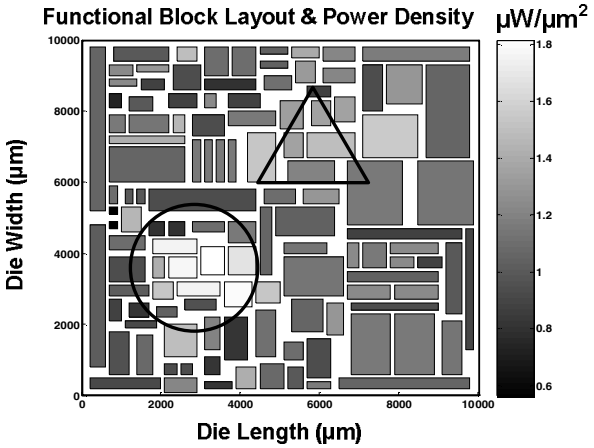

Figure 8. Functional block layout of a test chip. Power densities associated with functional blocks are also shown. The circle indicates a region where functional blocks have highest power density. The triangle indicates the functional blocks that have higher leakage power dissipation than all other blocks.

5. SIMULATION RESULTS AND IMPLICATIONS

5.1 IMPACT OF ELECTROTHERMAL COUPLINGS

First we demonstrate the importance of incorporating electrothermal couplings for estimating substrate temperature profile. Although the results are specific to the above mentioned IC, the conclusions are more generic. It can be observed that there is a region indicated by a circle in **Fig. 8** where blocks within this region have highest power density. In addition, there is a region indicated by a triangle in **Fig. 8** where blocks have *10X* leakage power dissipation with respect to the values of all other functional blocks.

The substrate temperature profile shown in **Fig. 9** is generated using traditional thermal simulator without considering electrothermal couplings. The highest temperature (hot-spot) is around 64.2°*C* and located in a region with highest power density (indicated by a circle in **Fig. 8**). However, based on the same packaging and system cooling condition (model shown in **Fig. 4(b)**), a different substrate temperature profile (**Fig. 10**) is obtained by the proposed electrothermally-aware simulator. From the temperature profile, two hot-spots can be observed--one in the region with highest power density and the other in the region with higher percentage of leakage power. Unlike traditional evaluation, the highest temperature is around *63.8°C* and is located in a region with higher percentage of leakage power (indicated by a triangle in **Fig. 8**).

As explained in Section 2, a region with higher switching power density does not necessarily yield a higher temperature due to the various electrothermal couplings. It is easy to observe the impact of electrothermal couplings on substrate temperature evaluation by comparing **Fig. 9** and **Fig. 10**. The substrate temperature profile obtained by electrothermally-aware evaluation shows an additional hot-spot and also 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.

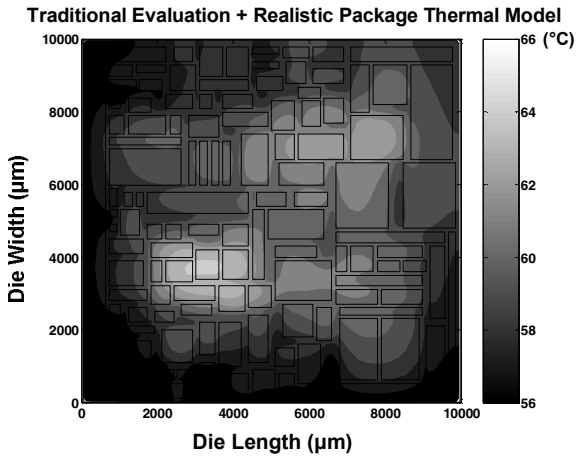

Figure 9. Substrate temperature profile generated by traditional thermal simulator without considering electrothermal couplings. The highest temperature (T_{max}) is around 64.2°C and located in the region with highest power density.

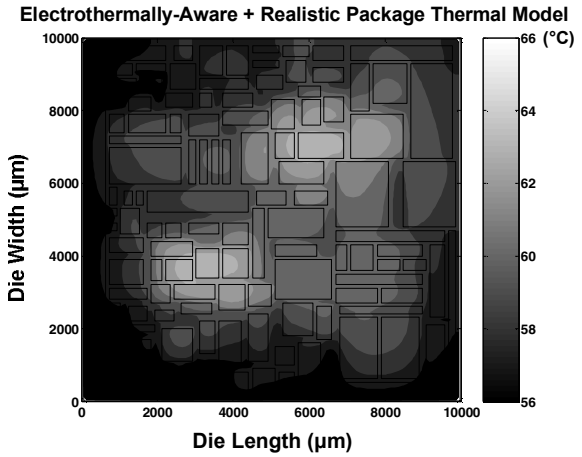

Figure 10. Substrate temperature profile generated by proposed electrothermally-aware simulator. Two hot-spot regions can be observed. The highest temperature (T_{max}) is around *63.8°C* and located in the region with higher percentage of leakage power.

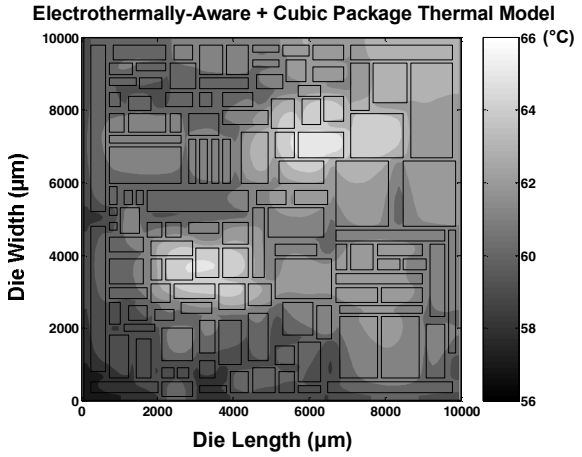

Figure 11. Substrate temperature profile generated by proposed electrothermally-aware simulator employing a cubic packaging stack-up as shown in **Fig. 4(a)**. The highest temperature (T_{max}) is around 65.7°C and located in the region with higher percentage of leakage power.

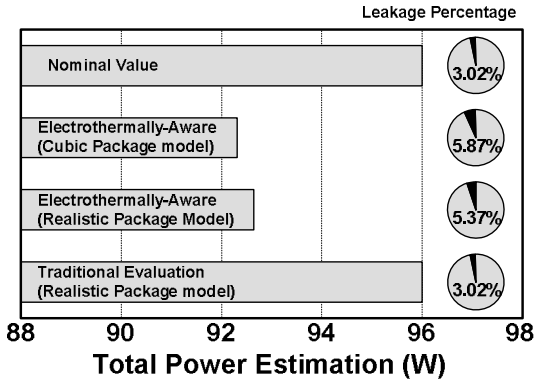

Figure 12. Estimation of total power dissipation under three different scenarios. Nominal power dissipation at *25°C* is also shown for comparison. The pie chart shows the percentage of the leakage power dissipation under each scenario.

5.2 Impact of Considering Realistic Package Thermal Model

Here we demonstrate the impact of employing two different package thermal models for the cooling path on substrate temperature profile estimation. For fair comparison, the layout, power density distribution, and discretization of the die are kept identical. Also, the physical and thermal properties of each packaging layer material are kept constant in both models.

Fig. 11 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 packaging layers (integrated heat spreader and heatsink). By comparing the two substrate temperature profiles (**Fig. 10** and **Fig. 11**), the maximum and average substrate temperature in the estimation is higher with unrealistic package thermal model. However, it is also important to notice that the temperature gradient from the hot-spot to the borders of the chip surface is higher while considering lateral heat spreading. This in turn, will impact physical design issues such as partitioning and placement schemes for high-performance microprocessors including multi-core designs.

5.3 Implications for Power Estimation & Thermal Management

Since power dissipation and temperature are strongly coupled and the coupling becomes more prominent as technology scales, it is critical to incorporate electrothermal couplings for accurate power estimation. **Fig. 12** compares the power estimation including active and leakage power under three scenarios as already shown in **Fig. 9**, **Fig. 10**, and **Fig. 11** respectively. It can be observed that both traditional evaluation and cubic (unrealistic) package thermal model lead to erroneous leakage power or total power estimation.

Fig. 13 shows the increase in maximum substrate temperature (ΔT_{max}) with an increase in leakage power consumption. 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 increase with scaling, the significance of employing the proposed methodology is expected to increase as technology scales.

Furthermore, hot-spots are known to determine system level thermal management choices since packaging and cooling solutions have to meet the maximum heat-flux requirements at the silicon-package interface. As we already show in **Fig. 13**, the two curves with different junction-to-ambient thermal resistance (θ_{ja}) have different slopes as the technology becomes leaky, i.e., impact of lowering θ_{ja} on

hot-spot temperature by packaging and cooling solutions will increase for leakage dominant technologies.

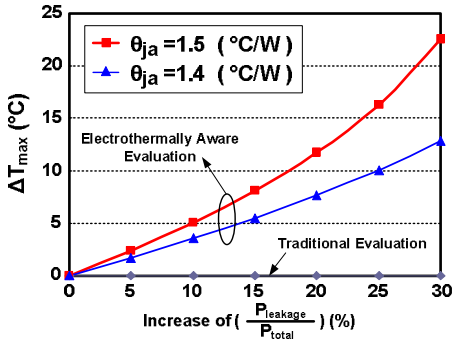

Figure 13. Increase in maximum substrate temperature (ΔT_{max}) as function of leakage power dissipation increase. $P_{leakage}$ and P_{total} denote the leakage and total power consumption respectively. The numbers in the x-axis represent the percentage increase of the ratio ($P_{leakage}/P_{total}$). ΔT_{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 to ambient. Curves for traditional evaluation and different θ_{ja} are also shown for comparison.

6. CONCLUSIONS

An electrothermally-aware substrate temperature profile estimation methodology has been introduced in this paper for leakage dominant technologies that takes various electrothermal couplings and realistic package thermal models into account. While traditional methodologies neglect electrothermal couplings and mislead hot-spot and thermal gradient evaluation, it is demonstrated that the proposed methodology provides accurate silicon substrate temperature profile with an efficient numerical approach. In addition, the significance of employing electrothermal couplings is expected to increase as technology scales. Moreover, it is shown that considering a realistic package thermal model not only improves the accuracy of estimating heat distribution and power dissipation but also has significant implications for hot-spot and thermal gradient management. As power and thermal problems increasingly impact the scalability of CMOS devices and architecture of high-performance ICs products including microprocessors, the proposed methodology will be invaluable for incorporating temperature awareness in IC design.

7. ACKNOWLEDGEMENT

This work was supported by Intel Corporation and the University of California-MICRO program. The authors would also like to acknowledge Dr. Ravi Mahajan and Dr. Greg Chrysler at Intel Corporation, Chandler, Arizona, for providing valuable technical feedback.

8. REFERENCES

[1] K. Banerjee, PhD Thesis, University of California, Berkeley, 1999.
[2] A. H. Ajami et al., "Analysis of Non-Uniform Temperature-Dependent Interconnect Performance in High Performance ICs," in *Proc. DAC*, pp. 567-572, 2001.
[3] *International Technology Roadmap for Semiconductors* (ITRS)
[4] R. S. Prasher et al., "Nano and Micro Technology-Based Next-Generation Package-Level Cooling Solutions" *Intel Technology Journal 4th quarter*, 2005.
[5] K. Banerjee et al., "The Effect of Interconnect Scaling and Low-K Dielectric on the Thermal Characteristics of the IC Metal," in *Proc. IEDM*, 1996, pp. 65-68.
[6] S. Im and K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs," in *Proc. IEDM*, 2000, pp. 727-730.
[7] K. Banerjee and A. Mehrotra, "Global (Interconnect) Warming," *IEEE Circuits Devices Magazine*, Vol. 17, pp. 16–32, 2001.
[8] A. H. Ajami et al., "Scaling Analysis of On-Chip Power Grid Voltage Variations in Nanometer Scale ULSI," *Journal of Analog Integrated Circuits and Signal Processing*, Vol. 42, pp. 277-290, 2005.
[9] J. Lee, "Thermal Placement Algorithm Based on Heat Conduction Analogy," *IEEE Trans. Components and Packaging Technologies*, Vol. 26, pp.473-482, 2003.
[10] B. Goplen and S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs Using a Force Directed Approach," in *Proc. ICCAD*, 2003, pp. 86-89.
[11] C. N. Chu and D. F. Wong, "A Matrix Synthesis Approach to Thermal Placement," in *Proc. ISPD*, 1997, pp. 163–168.
[12] S-C. Lin et al., "Analysis and Implications of IC Cooling for Deep Nanometer Scale CMOS Technologies," in *Proc. IEDM*, 2005, pp. 1041-1044.
[13] Y. Cheng et al., "A Chip-Level Electrothermal Simulator for Temperature Profile Estimation of CMOS VLSI Chips," in *Proc. ISCAS*, 1996, pp. 580-583.
[14] Y. Cheng et al., "ILLIADS-T: An Electrothermal Timing Simulator for Temperature-Sensitive Reliability Diagnosis of CMOS VLSI Chips," *IEEE Trans. on Computer-Aided Design (TCAD)*, Vol. 17, pp. 668-681, 1998.
[15] T. Wang et al., "3-D Thermal-ADI: A Linear-Time Chip Level Transient Thermal Simulator," *IEEE Trans. on Computer-Aided Design (TCAD)*, Vol. 21, pp. 1434-1445, 2002.
[16] T. Wang et al., "Thermal-ADI - A Linear-Time Chip-Level Dynamic Thermal-Simulation Algorithm Based on Alternating-Direction-Implicit (ADI) Method," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, Vol. 11, pp. 691-700, 2003.
[17] P. Li et al., "Efficient Full-Chip Thermal Modeling and Analysis," in *Proc. ICCAD*, 2004, pp. 319-326.
[18] Y. Zhan and S.S. Sapatnekar, "A High Efficiency Full-Chip Thermal Simulation Algorithm," in *Proc. ICCAD*, 2004, pp. 634-637.
[19] Z. Yu et al., "Fast Placement-Dependent Full Chip Thermal Simulation," in *Proc. VLSI-TSA*, 2001, pp. 249-252.
[20] W. Huang et al., "Compact Thermal Modeling for Temperature-Aware Design," in *Proc. DAC*, 2004, pp. 878-883.
[21] S. Borkar et al., "Parameter Variations and Impact on Circuits and Microarchitecture," in *Proc. DAC*, 2003, pp. 338-342.
[22] S. Narendra et al., "Full-Chip Sub-Threshold Leakage Power Prediction Model for Sub-0.18µm CMOS," in *Proc. ISLPED*, 2002, pp. 19-23.
[23] P. Gelsinger, 41st Design Automation Conference (DAC) Keynote, 2004.
[24] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge Univ. Press, 1998.
[25] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," in *Proc. ISLPED*, 1999, pp. 163-168.
[26] K. Banerjee et al., "A Self-Consistent Junction Temperature Estimation Methodology for Nanometer Scale ICs with Implications for Performance and Thermal Management," in *Proc. IEDM*, 2003, pp. 887-890.
[27] S. Im et al., "Scaling Analysis of Multilevel Interconnect Temperatures for High Performance ICs," *IEEE Transactions on Electron Devices (TED)*, Vol. 52, pp. 2710-2719, 2005.
[28] M. N. Özişik, *Boundary value problems of Heat Conduction*, Dover Publications, 2002.
[29] R. Haberman, "Elementary Applied Partial Differential Equations with Fourier Series and Boundary Value Problems," Prentice Hall, 1983.
[30] D. W. Peaceman and H. H. Rachford, "The Numerical Solution of Parabolic and Elliptic Differential Equations," *Journal of the Society for Industrial and Applied Mathematics (SIAM)*, pp. 28-41, 1995.
[31] J. Douglas and H. H. Rachford, "On the Numerical Solution of Heat Conduction Problems in Two or Three Space Variables," *Trans. American Mathematical Society*, pp. 421-439, 1956.
[32] A. H. Ajami et al., "Effects of Non-Uniform Substrate Temperature on the Clock Signal Integrity in High Performance Designs," in *Proc. CICC*, 2001, pp. 233-236.
[33] A. H. Ajami et al., "Analysis of Substrate Thermal Gradient Effects on Optimal Buffer Insertion," in *Proc. ICCAD*, 2001, pp. 44-48.