Power Blurring: Fast Static and Transient Thermal Analysis Method for Packaged Integrated Circuits and Power Devices

Amirkoushyar Ziabari, *Member, IEEE*, Je-Hyoung Park, Ehsan K. Ardestani, Jose Renau, *Associate Member, IEEE*, Sung-Mo Kang, *Fellow, IEEE*, and Ali Shakouri, *Member, IEEE*

***Abstract*— High-temperature and temperature nonuniformity in high-performance integrated circuits (ICs) can significantly degrade chip performance and reliability. Thus, accurate temperature information is a critical factor in chip design and verification. Conventional volume grid-based techniques, such as finite-difference and finite-element methods (FEMs), are computationally expensive. In an effort to reduce the computation time, we have developed a new method, called power blurring (PB), for calculating temperature distributions using a matrix convolution technique in analogy with image blurring. The PB method considers the finite size and boundaries of the chip as well as 3-D heat spreading in the heat sink. PB is applicable to both static and transient thermal simulations. Comparative studies with a commercial FEM tool show that the PB method is accurate within 2%, with orders of magnitude speedup compared with FEM methods. PB can be applied to very fine power maps with a grid size as small as 10 μm for a fully packaged IC or submicrometer heat sources in power electronic transistor arrays. In comparison with architecture-level thermal simulators, such as HotSpot, PB provides much more accurate temperature profiles with reduced computation time.**

***Index Terms*— Finite-element method (FEM), heat transfer, integrated circuits (ICs), package, power electronics, temperature, thermal management, thermal simulation.**

I. Introduction

IN RECENT years, the scaling of supply voltage has departed from the ideal scaling predicted in [1] and [2]. The threshold voltage ($V_{\rm th}$) has stopped scaling. This in turn has stopped the scaling of supply voltage ($V_{\rm dd}$) to maintain circuit performance. This results in higher power densities, which promotes temperature as one of the primary design parameters as the scaling advances to smaller technology sizes.

In addition, nonuniform activities in an integrated circuit (IC) chip yield nonuniform surface temperature distributions, since localized heating occurs much faster than chip-wide heating [3]. Thus, the temperature of certain regions can become much higher than those in the neighboring areas, which are called hotspots. Hotspots and spatial temperature gradients in VLSI ICs have become critical issues due to their limiting effects on both the performance and the reliability of IC chips in packages [4]. In transistor arrays used in power electronics, nonuniform temperature distribution affects device characteristics and the switching behavior.

Moreover, the increasing leakage power and its exponential dependence on the temperature require more attention to thermal-aware simulations and optimizations. Hence, precise estimation of temperature distribution is essential for accurate analysis of performance, reliability, and power management.

Generally, thermal simulations and design optimizations are done under steady-state worst case conditions due to the high computational cost, causing reliance on the use of conservative margins in thermal designs. However, the temperature nonuniformity evolves over time and thus hotspots are of spatiotemporal nature. The transient temperature spike or localized heating also can cause timing errors, nonuniform current flow, or reliability failures. As the thermal budget becomes increasingly tight, the worst case approach becomes too costly and ineffective. It is also known that the simulated worst case peak power and its corresponding peak temperature are rarely observed [3], [5].

Even with the state-of-the-art tools, the chip-level transient thermal simulation with a realistic package configuration is too costly for physical design optimization or performance verification in the packaged environment. In addition, in the early stages of chip design, i.e., architectural specification stage, specific package information, and thermal boundary conditions may not be available. At these stages, designers rely on simulation to consider the tradeoffs. However, slow simulation limits the scope of analysis. For these reasons, a fast thermal analysis method is highly desired [6]. In this paper, we present a fast, yet accurate steady-state and transient temperature computation method suitable for VLSI ICs in packages. To validate the new method, called power blurring (PB), the simulation results are compared with those of a

Manuscript received June 1, 2013; revised September 26, 2013; accepted November 10, 2013. Date of publication January 30, 2014; date of current version October 21, 2014. A. Ziabari and J.-H. Park contributed equally to this paper.

A. Ziabari is with the Department of Electrical and Computer Engineering and the Birck Nanotechnology Center, Purdue University, IN 47907 USA (e-mail: aziabari@purdue.edu).

J.-H. Park is with Samsung Electronics, Hwaseong 445-330, Korea (e-mail: jh2010.park@samsung.com).

E. K. Ardestani and J. Renau are with the Department of Computer Engineering, University of California Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: eka@soe.ucsc.edu; renau@soe.ucsc.edu).

S.-M. Kang is with the Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: kangsm@kaist.ac.kr).

A. Shakouri is with the Department of Electrical and Computer Engineering and the Birck Nanotechnology Center, Purdue University, IN 47907 USA (e-mail: shakouri@purdue.edu).

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/TVLSI.2013.2293422

finite-element analysis (FEA) software as well as HotSpot, an architectural level simulator [7], [8]. The evaluation results demonstrate PBs high speed and accuracy. PB calculates the temperature profile of the ICs at least an order of magnitude faster than other methods, while maintaining the accuracy within 2%.

It is well known that the electrical characteristics of many devices are a function of their temperature. Thus, it is necessary to perform self-consistent electrothermal simulations. Temperature profiles from an initial power distribution map can be used as an input parameter to recalculate the power dissipation in individual devices, which in turn can be used to calculate a refined power dissipation profile and temperature distribution. This process can be repeated until no more refinement is observed. Such iterative solutions are particularly important in deep submicrometer devices, in which the leakage power or the subthreshold characteristics are a strong function of the local temperature, as well as in power electronic transistor arrays. Although we do not handle the coupled electrothermal simulations to calculate the power consumption profile with realistic electrical input data in this paper, the fast method to calculate the temperature profile from an input power map in a realistic package (i.e., PB technique) is a key ingredient for full self-consistent simulations.

The remainder of this paper is organized as follows. In Section II, the PB method is discussed in detail. The related works are also explained in this section, and a qualitative comparison with the PB method is presented. The methodologies are explained in Section III. In Section IV, the simulation results for different static power maps are described, and a quantitative comparison between the results obtained with PB and some conventional methods is presented. Transient simulation methodology based on the PB method and exemplary simulation results are presented in Section V, followed by a conclusion in Section VI.

II. PB Method

A. Available Methods and Their Limitations

To obtain thermal profiles for a region of interest, the heat diffusion equation shown in (1) has to be solved with conditions imposed on boundaries [9]

$$k\frac{\partial^2 T}{\partial x^2}+k\frac{\partial^2 T}{\partial y^2}+k\frac{\partial^2 T}{\partial z^2}+q*=\rho c_p\frac{\partial T}{\partial t} \tag{1}$$

where k is the thermal conductivity (W/m.K), ρ is the density (kg/m^3), c_P is the specific heat (J/kg.K), $q*$ is the heat generation per unit volume (W/m^3), and T is the temperature of the location (x, y, z) at time t. In IC thermal analysis, the heat diffusion equation has been conventionally handled by grid-based methods, such as the finite-difference method (FDM) or the finite-element method (FEM), which generate 3-D volumetric meshes of solid structure [10]. The accuracy of FDM and FEM comes at a price of long execution time, exhaustive CPU and memory usage. Since the computation time increases superlinearly with the number of meshed elements, these approaches are impractical to be integrated into an interactive place-and-route IC design program. In addition,

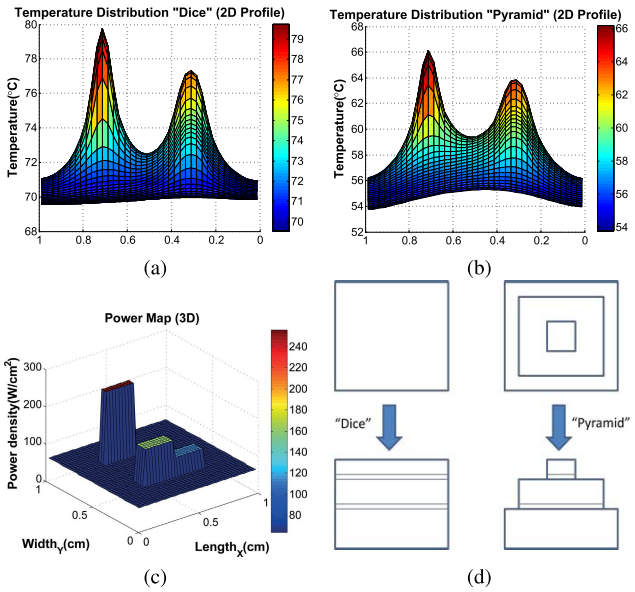


Fig. 1. Effect of the thermal mount geometry on the shape and the average value of temperature profile. 2-D temperature profile obtained by the (a) dice and (b) pyramid geometries. (c) Applied power density map on top surface of Si die. (d) Dice and pyramid geometries.

FEA programs require a thorough design of the meshes to attain convergence, a procedure that cannot be easily automated for a complex geometry or power dissipation loads.

A Green's function-based method was introduced for thermal analysis as an alternative [6], [10]. This method has an advantage over FEM or FDM due to its lower dependency on the volumetric mesh. However, the analytical expression of the Green's function could be found only for simple 1-D planar geometries and requires an infinite series to be accurate. The Green's function for multilayer structures has been calculated analytically by several groups and is readily available online. For example, in [11], a fast algorithm for calculating temperature profiles based on analytical Green's function of the bounded box is presented. However, most of these techniques fail to handle the real geometry of the heat spreader underneath the chip. They assumed the same dimensions for both the heat spreader and the silicon die. However, in reality, the heat spreading layer closer to a heat source has a smaller surface area than the one farther away, similar to a reversed pyramid structure. Neglecting the real heat spreading to the heat sink can result in overestimation of the chip temperature. In addition, the shape of the temperature profile is also affected significantly by the size of the heat sinks. Whereas the average chip temperature can be scaled by an appropriate scaling of the convection coefficient, changing the temperature profile is not straightforward [10]. Fig. 1 shows how the geometry of the package can affect the temperature profile. In this figure, the temperature fields, shown in Fig. 1(a) and (b), result from applying the power density map shown in Fig. 1(c) on two geometries drawn in Fig. 1(d). The dimensions of the two geometries are the same as those shown in Table I. The geometries contain an Si die on the top, a heat spreader in the middle (Cu), a heat sink in the bottom (Cu), and two layers of thermal interface materials (TIMs) between the layers to bind them together. This structure is discussed in

TABLE I
MATERIAL PROPERTIES AND DIMENSIONS OF THE PACKAGE MODEL [12]

	Area (mm^2)	Thickness (mm)	Thermal Conductivity (W/m-K)	Density (kg/m^2)	Specific Heat (J/kg-K)
Si Die	10×10	0.775	117.5	2330	700
TIM1	10×10	0.025	5.91	1930	15
Heat Spreader	28×28	1.8	395	8933	397
TIM2	28×28	0.025	3.5	1100	1050
Heat Sink	60×80	6	395	8933	397

more detail in Section II-C. As shown in Fig. 1(d), the two geometries are different in that: in the dice geometry, all the layers have the same size as the silicon die, whereas in the pyramid geometry, the heat spreader and the heat sink have about 10 and 50 times larger surface than the chip. Perfect convection is assumed on the top surface of the heat sink so that the total convection remains the same for both cases. Even with this assumption the average temperature will be significantly different, since the mounts are different in shape. Bagnoli *et al.* [13] has managed to find an analytical system of equations that relates the temperature and heat flux at the material interfaces of the pyramid structure. These equations can be discretized and solved by the usual matrix inversion. While this technique can be applied to the pyramid multilayer structure, it still requires significant computational resources. The main drawback of tackling the more realistic pyramid geometry is that a simple analytic expression no longer exists and thus requires analytical methods, where a part of the algorithm must rely on either the empirical parameters or previous simulations for reasonable speed.

Acceleration techniques for grid-based methods are introduced in [14] and [15], in which acceleration was achieved by either decomposing a multiple dimensional problem into 1-D problems or reducing the thermal network. However, both failed to include a realistic package model. Ignoring the lateral heat spreading in realistic packages can result in an overestimated chip temperature [16]. In [18], a thermal circuit composed of thermal resistors and capacitors was built based on duality between the heat transfer and the electrical circuit. In the model, a secondary heat transfer path including the interconnect layers and the path between I/O pads and printed-circuit board were included. However, the heat spreader and the heat sink are too coarsely meshed to represent 3-D heat spreading accurately in those regions. Using a simple model for packages can introduce a significant error. In Fig. 2, using the results obtained by ANSYS, it is shown that neither the temperature profile in the bottom surface of a silicon die nor the heat spreader thermal profiles are uniform for a given structure with a given power map. This implies that thermal packages cannot be represented by a simple thermal resistor. In Fig. 2(d), one can also observe that the peak temperature on top of the heat spreader is shifted toward the center due to 3-D heat spreading at the corners of the chip. This cannot be accurately considered by 1-D models assuming that the heat sink and heat spreader of the Si die the same size. To achieve higher accuracy, more elaborate methods have been

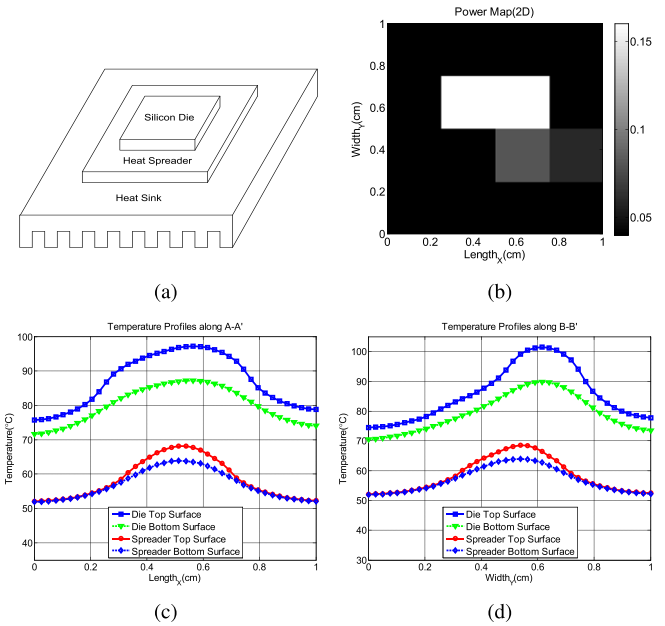


Fig. 2. Thermal profiles in a flip-chip thermal package. (a) Picture of the package. (b) Power dissipation map. (c) and (d) Corresponding thermal profiles in the silicon die and the heat spreader along two arbitrary paths. A–A' is along a diagonal of the die, whereas B–B' is along a vertical line through the center of the die.

developed but they have become as complicated as the grid-based techniques.

In [17], the diffusive representation (DR) is used to construct a compact thermal model as a state-space representation. The method is dedicated to the representation of nonrational systems based on infinite contributions. A discrete formulation of the DR yields a practical engineering modeling method. DR enables modeling of a given system, provided that the temperature to be monitored is observable for the purpose of prior validation. The main drawback of this method is that it cannot construct geometry-dependent models. The formal construction of thermal compact models from the geometry description of the system is limited to regular and simple shapes.

HotSpot is a widely used simulator for staic and dynamic thermal analysis in IC architecture community [7]. We have shown in [19] that PB can surpass HotSpot by an order of magnitude in both speed and accuracy terms. In Sections III and IV, a detailed comparison of HotSpot with the PB method, in both static and transient thermal analysis, is presented. Han *et al.* [20] suggest improvements to the thermal solver of HotSpot. However, the model used is only for architectural-level thermal simulator and is not able to capture the device-level thermal profile of ICs. The speedup is possible only for thermal simulation at architecture block level. Therefore, the accuracy of the model is at the same level as the HotSpot simulator in block-level mode. In the following sections, we present the accuracy and speed of PB compared with HotSpot at the grid level.

Yang *et al.* [21] introduced ISAC for static and dynamic thermal analysis. The simulator provides one to two orders of magnitude speedup compared with the COMSOL finite-element software [22] for static thermal analysis, while the error is within 3% on average. In Section IV, our evaluation

results show that PB outperformed ISAC by providing three orders of magnitude speedup over the FEM softwares. In addition, using the same error equation defined in [21], the average error of PB is under 0.7%. Using (10), which is independent of the temperature unit, the average error for the two cases of ISAC and PB would be about 6% and 2.5%, respectively.

One should note that both HotSpot and ISAC are implemented in C++ while PB is implemented in MATLAB. The implementation of PB in C++ or a similar programming language would yield even more speedup. In addition, speed and accuracy are not the only factors that distinguish PB. It is shown in [23] and [24] that PB as a fast and accurate method is applicable for vertically 3-D ICs (3-DICs) including thermal vias. In [25], PB employed to obtain temperature profiles of power electronic array transistors. Solving complex nonlinear problems has been tackled by PB. The PB technique is based on the superposition principle, which requires linearity of the heat conduction equation. An adaptive PB technique is developed in [26], which can solve nonlinear problems, e.g., when the thermal conductivity of the silicon is modified based on the local temperature of the chip, using two or three iterations. Excellent agreements with self-consistent finite-element simulations have been obtained. Another unique feature of PB is that it can be applied to the inverse problem [27], i.e., extracting power map from temperature profile. This is possible since PB is essentially an image blurring technique. Advanced image deblurring algorithms enable solving the ill-posed inverse problem in many configurations. These points illustrate that the versatility of PB is also a significant trait in addition to its speed and accuracy, which distinguishes PB from other available methods.

In the following section, we will describe the PB method applied to accurately predict the static and dynamic temperature profiles of IC chips and a power transistor array.

B. Background of PB

We developed a matrix convolution technique, called *PB*, to expedite the computation of temperature distribution in IC chips. The PB method has its theoretical basis on the Green's function method. Implementation is similar to image blurring used for image processing.

The Green's function method finds a solution to the partial differential equation with a point source as the driving function in the first step ($G(r, r')$). This solution is called the Green's function, which is equivalent to the impulse response of a system. Subsequently, a solution to an actual source is represented as a superposition of the impulse responses to the point sources at different locations [28]. This is expressed in (2), in which V is the volume over which the heat $q(x', y', z')$ is generated

$$\iiint_V G(r, r')q(x', y', z')dv'. \tag{2}$$

Thus, the Green's function is used as a building block for constructing an actual solution.

In image processing, an image is blurred by a convolution with a filter mask. The filter mask is a matrix whose elements define a process in which the modification (i.e., blurring) of an image occurs. For instance, an image, f, is convoluted with a filter mask, w, to produce a blurred image g by

$$g(x, y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s, t) f(x + s, y + f) \tag{3}$$

where $a = (m-1)/2$ and $b = (n-1)/2$ for a $m \times n$ mask [29].

In thermal analysis, the impulse response (i.e., Green's function) of a system corresponds to a heat-spread function, namely its thermal mask. The thermal mask represents the amount of temperature rise that occurs in a solid due to a unit point heat source. A 2-D spatial distribution of heat dissipation in an IC chip is called power map. The power map for a given IC chip can be estimated using voltages and currents in each device or in circuit blocks [30]. If one thinks of the power map as an image, f, the thermal profile of the IC chip resulting from its power map can be regarded as a blurred image, g, when the filter mask, w, conforms to the thermal mask.

In principle, once the thermal mask is obtained for a given package assembly, a temperature profile can be easily obtained by a simple matrix convolution for an arbitrary power distribution map. Table II summarizes the analogy between the image blurring and PB methods.

TABLE II
ANALOGY BETWEEN IMAGE PROCESSING AND PB

Image Processing	Power Blurring
Image (f)	Power Map
Filter Mask (w)	Impulse Response
Blurred Image (g)	Temperature Profile

C. Full-Chip Package Model

Fig. 2(a) shows a simplified thermal model of a typical flip-chip package structure, which is composed of three main components: 1) Si die; 2) the heat spreader (Cu); and 3) heat sink (Cu). Heat spreader improves heat transfer between the silicon die and the heat sink. A TIM serves as a bonding material and also enhances thermal conductivity at the interface [31]. The bottom layer of the die, where devices are fabricated, is very thin compared with the substrate of the die. Hence, the die is considered a bulk silicon. ICs have passivation layers between interconnects [32], and thus the thermal resistance of minor heat transfer path is much higher than that of the major heat transfer path (through the substrate). Most of the heat is assumed to be transferred into the ambient (35 °C) by conduction and convection along the major heat transfer path. Other minor heat transfer paths are neglected in this paper. To this end, adiabatic boundary conditions are imposed on four sides and bottom surface of the Si die. The material properties and dimensions are listed in Table I. The die thickness is assumed to be 0.775 mm as in most of the chips in the 90-nm copper CMOS technology.

D. Thermal Mask

As mentioned earlier, the thermal mask is an impulse response of a system in the space domain. According to the Green's function method, we can build a solution to a partial

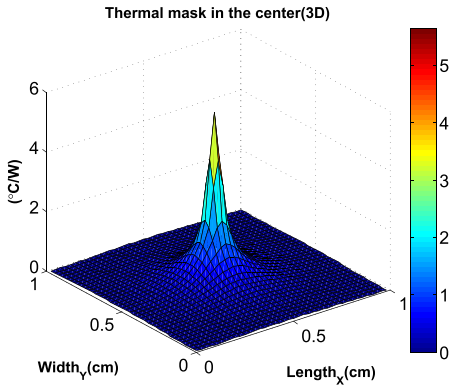


Fig. 3. Thermal mask. The surface of the Si die is discretized into 40×40 grids.

differential equation with an arbitrary driving function once we have a solution corresponding to an impulse (a point source). In the thermal analysis of IC chips, temperature distribution is the physical quantity of interest, and heat (i.e., power consumption in ICs) is the driving function. Thus, the thermal mask conforms to a steady-state temperature distribution induced by a point heat source, which is applied to the center of the Si die. In practice, the die area is discretized into grids and an approximate delta function simulating a point heat source is applied to the center element of the grid. Subsequently, the difference between the resulting and the ambient temperatures (net temperature rise due to the heat source applied) is normalized with respect to the amount of the input power. Although the thermal mask can be obtained in analytical form for a simple 1-D geometry, measurement, or a 3-D FEA simulator, such as ANSYS [33], is needed for a realistic structure, as shown in Fig. 2, where heat spreading plays an important role. Fig. 3 shows the thermal mask for the given package model. Its units are in thermal resistance (°C/W), hence the thermal mask generates a temperature profile when convoluted with a power distribution map (W).

However, when the thermal mask shown in Fig. 3 is convoluted with a power map, the direct convolution result is far from the correct thermal profile. The shape of the thermal mask depends on the location of the point heat source. In other words, point heat sources at the corners and along the edges of the Si die surface produce different temperature profiles and, therefore, different thermal masks (Fig. 4). For those regions with proximity to the boundaries, the thermal mask shown in Fig. 3 is not appropriate to be used in the convolution. This source-location dependence of the thermal mask prevents us from using a single thermal mask for the convolution with a power map.

In [12], the surface area was divided into three different regions: 1) center; 2) edge; and 3) corner (see Fig. 4). Multiple thermal masks corresponding to different regions were obtained by the repeated execution of FEA simulations. In addition, interpolated masks were used at boundary regions to handle abrupt changes among the thermal masks. Thus, at least three FEA simulations are required to characterize an arbitrary package and calculate the temperature profile for any input power distribution. The estimation errors are in the order

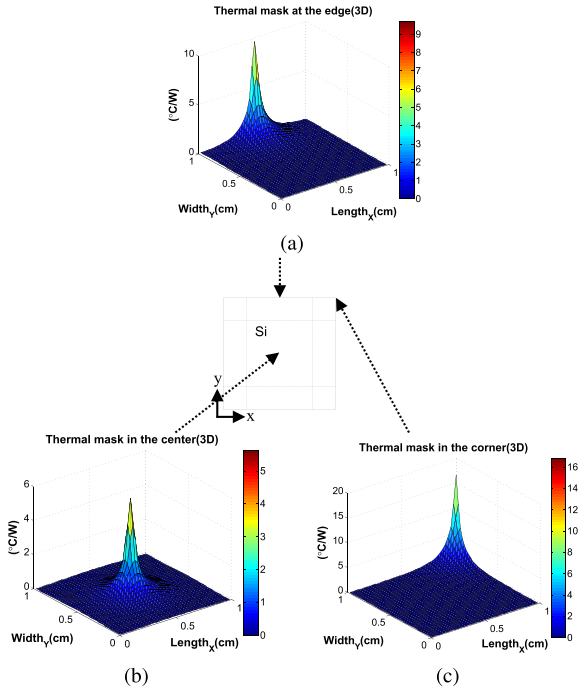


Fig. 4. Three different shapes of the thermal masks depending on the location of a point heat source. (a) Edge. (b) Center. (c) Corner.

of a few percents and they occur mainly at the boundaries between the different regions. For simplicity, it is desirable to avoid dividing the chip into arbitrary regions at the edges and corners. In the following section, a method to reduce errors from a single thermal mask will be discussed in detail.

E. Error Reduction Steps

It is desirable to use a single thermal mask (Fig. 3) to obtain the temperature profile. This temperature profile can be very inaccurate compared with the correct temperature map of the IC. The discrepancy is attributed to the two main causes. First, the regions close to the edges produce a higher temperature profile than those far from the edges in response to a single point heat source for the same amount of heat. Thus, if the region of interest is far from the edges (e.g., center region), proximity to the boundaries is of no concern when the impulse response is acquired. However, due to the finite dimensions of the Si die and our interest in the thermal profile of the whole surface area, the shape of the impulse response subjected to the edge effect is important. Second, 3-D heat spreading in the package (i.e., heat spreader and heat sink) is another source of discrepancy. Our package model has a pyramid shape (a realistic package model), in which each layer has different lateral dimensions, as shown in Fig. 2. Unlike the thermal model of a cubic package, where each layer has the same lateral dimensions, the realistic package model considers heat spreading. 3-D heat spreading significantly influences the thermal profile [34] and requires an additional compensation in the postprocessing of the convolution result.

To resolve this discrepancy, two additional processes are introduced into the direct convolution: 1) the method of image and 2) intrinsic error compensation.

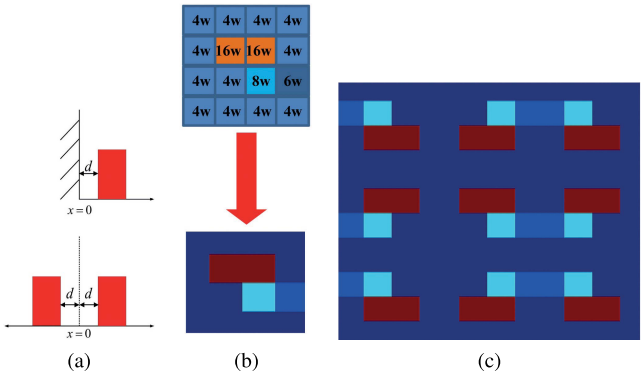


Fig. 5. Method of image. (a) Equivalence of the adiabatic boundary and the mirrored image of a heat source. (b) Converting a power map to an image. (c) Mirror replicating of the image at the boundaries (method of image).

1) Method of Image: Thermal problems are often converted into electrical problems based on the duality to take advantage of well developed methods of analysis in electrical systems. One of these is the method of image from the electromagnetism [6]. Electromagnetic problems associated with a planar perfect electric conductor can be solved through the method of image in which the boundary is replaced by the mirror image (equivalent) sources with appropriate signs [28], [35].

A similar principle can be applied to the thermal problems involving adiabatic boundary conditions. Consider the case of a heat source located at a distance, d, from an adiabatic surface, as shown in Fig. 5(a). No heat transfer can occur at the adiabatic boundary. This means the net outgoing heat flux at $x = 0$ and $x = L$ is zero. If we replace the adiabatic boundary with an image source, the adiabatic boundary condition is satisfied. The problem then becomes more manageable with this approach.

Since majority of the heat is dissipated through the heat spreader and the heat sink, the heat flux in the plane of the silicon die (perpendicular to the edges) is negligible. The net outgoing heat flux in the plane of the silicon die at the edges and corners of the chip is zero. Thus, in our thermal package model, adiabatic boundary conditions are imposed on the four sides of the Si die

$$\left.\frac{\partial T}{\partial x}\right|_{x=0} = 0; \quad \left.\frac{\partial T}{\partial x}\right|_{x=L} = 0$$
$$\left.\frac{\partial T}{\partial y}\right|_{y=0} = 0; \quad \left.\frac{\partial T}{\partial y}\right|_{y=W} = 0. \tag{4}$$

Here, T is the thermal profile, point (0, 0) is assumed to be the bottom corner on the top surface, and W and L are the width and the length of the die. As a result, the power distribution on the Si die can be extended with mirror images [Fig. 5(b) and (c)]. A single thermal mask, as shown in Fig. 3, can then be employed for the direct convolution and the edge effect is simultaneously considered. This is called direct convolution with method of image (DCMI). The disadvantage is that the size of the power dissipation matrix is now nine times bigger [Fig. 5(c)]. However, since the spatial convolution could be done very fast in the Fourier domain, this is not a factor, which will significantly increase the PB computation time

$$\begin{aligned} P_{\text{image}} = [&P(x,y) + P(x,-y) + P(-x,y) + P(-x,-y) \\ &+ P(x,2W-y) + P(-x,2W-y) + P(2L-x,y) \\ &+ P(2L-x,-y) + P(2L-x,2W-y)]. \end{aligned} \tag{5}$$

Here, P_{image} is the extended power map, P is the power map, and W and L are the width and the length of the power map, respectively. P_{image} extends from $(-W, -L)$ to $(2W, 2L)$, which is 9 times larger than P. Even though the new matrix is nine times larger, mathematically, only two-third of this matrix is needed for the final temperature calculation [36].

The direct convolution result is given by

$$T(x,y) = h(x,y) * P_{\text{image}}\left[-\frac{L}{2} : \frac{3L}{2}, -\frac{W}{2} : \frac{3W}{2}\right] \tag{6}$$

where T is the thermal profile and h is the thermal mask. The term in brackets is the portion of the extended power map needed for the final temperature calculation. In the post-process, the center region of the thermal profile corresponding to the original power map needs to be retrieved.

2) Intrinsic Error Compensation: For an accurate IC temperature calculation, the thermal profile obtained by the DCMI requires an additional step, namely the intrinsic error compensation. Due to the pyramid structure of our thermal package model, 3-D heat spreading plays an important role in heat transfer. In spite of the adiabatic boundary conditions imposed on the four sides and the bottom surface of the die, those regions along the die perimeter have better chances of heat removal than the center region. This situation can be more clearly represented with an example of a uniform power distribution.

Consider a uniform power distribution, as shown in Fig. 6(a). The thermal profile corresponding to the power distribution should be bell shaped. An FEA simulation result predicts such a profile, as shown in Fig. 6(b). However, DCMI generates a uniform temperature profile, as shown in Fig. 6(c). The temperature difference between Fig. 6(b) and (c) is shown in Fig. 6(d). This deviation is an artifact of the method of images, since this method does not consider the larger size of the heat spreader and the heat sink with respect to the silicon die. When the method of image is applied to the uniform power map, the resulting power map is another uniform power map with enlarged dimensions.

Although the thermal mask (the impulse response) is obtained for a 3-D geometry, the convolution is processed in 2-Ds. The effect of 3-D heat spreading is not appropriately handled with DCMI. Thus, the temperature deviation along the perimeter of the die is intrinsic to DCMI, and these intrinsic errors need to be compensated to obtain the final result.

The temperature rise due to uniform power input is linearly proportional to the input power level. Therefore, the intrinsic error in Fig. 6(d) is a linear function of the input power level. On the other hand, the relative deviation given in the following equation is a constant function regardless of the input power level:

$$E_r = \frac{T_{\text{DCMA}} - T_{\text{ANSYS}}}{T_{\text{ANSYS}}} \tag{7}$$

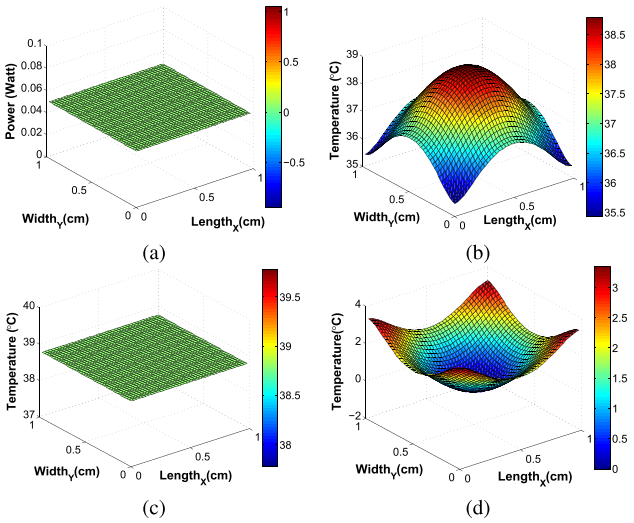


Fig. 6. Intrinsic error. (a) Uniform power map. (b) Thermal profile by ANSYS. (c) Thermal profile by the PB method. (d) Temperature deviation.

where T_{DCMI} and T_{ANSYS} are the thermal profiles for a uniform power map obtained by DCMI and ANSYS, respectively. Both T_{DCMI} and T_{ANSYS} are representing the difference between the temperature profile of the chip and the ambient temperature. E_r can serve as a position-dependent scaling factor, and it can be employed for any kind of power dissipation profile. After the calculation of E_r, the final temperature can be obtained using the following formula:

$$T_{\mathrm{final}} = \frac{T_{\mathrm{DCMI}}}{1 + E_r} + T_{\mathrm{ambient}}. \tag{8}$$

Equation (8) obviously compensates for the error of the method of image.

III. Evaluation Methodology

A. ANSYS Simulation Setup

In the following sections, ANSYS FEM software has been used to validate the PB method. Therefore, it is necessary to dedicate a part of this paper to the ANSYS simulation setup. Steady-state and transient thermal simulations are carried out in ANSYS APDL using Solid70 elements. Material properties and dimensions are based on Table I for the first static and transient case studies, and based on Table IV for the second static case study. The IC is meshed uniformly with various sizes in different case studies and the rest of the geometry is swept accordingly. In all of the simulations, the ambient temperature is set to 35 °C. The convection coefficient in the bottom surface of the heat sink is determined based on the convection resistance of the package. The convection resistance for the packages used in this paper is 0.1 K/W. In the first package (Table I), the heat sink surface area is 48 cm^2, and this value is 36 cm^2 for the second package (Table II). Using 9 and the aforementioned values, convection coefficients are 0.2083 and 0.2778 $\mathrm{W/cm}^2$-K for the first and second packages, respectively. In this equation, R is the convection resistance between the heat sink and air, A is the surface area of the heat sink, and h is the equivalent convection coefficient

$$h = \frac{1}{R \times A}. \tag{9}$$

After creating the geometry, meshing, setting ambient temperature, and convection coefficient, the temperature profile can be calculated for a given power map.

B. Error Metrics

To study the accuracy of the methods discussed, we calculated the relative error compared with that of ANSYS, which is a standard FEM tool for thermal analysis, using

$$\mathrm{Error} = \frac{T_{\mathrm{Method}} - T_{\mathrm{ANSYS}}}{T_{\mathrm{ANSYS}} - T_{\mathrm{Ambient}}}. \tag{10}$$

Subtracting ambient temperature in denumerator ensures that the error would remain the same in different unit systems. However, this results in a higher error numbers compared with the error reported in the previous works for the same experiment (see [21]).

Maximum Error: For each grid across the entire chip, (10) is used to calculate the error and the maximum error is then reported.

Hot-Spot Error: Equation (10) is used to calculate the error in the hottest spot for the temperature profile.

Average Error: Equation (10) is applied to the average temperature across the chip.

Absolute Temperature Error Range: This error is obtained using

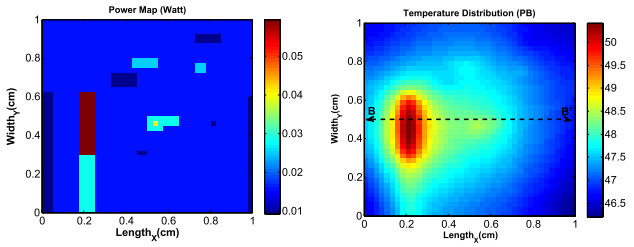
$$\mathrm{Error} = \left(T_{\mathrm{Method}} - T_{\mathrm{ANSYS}}\right). \tag{11}$$

C. HotSpot Calibration

In ANSYS and PB simulations, 0.1 K/W is used for convection resistance. To ensure the overall chip and package models for these methods and HotSpot match, we perform one step of calibration before doing the evaluation. Thus, instead of setting the parameters in HotSpot to match the 0.1 K/W, we try to adjust the convection resistance to a value such that the overall average error is minimized. We then evaluate the relative error values. For example, for the steady-state case study, the initial average error obtained for HotSpot was 19% with convection resistance of 0.1 K/W. However, after applying this adjustment procedure, the optimized value for convection resistance becomes 0.13 K/W, which minimized the average error to 6% (Table IV). Then, for the same optimized value, the transient case study results are obtained. Therefore, in all the following ANSYS and PB simulations, the value for convection resistance is 0.1 K/W, while simulations in HotSpot employ 0.13 K/W for convection resistance.

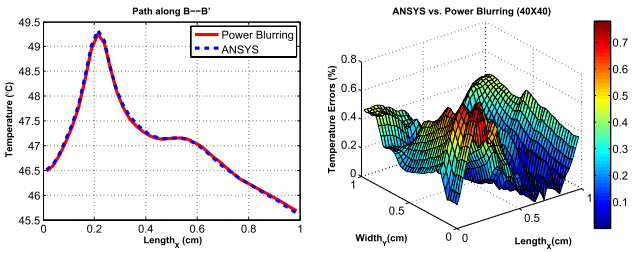
IV. Static Simulation Result

Static simulation results are separated into two parts. In the first part, the simulation results are used to validate the PB method against a FEM software, ANSYS. In the second part, the PB method is compared with a widely used standard architecture-level thermal simulator, HotSpot. The PB method



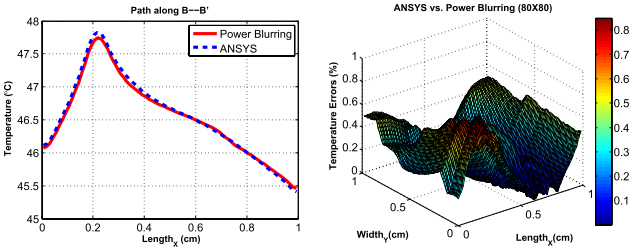
(a) (b)

Fig. 7. (a) Typical power map of an IC. (b) Corresponding thermal profile.



(a) (b)

Fig. 8. Comparisons between the PB method and ANSYS results (40 × 40 grid size). (a) Thermal profile along B–B'. (b) Overall relative error.



(a) (b)

Fig. 9. Comparisons between PB method and ANSYS results (80 × 80 grid size). (a) Thermal profile along A–A'. (b) Overall relative error.

is implemented using MATLAB [37]. The input power map for an IC (1 × 1 cm^2) is shown in Fig. 7(a). The results have been validated against ANSYS, which has been widely used in the industry.

Fig. 7(b) shows the thermal profile obtained by the PB method using the power map shown in Fig. 7(a). The comparisons of the PB method results and the ANSYS simulation results are shown in Figs. 8 and 9 for 40 × 40 and 80 × 80 grid sizes, respectively. Figs. 8(a) and 10(a), are thermal profiles along the B–B' path specified in Fig. 7(b). Figs. 8(b) and 9(b) show the overall relative temperature error. As observed, the PB method renders accurate thermal profiles with maximum temperature error of less than 0.85% with respect to the ANSYS simulation results. ANSYS parameters for these simulations are set according to Table I.

In Table III, a comparison between ANSYS and PB, is presented. In this table, the execution time for each of these methods for different meshing size, as well as maximum error, error in the hottest spot, and average error of the results obtained by PB relative to ANSYS is shown. The advantage of the reduction in computation time becomes more

TABLE III
COMPARISON BETWEEN THE PB AND ANSYS RESULTS FOR TWO DIFFERENT GRID SIZES

	ANSYS execution time	PB execution time	Maximum relative error	Relative hot-spot error	Average relative error
(40×40)	14s	0.035s	0.78%	0.47%	0.27%
(80×80)	33s	0.051s	0.85%	0.47%	0.29%

prominent when we handle power maps with finer grid sizes. For example, as shown in Table III, although PB in a finer grid size (80 × 80) is slightly slower (0.016 s) than a smaller grid size (40 × 40), the ratio of computation time reduction, compared with ANSYS, increases from 345 times faster for the 40 × 40 grid size to 650 times faster for the 80 × 80 grid size, while the maximum relative error is under 0.85%. This is because the number of meshed elements in a volume grid increases significantly; but simultaneously, the fast Fourier transform (FFT) for spatial convolution is very fast [38].

Generating temperature profiles from very fine power maps using conventional numerical analysis methods is extremely difficult as we are often limited by the number of mesh points and the long execution times. However, the PB method only requires a thermal mask, which can have a user-specified resolution using curve fitting based on the circular symmetry near a point heat source [34]. Thus, it can be applied to any level of resolution. In addition, as it has been mentioned, we employ the FFT algorithm in MATLAB to handle the convolution of large-sized matrices. In [34], we pointed out that very fine meshing with a grid of 5–10 μm are needed to obtain an accurate average temperature of the chip. We did ANSYS simulations for power map shown in Fig. 7 with a grid size of 10 × 10 μm^2 (more than 20 million elements in the structure). We obtained corrections less than 0.1 °C) compared with 100 × 100 μm square grid. This means that very small grid size is not necessary to obtain the coarse temperature profile in the chip. However, if there are very small transistors, which dissipate significant power (e.g., in ESD protection circuit), very fine grid is necessary to accurately predict the local temperature field.

A. Comparison With Architecture-Level Simulators

Thermal simulators, such as HotSpot [7], are designed to calculate temperature profiles, which are accurate for the experiments at the architecture level (block sizes in tens of micro to millimeter range), and still fast enough to allow for the simulation of long dynamic temperature traces on the order of seconds. Their main feature, small computation time, compared with detailed finite-element models, comes at a cost of accuracy. Nevertheless, this does allow architects to study thermal and performance tradeoffs in their system design.

HotSpot is based on an equivalent circuit of thermal resistances and capacitances that correspond to the microarchitecture blocks. The essential aspects of the thermal package are also considered [8].

HotSpot can model steady state as well as transient cases. It can be run in two modes: 1) the block model mode

TABLE IV
MATERIAL PROPERTIES AND DIMENSIONS OF THE PACKAGE MODEL FOR COMPARISON WITH HOTSPOT

	Area (mm^2)	Thickness (mm)	Thermal Conductivity (W/m-K)	Density (kg/m^2)	Specific Heat (J/kg-K)
Si Die	17×11.35	0.15	100	2330	751
TIM1	17×11.35	0.020	4	1930	2072.5
Heat Spreader	37.5×37.5	1	400	8933	397.4
TIM2	37.5×37.5	0.020	4	1930	2072.5
Heat Sink	60×60	6.9	400	8933	397.4

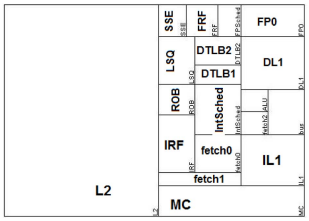
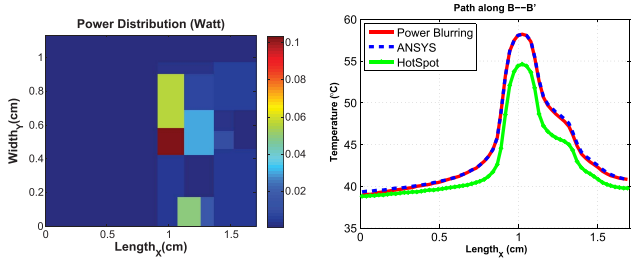


Fig. 10. Floor plan of the architecture blocks.



(a) (b)

Fig. 11. Comparison between three methods. (a) Power map. (b) Temperature profiles along the device width.

and 2) the grid model mode. While block-level simulation has higher speed, the grid-level simulation is more accurate. We compare our method with the HotSpot in grid model mode. Dimensions of the packaged IC and specification of the materials are chosen in accordance with the HotSpot simulator default values. These values and the floor plan of the processor model are shown in Table IV and Fig. 10, respectively. Here, we consider a typical Bobcat mobile processor model [39].

In this paper, we chose a 64 × 64 grid size, which is the default grid size for the HotSpot simulator in grid model mode. As mentioned in Section III-C, the convection resistance for HotSpot is adjusted so that the average error is minimized. The adjusted value of the convection resistance between the heat sink and air is 0.13 K/W. For the power map shown in Fig. 11(a), which is the power map of a typical Bobcat mobile processor, ANSYS, PB, and HotSpot results are obtained and compared. Table V shows the computation time, maximum relative error, and relative error in the hottest spot of the chip for each of these methods.

As shown in Fig. 11, the PB method offers a more accurate result while its execution time is shorter than HotSpot. In the

TABLE V
COMPARISON BETWEEN THE HOTSPOT AND PB

	ANSYS	Hotspot	PB
Computation Time	56s	0.11s	0.041s
Err. in hot-spot	-	12.9%	0.14%
Max. Err.	-	25.7%	13.7%
Avg. Err.	-	6.5%	2.5%
Abs. Err. range	-	0-4.2°C	0-0.56°C

hottest spot of the chip, the relative error for the PB method is only 0.14% compared with ANSYS, while it is 12.9% for the HotSpot simulator. The maximum error of 14% in the entire temperature profile relative to ANSYS is due to a temperature difference of less than 0.6° (39.1 °C−38.5 °C). However, because this small change of temperature occurs at the very edge of the chip, in which the temperature is much lower than the center and very close to ambient temperature (35 °C), it will result in a large % error value even though it is a negligible change (10). The computation time for ANSYS is 56 s, whereas it is 0.11 and 0.04 s for HotSpot and PB, respectively. In ANSYS, we have used the sparse equation solver algorithm in which the time complexity is of the order of $O(n^2)$, for a thermal circuit with n nodes, while the time complexity of the FFT algorithm used in PB method is of the order of $O(\text{nlog}(n))$. HotSpot uses traditional integration-based solvers for which the lower bound of the computation time complexity is of the order of $O(n^c)$, where c is a number between 1.5 and 2 [40]. Considering these orders, one can see that by increasing the number of nodes in the thermal grid model, the computation time of the PB method increases with asymptotically slower rate than the other methods.

V. TRANSIENT THERMAL SIMULATION

A. Simulation Methodology

For steady-state thermal simulations, the thermal mask is obtained with a spatial impulse (i.e., point heat source). The PB method can be applied to the transient thermal simulation with a minor adjustment. The difference is that the time evolution of the thermal mask resulting from spatiotemporal impulse is employed for the transient simulation. To obtain an impulse response in the time domain, a delta function needs to be applied to the center of the die. In practice, a point heat source is applied for a very short time (approximate delta function), and the corresponding thermal response is recorded at each time step, which is shorter than the width of the approximate delta function. The width of the delta function is determined by the desired level of temporal resolution. The resulting thermal responses are normalized with respect to the amount of applied power. The series of thermal masks acquired at the end of this procedure constitutes a transient thermal mask (i.e., time evolution of the thermal mask). Once the transient thermal mask is prepared, the transient temperature profile is obtained by means of the superposition principle. A schematic overview of the transient thermal simulation process is shown in Fig. 12.

We discretize the power pulses to be the same width as the width of the time steps of the transient thermal mask. If we assume the point heat source is divided into r time

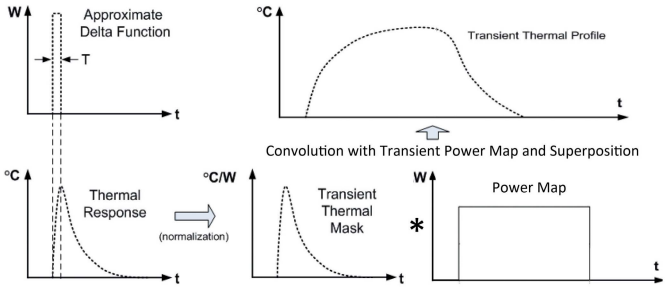


Fig. 12. Schematic overview of transient thermal simulation. The superposition is performed according to the algorithm shown in Fig. 13.

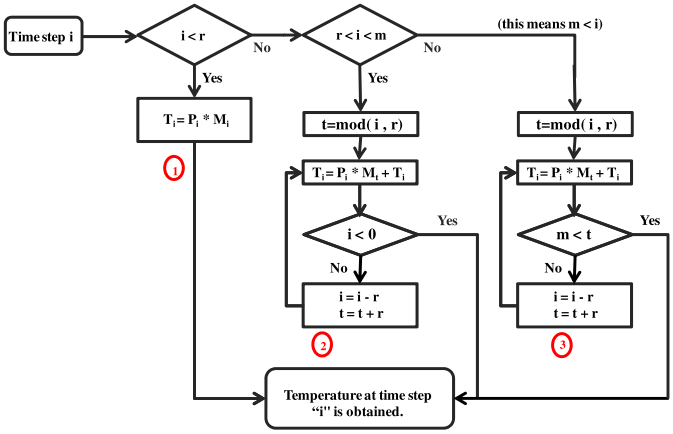


Fig. 13. Algorithm for transient PB. Upon user request, the temperature profile at any time step, i, can be computed. r: number of time steps in impulse heat source. m: number of time steps in total length of thermal mask. T_i: temperature at time step i. M: thermal mask. P_i: power at time step i.

steps, the power map is discretized into n time steps, and the thermal mask has m time steps. The flowchart shown in Fig. 13 is then employed to calculate the temperature profile at each time step. This algorithm efficiently minimizes the number of convolutions as well as aggregation operations required to calculate the temperature profile at each time step. The flowchart is divided into three parts. In the first part, the temperature responses at the times smaller than the width of the point heat source impulse are calculated. As can be observed, at those times, we need only one convolution to calculate the temperature profile. At the times longer than the width of the point heat source impulse, but shorter than the full length of the transient thermal mask, the number of convolutions increases one term at each r time step. At the times longer than the length of the transient thermal mask, the number of convolutions is fixed. The method of image and intrinsic error compensation mentioned previously are applied at each and every time step when the convolution is performed.

Let us assume $r = 3$, $n = 15$, and $m = 9$ time steps. This means we have 15 power cycles and need to calculate temperature at those cycles and afterward. The calculation procedure shown in Fig. 13 results in

Part 1 :

$$T_1 = P_1 * M_1 \quad T_2 = P_2 * M_2 \quad T_3 = P_3 * M_3$$

Part 2 :

$$T_4 = P_4 * M_1 + P_1 * M_4$$
$$T_5 = P_5 * M_2 + P_2 * M_5$$

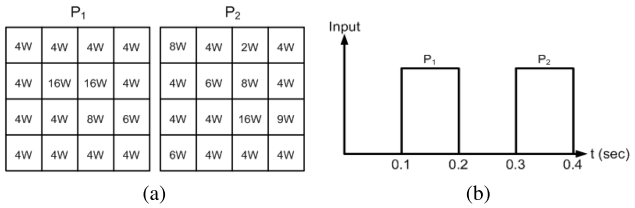


Fig. 14. Pulse input of the coarse power maps. (a) Coarse power maps. (b) Power dissipation pattern.

$$T_6 = P_6 * M_3 + P_3 * M_6$$
$$T_7 = P_7 * M_1 + P_4 * M_4 + P_1 * M_7$$
$$T_8 = P_8 * M_2 + P_5 * M_5 + P_2 * M_8$$
$$T_9 = P_9 * M_3 + P_6 * M_6 + P_3 * M_9$$

Part 3 :

$$T_{10} = P_{10} * M_1 + P_7 * M_4 + P_4 * M_7 + \mathbf{P_1} * (\mathbf{M_{10}} = \mathbf{0})$$
$$= P_{10} * M_1 + P_7 * M_4 + P_4 * M_7$$
$$T_{11} = P_{11} * M_2 + P_8 * M_5 + P_5 * M_8.$$

It can be seen that the temperature profile at each instant of time can be calculated separately. Compared with a previous method for transient PB described in [41], this new algorithm eliminates all the unnecessary convolutions and aggregations. In addition, there is no need to record a thermal basis, proposed in [41], for the calculation of the temperature profile, which improves the performance and speed of this transient PB. In the following section, PB, ANSYS, and HotSpot are compared for two transient case studies.

B. Transient Simulation Results

1) Case Study 1—Contrived Power Map: Transient simulations are performed for two case studies, a simple contrived power map as well as a real processor workload. The first case study is for the power train input shown in Fig. 14(a). Two different power maps (P1 and P2) in Fig. 14(a) are applied at $t = 0.1$ s and $t = 0.3$ s for 0.1-s duration, respectively. For this case study, a 1×1 cm^2 chip with its cooling solution using the dimensions and the material properties presented in Table I, is employed. The chip has been meshed with 64×64 grid size. Resulting temperature profiles on the top of the silicon die at different time instances and the transient thermal trace of the center element are shown in Fig. 15.

The temperature profiles at $t = 0.15$ and 0.35 s are shown in Fig. 16(a) and (b), respectively. The maximum absolute error in the temperature profile over the entire chip as well as the error in the hottest spot of the chip at each time step is shown in Fig. 17. A detailed comparison between PB and HotSpot at times $t = 0.15$ s and $t = 0.35$ s is shown in Table VI. The absolute error in the hottest spot on the chip is less than 0.2 °C throughout the case study. In addition, the maximum error is about 3.5 °C, which occurs at the time when there is no power being dissipated in the chip and the obtained temperature of both methods is very close to the ambient temperature. Therefore, a very small change leads to a large relative error value. In terms of computational efficiency, the PB simulation was completed in 23 s whereas this value was 697 and 27 858 s for HotSpot and ANSYS, respectively.

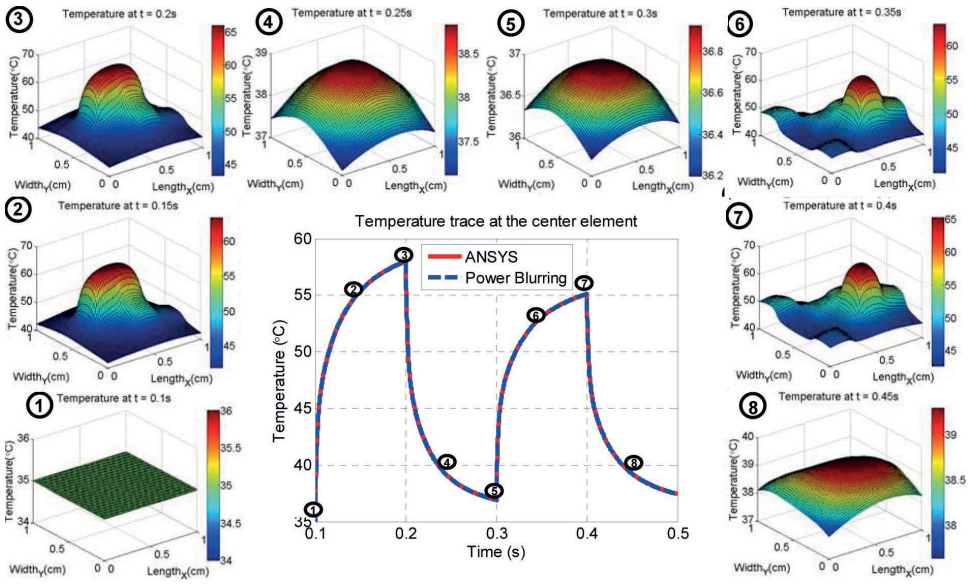
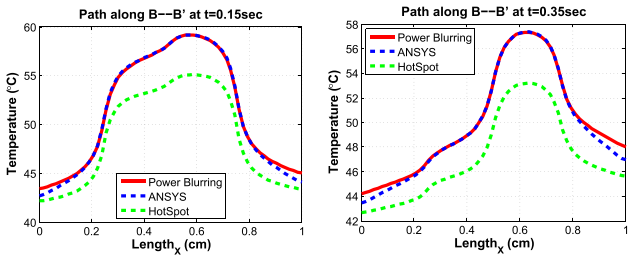
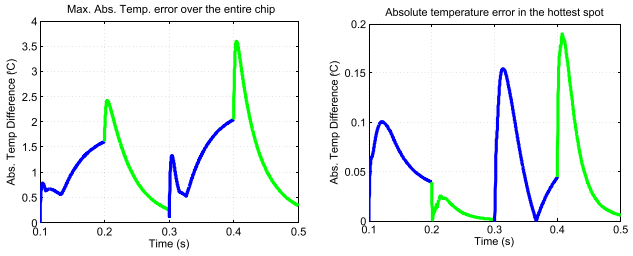


Fig. 15. Transient temperature profile at the center of the IC chip for the time-dependent inputs of the two different power maps in Fig. 14(a).



(a) (b)

Fig. 16. Profile B–B' at (a) $t = 0.15$ s and (b) $t = 0.35$ s.



(a) (b)

Fig. 17. Evaluating absolute error over time for the PB and ANSYS methods in the transient state. Blue lines: when the device is ON and power is being dissipated. (a) Maximum absolute temperature error. (b) Absolute temperature error in the hottest spot on the chip.

As this case study demonstrates, the PB method is about 28 times faster than HotSpot while it provides considerably more accurate results. PB also rendered its results three orders of magnitude faster than ANSYS.

2) Case Study 2—Real Processor Workload: In the second case study, we evaluated the transient response of the mobile processor executing *gcc* workload from SPEC CPU 2000 benchmark suite. We evaluate only one workload because the goal is to show the capability of running transient simulation

TABLE VI

COMPARISON BETWEEN THE PB AND HOTSPOT (TRANSIENT)

	Hotspot	PB
Computation Time	697	23s
Err. in hot-spot @0.15s	16.7%	0.24%
Max. Err. @0.15s	22.2%	16%
Avg. Err. @0.15s	11.7%	2.1%
Abs. Err. range @0.15s (°C)	0-4.7	0-1
Err. in hot-spot @0.35s	16%	0.13%
Max. Err. @0.35s	22.3%	18.6%
Avg. Err. @0.35s	15.3%	3%
Abs. Err. range @0.35s (°C)	0-4.5	0-1.34

with the PB method. As shown in [44], not all the workloads show varied enough thermal transients, so we made certain to pick the selected workload, i.e., gcc, from the thermally interesting category introduced in that work. We used SESC architectural performance simulator [42] to run the workload. To obtain the power trace, we modified SESC to send activity counters of each microarchitectural block to McPAT microarchitectural power model [43] every 3 μs (around every 10K cycles at 3 GHz, as recommended in [7]). The floor plan and dimensions of the chip are shown in Fig. 10 and Table IV, respectively. The chip is meshed with grid size of 64 $\times$ 64. For PB method in this case, the width of the delta function is 6 μs and the time step is 3 μs.

To obtain the transient thermal mask, an impulse heat with the width of 100 μs and a time step of 33 μs is applied on the center element of the chip. The result is recorded for 60-ms duration.

Evolution of the hotspot as well as the average resulting temperature acquired by PB and HotSpot are compared in Fig. 18(a), and (b), respectively. The difference between the values calculated by the two methods is also plotted in Fig. 18(c), and (d). For the PB results in these images,

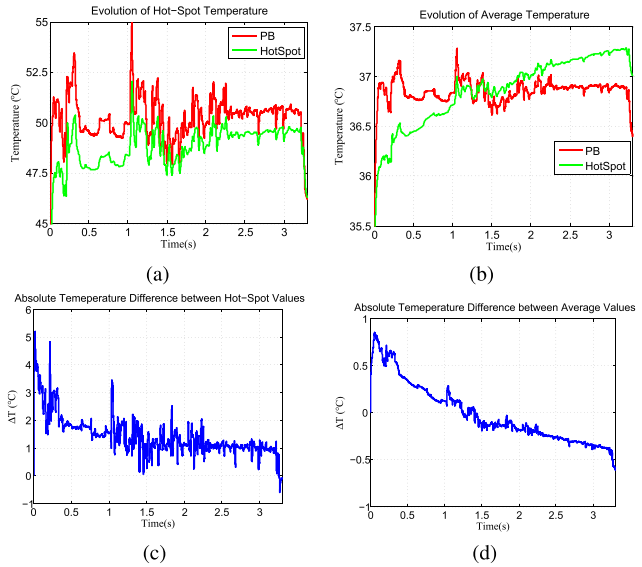


Fig. 18. Comparison between HotSpot and PB for *gcc* workload. Evolution of (a) maximum and (b) average temperatures, in floor-plan blocks over time. Absolute temperature difference between (c) maximum and (d) average values.

a 60-ms-length thermal mask is employed. In both methods, we averaged the power trace every 400 cycles. We cannot report the error for each method because we are unable to run such a long a transient simulation with ANSYS to obtain the reference temperature profile. However, this experiment emphasizes the capability of the PB method for integration with an architectural performance simulator and performing a transient simulation at grid level for real-time workloads. For this simulation, HotSpot took 193 min to obtain the results while PB took 67 min. The absolute temperature difference between the maximum and average values obtained by the two methods is also plotted in Fig. 18(c) and (d). These two figures are plotted to illustrate that the difference in accuracy between these two methods is beyond a simple fitting parameter. Even though the main emphasis of this paper is to improve the computation times of static and transient thermal simulations, the case studies presented in this paper are indicative of the high accuracy of the PB method. This high accuracy can benefit some dynamic thermal management studies in which the absolute error is as important as the trend of temperature progress. For instance, this high accuracy of estimating maximum temperature can be employed in temperature throttling and dynamic thermal management techniques. Considering the impact of thermal throttling on performance, a few degrees difference can lead to considerably different performance results. For example, Ardestani *et al.* [45] evaluate the impact of accurate temperature estimation on performance estimation. They compare two integrated temperature simulation methods (TASS and TAPS), with a slight different absolute accuracy. In their work, considering gcc00, TAPS with slight more error compared with TASS (both compared against oracle simulation called full), results in around 53% error in performance estimation, while TASS with lower error in temperature estimation results in 8% performance error. Note that both methods capture the trend. In addition, a transient thermal method with such an accuracy and speed is proposed to be integrated with an architectural simulator for thermal-aware floor-planning applications [46].

TABLE VII
EFFECT OF THE TRANSIENT THERMAL MASK TRUNCATION ON THE SPEED OF THE PB METHOD

	PB-60ms	PB-30ms	PB-15ms	HotSpot
Execuation Time (min)	67.7	35.4	19.2	193

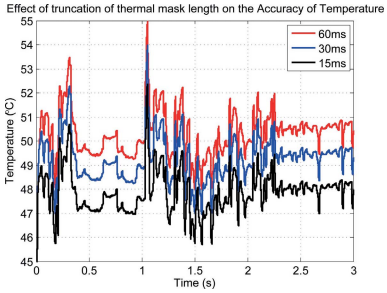


Fig. 19. Effect of the transient thermal mask length's truncation on the accuracy of the PB method.

TABLE VIII
EFFECT OF NEGLECTING METHOD OF IMAGE ON THE SPEED OF THE PB METHOD

	PB-60ms	PB-30ms	PB-15ms	HotSpot
Execuation Time (min)	20.1	10.3	5.5	193

C. Mask Truncation Effect

Using a truncated thermal mask, i.e., decreasing the number of time steps, and losing some accuracy, transient PB could be even faster. This is shown in Table VII as well as Fig. 19. By truncating the thermal mask at 15 ms and sacrificing less than 5% of accuracy, it would take only 19.2 min for PB to obtain the results at this grid level. By neglecting method of image, which is an additional step to increase the accuracy at the corners and edges, PB can be even faster without losing any accuracy in calculating the peak temperature. For the latter case, a comparison between the speed of computation PB and HotSpot is conducted and the results are presented in Table VIII.

It should be mentioned that the PB method relies on two FEA simulations (or measurements) giving the unit impulse response at the center of the chip and the additional correction factor from the uniform power dissipation profile. These calculations could be done offline, so the main advantage of PB is in multiple thermal simulations when different placements of the IC blocks are studied. All the matrix arithmetic calculations in the PB method have been done in MATLAB. While this is flexible and it allows the use of image processing tools, it is anticipated that direct implementation of the matrix convolution in a higher level program (e.g., C) can significantly increase the speed of the PB method.

VI. CONCLUSION

The demands for efficient thermal simulation for VLSI ICs in a thermal package as well as power electronic and

optoelectronic devices have raised as the CMOS technology scales down and power densities goes up. In this paper, we have presented a highly accurate yet fast thermal simulation method, namely PB. We have demonstrated that the PB method is suitable for both static and transient thermal simulations). Unlike the conventional Green's function based methods, realistic package models can be incorporated with the PB method. The current implementation of the algorithm uses MATLAB [37], [47]. Given efficient image processing tools and graphical IC chips, the calculation times could still be improved significantly. In addition, like other current methods, the accuracy of the temperature profile is dependent upon the accuracy of the estimated power dissipation map, which can be obtained through well-established power estimation methods [30]. In addition to its speed and accuracy, versatility of PB makes it distinctive compared with other available methods. Fast and accurate temperature calculation in 3-DICs is illustrated in [23]. Employing material-dependent thermal masks in [24] enabled PB to incorporate thermal vias as well as material nonuniformities of different layers in 3-DICs. Melamed *et al.* [48] conducted a junction-level thermal analysis of 3-DICs using PB. Three orders of magnitude improvement in runtime as well as six orders of magnitude speedup are obtained by their PB method in comparison with solving the thermal network of the chip directly. Acquiring temperature profiles of power electronic transistor arrays [25], solving nonlinear problems (calculating temperature profiles in ICs considering temperature dependence of material properties of the chip) using only two or three iterations [26], and solving the inverse problem (i.e., obtaining heat dissipated in ICs from their temperature profiles) [27] are some other examples, which are illustrative of versatility of the PB technique.

REFERENCES

[1] G. E. Moore, "Progress in digital integrated electronics," in *Proc. IEDM*, vol. 21. 1975, pp. 11–13.

[2] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[3] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware computer systems: Opportunities and challenges," *IEEE Micro*, vol. 23, no. 6, pp. 52–61, Nov./Dec. 2003.

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

[5] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in vlsi circuits: Principles and methods," *Proc. IEEE*, vol. 94, no. 8, pp. 1487–1501, Aug. 2006.

[6] Y. K. Cheng and S. M. Kang, "A temperature-aware simulation environment for reliable ULSI chip design," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 19, no. 10, pp. 1211–1220, Oct. 2000.

[7] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware micro-architecture: Extended discussion and results," Dept. Comput. Sci., Univ. Virginia, Charlottesville, VA, USA, Tech. Rep. CS-2003, 2003.

[8] (2012). *HotSpot Thermal Simulator* [Online]. Available: http://lava.cs.virginia.edu/HotSpot/index.htm

[9] A. J. Chapman, *Heat Transfer*, 4th ed. New York, NY, USA: Macmillan, 1984.

[10] B. Wang and P. Mazumder, "Accelerated chip-level thermal analysis using multilayer Green's function," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 26, no. 2, pp. 325–344, Feb. 2007.

[11] Y. Zhan and S. Sapatnekar, "A high efficiency full-chip thermal simulation algorithm," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, San Jose, CA, USA, Nov. 2005, pp. 635–638.

[12] T. Kemper, Y. Zhang, Z. Bian, and A. Shakouri, "Ultrafast temperature profile calculation in IC chips," in *Proc. 12th Int. Workshop Thermal Investigat. ICs*, Nice, France, 2006, pp. 133–137.

[13] P. E. Bagnoli, C. Bartoli, and F. Stefani, "Validation of the DJOSER analytical thermal simulator for electronic power devices and assembling structures," in *Proc. Workshop Thermal Investigat. ICs Syst.*, 2007, pp. 185–196.

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

[15] L. Codecasa, D. D'Amore, and P. Maffezzoni, "An Arnoldi based thermal network reduction method for electro-thermal analysis," *IEEE Trans. Compon. Packag. Technol.*, vol. 26, no. 1, pp. 186–192, Mar. 2003.

[16] S. C. Lin and K. Banerjee, "An electrothermally-aware full-chip substrate temperature gradient evaluation methodology for leakage dominant technologies with implications for power estimation and hot-spot management," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, San Jose, CA, USA, Nov. 2006, pp. 568–574.

[17] B. Allard, X. Jorda, P. Bidan, A. Rumeau, H. Morel, X. Perpina, *et al.*, "Reduced-order thermal behavioral model based on diffusive representation," *IEEE Trans. Power Electron.*, vol. 24, no. 12, pp. 2833–2846, Dec. 2009.

[18] W. Huang, M. R. Stan, K. Skadron, K. Sankaranarayanan, S. Ghosh, and S. Velusamy, "Compact thermal modeling for temperature-aware design," in *Proc. DAC*, San Diego, CA, USA, Jun. 2004, pp. 878–883.

[19] A. Ziabari, E. K. Ardestani, J. Renau, and A. Shakouri, "Fast thermal simulators for architectural level circuit design," in *Proc. 27th Annu. Thermal Meas., Model. Manag. Symp.*, San Jose, CA, USA, Mar. 2011, pp. 70–75.

[20] Y. Han, I. Koren, and C. M. Krishna, "TILTS: A fast architectural-level transient thermal simulation method," *J. Low Power Electron.*, vol. 3, no. 1, pp. 1–9, 2007.

[21] Y. Yang, Z. P. Gu, C. Zhu, R. P. Dick, and L. Shang, "ISAC: Integrated space-and-time-adaptive chip-package thermal analysis," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 26, no. 1, pp. 86–99, Jan. 2007.

[22] (2012, Apr.). *COMSOL Multiphysics* [Online]. Available: http//www.comsol.com/products/multiphysics

[23] J. H. Park, A. Shakouri, and S. M. Kang, "Fast thermal analysis of vertically integrated circuits (3D ICs) using power blurring method," in *Proc. InterPACK*, San Francisco, CA, USA, Jul. 2009, pp. 19–23.

[24] A. Ziabari and A. Shakouri, "Fast thermal simulations of vertically integrated circuits (3D ICs) including thermal vias," in *Proc. ITherm*, May 2012, pp. 588–596.

[25] K. Maize, X. Wang, D. Kendig, A. Shakouri, W. French, B. O'Connell, *et al.*, "Thermal characterization of high power transistor arrays," in *Proc. 25th Annu. IEEE Semicond. Thermal Meas. Manag. Symp.*, San Jose, CA, USA, Mar. 2009, pp. 50–54.

[26] A. Ziabari, Z. Bian, and A. Shakouri, "Adaptive power blurring techniques to calculate IC temperature profile under large temperature variations," in *Proc. IMAPS*, Sep. 2010, pp. 1–6.

[27] X. Wang, S. Farsiu, P. Milanfar, and A. Shakouri, "Power trace: An efficient method for extracting the power dissipation profile in an IC chip from its temperature map," *IEEE Trans. Compon. Packag. Technol.*, vol. 32, no. 3, pp. 309–317, Jun. 2009.

[28] C. A. Balanis, *Advanced Engineering Electromagnetics*, New York, NY, USA: Wiley, 1989, ch. 14.

[29] R. C. Conzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[30] F. N. Najm, "A survey of power estimation techniques in VLSI circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 2, no. 4, pp. 446–455, Dec. 1994.

[31] 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 I: Electrothermal couplings and full-chip package thermal model," *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3342–3350, Dec. 2007.

[32] C. H. Diaz, S. M. Kang, and C. Duvvury, "Circuit-level electrothermal simulation of electrical overstress failures in advanced MOS I/O protection devices," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 13, no. 4, pp. 482–493, Apr. 1994.

[33] *ANSYS R11.0*, Swanson ANSYS Inc., Philadelphia, PA, USA, 2007.

[34] V. M. Heriz, J. H. Park, A. Shakouri, and S. M. Kang, "Method of images for the fast calculation of temperature distributions in packaged VLSI chips," in *Proc. 13th Int. Workshop Thermal Investigat. ICs*, Budapest, Hungary, Sep. 2007, pp. 18–25.
[35] I. V. Lindell, "Image theory for electromagnetic sources in chiral medium above the soft and hard boundary," *IEEE Trans. Antennas Propag.*, vol. 49, no. 7, pp. 1065–1068, Jul. 2001.
[36] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.
[37] *MATLAB R2006b*, The MathWorks Inc., Natick, MA, USA, 2006.
[38] J. H. Park, X. Wang, A. Shakouri, and S. M. Kang, "Fast computation of temperature profiles of VLSI ICs with high spatial resolution," in *Proc. 24th Semi-Therm*, San Jose, CA, USA, Mar. 2008, pp. 50–54.
[39] B. Burgess, B. Cohen, M. Denman, J. Dundas, D. Kaplan, and J. Rupley, "Bobcat: AMD's new low-power x86 processor," *IEEE Micro*, vol. 31, no. 2, pp. 16–25, Jan. 2011.
[40] P. Liu, Z. Qi, H. Li, L. Jin, W. Wu, S. X.-D. Tan, *et al.*, "Fast thermal simulation for architecture level dynamic thermal management," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, Nov. 2005, pp. 639–644.
[41] J. H. Park, A. Shakouri, and S. M. Kang, "Fast evaluation method for transient hot spots in VLSI ICs in packages," in *Proc. 9th ISQED*, San Jose, CA, USA, 2008, pp. 600–603.
[42] J. Renau, B. Fraguela, J. Tuck, W. Liu, M. Prvulovic, L. Ceze, *et al.*. (2005). *SESC Simulator* [Online]. Available: http://sesc.sourceforge.net
[43] S. Li, J. Ho Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchit.*, Dec. 2009, pp. 469–480.
[44] F. J. Mesa-Martinez, E. K. Ardestani, and J. Renau, "Characterizing processor thermal behavior," in *Proc. 15th Ed. ASPLOS Archit. Support Program. Lang. Operat. Syst.*, 2010, pp. 193–204.
[45] E. K. Ardestani, E. Ebrahimi, G. Southern, and J. Renau, "Thermal-aware sampling in architectural simulation," in *Proc. ISLPED*, Aug. 2012, pp. 33–38.
[46] E. K. Ardestani, A. Ziabari, A. Shakouri, and J. Renau, "Enabling power density and thermal-aware floorplanning," in *Proc. 28th Annu. Thermal Meas., Model. Manag. Symp.*, San Jose, CA, USA, Mar. 2012, pp. 302–307.
[47] U.S. Patent 7 627 841.
[48] S. Melamed, T. Thorolfsson, T. R. Harris, S. Priyadarshi, P. Franzon, M. B. Steer, *et al.*, "Junction-level thermal analysis of 3-D integrated circuits using high definition power blurring," *IEEE Trans Comput.-Aided Des. Integr. Circuits Syst.*, vol. 31, no. 5, pp. 676–689, May 2012.

Amirkoushyar Ziabari (M'13) received the bachelor's and first master's degrees in microelectronic circuits from Amirkabir and Sharif Universities, Tehran, Iran, and the second master's degree in semiconductor devices and nanoelectronics from the University of California, Santa Cruz, CA, USA. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering and the Birck Nanotechnology Center, Purdue University, West Lafayette, IN, USA.

His current research interests include the linear and nonlinear thermoelectric effects in semiconductors, electro-thermal simulation and modeling of devices using FEM, high resolution thermal imaging, and thermal stress analysis in high temperature thermoelectrics.

Je-Hyoung Park received the B.S. degree from the Kyungpook National University at Daegu, Korea, in 2001, the M.S. degrees from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2004, and the Ph.D. degree from the University of California at Santa Cruz, Santa Cruz, CA, USA, in 2009, all in electrical engineering.

He has published over ten papers. He is currently with Memory Business of Samsung Electronics, Hwaseong, Korea. His current research interests include thermal modeling and simulation of semiconductor packages, memory modules, and servers.

Dr. Park is a Committee Member of the Joint Electron Device Engineering Council.

Ehsan K. Ardestani received the B.S. and M.S. degrees in computer engineering and computer architecture from the Isfahan University and Amirkabir University of Technology, Tehran, Iran. He is currently pursuing the Ph.D. degree in computer engineering with the University of California, Santa Cruz, CA, USA.

His current research interests include computer architecture, power modeling, power and thermal-aware design, and simulation methodology for scalable multicore systems.

Jose Renau (A'11) received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign, Urbana, IL, USA.

He is an Associate Professor of computer engineering with the University of California, Santa Cruz, CA, USA. His current research interests include computer architecture, including design effort metrics and models, infrared thermal measurements, low-power and thermal-aware designs, process variability, thread level speculation, and FPGA/ASIC design.

Sung-Mo Kang (M'03) is the 15th President of Korea Advanced Institute of Science and Technology (KAIST) in Daejeon, Korea, a Distinguished Chair Professor on leave of the Jack Baskin School of Engineering, UC Santa Cruz, and Chancellor Emeritus of UC Merced. He received the B.S. degree from Fairleigh Dickinson University, Teaneck, New Jersey in 1970, the M.S. degree from the State University of New York at Buffalo in 1972, and the Ph.D. degree from the University of California at Berkeley in 1975, all in electrical engineering. From 1995 to 2000, he was the Department Head of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign (UIUC). Prior to UIUC, he was the Supervisor of High-End Microprocessor Design, responsible for development of world's premier full-CMOS 32b BELLMAC-32, at AT&T Bell Laboratories, Murray Hill, NJ, and also served as a faculty member of Rutgers University, New Brunswick, NJ. From 2001 to 2007, he was the Dean of the Jack Baskin School of Engineering and Professor of Electrical Engineering at UC Santa Cruz. From 2007 to 2011 he served as the Chancellor and Professor of Engineering at UC Merced, the 10th campus of the UC System. Dr. Kang is a Fellow of the IEEE, the Association for Computing Machinery (ACM), and the American Association for the Advancement of Science (AAAS). He has served on the Presidential Advisory Council on Science and Technology of Korea as Chair of Creative Economy. His research interest includes modeling and simulation of semiconductor devices and circuits; memristors, memristive devices and systems; low-power very large scale integration design and optimization for power, performance, reliability and manufacturability; and nanobioelectronic circuits and systems.

Ali Shakouri (M'12) received the Engineering degree from Telecom Paris, Paris, France, in 1990, and the Ph.D. degree from the California Institute of Technology, Pasadena, CA, USA, in 1995.

He is the Mary Jo and Robert L. Kirk Director with the Birck Nanotechnology Center and a Professor of electrical and computer engineering with Purdue University, West Lafayette, IN, USA. He is involved in research on a new interdisciplinary sustainability curriculum in collaboration with colleagues in engineering and social sciences. His current research interests include nanoscale heat and current transport in semiconductor devices, high resolution thermal imaging, and waste heat recovery systems.

Dr. Shakouri received the Packard Fellowship in Science and Engineering in 1999 and the National Science Foundation Career Award in 2000.