Multiscale Thermal Analysis for Nanometer-Scale Integrated Circuits

Zyad Hassan, *Student Member, IEEE*, Nicholas Allec, *Student Member, IEEE*, Li Shang, *Member, IEEE*, Robert P. Dick, *Member, IEEE*, Vishak Venkatraman, *Member, IEEE*, and Ronggui Yang, *Member, IEEE*

Abstract—Thermal analysis has long been essential for designing reliable high-performance cost-effective integrated circuits (ICs). Increasing power densities are making this problem more important. Characterizing the thermal profile of an IC quickly enough to allow feedback on the thermal effects of tentative design changes is a daunting problem, and its complexity is increasing. The move to nanometer-scale fabrication processes is increasing the importance of thermal phenomena such as ballistic phonon transport. The accurate thermal analysis of nanometer-scale ICs containing hundreds of millions of devices requires characterization of heat transport across multiple length scales. These scales range from the nanometer scale (device-level impact) to the centimeter scale (cooling package impact). Existing chip–package thermal analysis methods based on classical Fourier heat transfer cannot capture nanometer-scale thermal effects. However, accurate device-level modeling techniques, such as molecular dynamics methods, are far too slow for use in full-chip IC thermal analysis. In this paper, we propose and develop ThermalScope, a multiscale thermal analysis method for nanometer-scale IC design. It unifies microscopic and macroscopic thermal modeling methods, i.e., the Boltzmann transport equation and Fourier modeling methods. Moreover, it supports adaptive multiresolution modeling. Together, these ideas enable the efficient and accurate characterization of nanometer-scale heat transport as well as the chip–package-level heat flow. ThermalScope is designed for full-chip thermal analysis of billion-transistor nanometer-scale IC designs, with accuracy at the scale of individual devices. ThermalScope enables the accurate characterization of various temperature-related effects, such as temperature-dependent leakage power and temperature–timing dependences. ThermalScope has been implemented in software and used for the full-chip thermal analysis and temperature-dependent leakage analysis of an IC design with more than 150 million transistors. ThermalScope will be publicly released for free academic and personal use.

Index Terms—Integrated-circuit (IC) thermal factors, leakage-power estimation, nanoscale heat flow, simulation.

I. Introduction

PROCESS scaling and increasing device density increase integrated circuit (IC) power density and thermal effects. Increased IC power consumption and temperature affect circuit performance (via reduced transistor carrier mobility [1], decreased threshold voltage, and increased interconnect resistance), reliability (via electromigration [2], dielectric breakdown, and negative body biasing), power consumption (via increased subthreshold current [3]), and cooling cost. IC thermal analysis is thus critical because it provides valuable guidance for IC design- and run-time thermal optimization.

CMOS technology is fast approaching the nanometer-scale regime. The 45-nm CMOS fabrication technology is entering mainstream use. In the coming five years and beyond, ultrathin-body device structures, such as multigate MOSFET (FinFET) and silicon-on-insulator, will be used for mainstream ICs. As the device feature size approaches the nanometer scale, thermal effects will become prominent. When the mean free path of phonons (lattice vibrations) is comparable to the device feature scale, ballistic phonon transport serves as the main mechanism of heat transfer. Heat transport within nanometer-scale devices is strongly affected by interface scattering and reflection effects.

The thermal profile of a nanometer-scale IC depends on power consumption variation at multiple scales: The chip average temperature is determined by the IC average power density and cooling efficiency. Hot spots in the active layer are often caused by high-power density functional units, e.g., a floating point unit. Inside a transistor, a hot spot often occurs near the drain terminal region, mainly due to the accumulation of slow-moving (optical) phonons. IC thermal analysis thus requires the accurate modeling of heat transport across multiple scales, from nanometer-scale on-chip devices through millimeter-scale silicon chips and centimeter-scale cooling packages to the ambient environment.

IC thermal analysis is the process of characterizing the 3-D thermal profile of an IC chip and cooling package. As previously stated, the IC thermal profile is a complex function of its design, fabrication technology, cooling-package configuration, power consumption, and ambient environment. Conventional chip–package thermal analysis techniques have been so slow that evaluating numerous design alternatives was prohibitively expensive [4]–[6]. As a result, most thermal optimization was done after the packaging and cooling solution design, and by that time, the design is already tightly constrained. Recently, a number of researchers have developed fast thermal analysis techniques for use during the IC design process [7]–[15]. Using
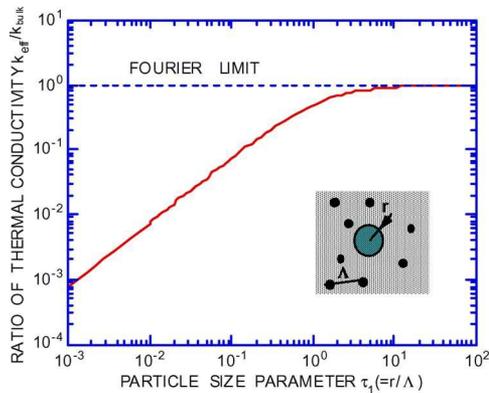
Fig. 1. Effective thermal conductivity surrounding nanoscale heated spheres normalized to the bulk thermal conductivity of the media, plotted as a function of the sphere size r normalized to the phonon mean free path Λ [18], [19].

these methods, heat transfer through the chip and cooling package is modeled using the classical Fourier heat diffusion model. IC chip and cooling packages are virtually partitioned (spatially discretized) into discrete 3-D thermal elements. Compact heat-transfer equations are then derived and solved using numerical methods to characterize the thermal profile of the IC chip and cooling package.

Although some of these techniques are fast enough for use in IC design- and run-time thermal management, they are all based on the Fourier heat flow model. This model cannot capture nanometer-scale thermal effects and produces inaccurate results when used at length scales on the order of the phonon mean free path (i.e., the average distance between interactions) [16], [17]. Ballistic phonon transport implies reduced effective thermal conductivity in proportion to the ratio of the hot-spot size to the phonon mean free path (see Fig. 1). It is expected that heat conduction in some nanometer-scale circuits will deviate considerably from that predicted by the Fourier model due to ballistic phonon transport and the finite relaxation time of heat carriers, and this is supported by the data presented in Section VI-A.

Techniques with different fidelities and efficiencies have been developed to model nanometer-scale device-level heat transport, including molecular dynamics methods [20], the Boltzmann transport equation (BTE) [21], and the ballistic-diffusion model [22]. Computational complexity has been the primary challenge to considering nanometer-scale heat transfer for large-scale IC chip–package thermal analysis.

In summary, the thermal analysis for nanometer-scale ICs raises the following challenges.

1) The numerical thermal analysis of nanoscale device ICs has high computational complexity and memory usage. Accurate thermal analysis requires the use of detailed numerical analysis methods with fine-grain models. The modeling granularities required for nanometer-scale ICs vary by several orders of magnitude. IC chip–package-level thermal analysis with accurate characterization of individual on-chip devices will introduce tremendous computation and memory overhead.

2) Accurate thermal analysis requires unified heat transport modeling from nanoscale devices to the chip–package level. However, chip–package- and device-level thermal analyses are currently two isolated research fields. The

Fourier heat diffusion model has been widely used for fast chip–package-level thermal analysis. However, it does not accurately capture nanometer-scale thermal effects. Device-level modeling techniques, such as molecular dynamics and BTE, model nanoscale thermal effects. However, their usage has been limited to individual devices due to their high computational complexity.

To close the gap between the efficiency and accuracy of nanoscale and chip–package thermal analysis techniques, we propose and develop a multiscale solution, named Thermal-Scope, for a unified device–chip–package thermal analysis targeting billion-transistor nanometer-scale ICs. The proposed multiscale solution integrates *microscopic and macroscopic thermal physics modeling methods*, enabling the characterization of nanometer-scale heat transport as well as chip–package-level heat flow, *detailed and compact numerical analysis techniques*, allowing the use of computationally intensive device-level modeling within full-chip thermal characterization, and *multiresolution adaptive modeling granularities*, permitting modeling thermal effects on length scales ranging from nanometer-scale devices to centimeter-scale packaging and cooling structures. The proposed solution overcomes the limitations of existing chip–package- and device-level thermal analysis methods. It provides a unified modeling infrastructure for IC heat flow analysis from nanometer-scale devices to billion-device IC chips. ThermalScope has been implemented in software and used for the full-chip thermal analysis and temperature-dependent leakage analysis of an IC design with more than 150 million transistors.

The rest of this paper is organized as follows. Section II describes existing methods used to model heat transport and indicates the scales at which they are valid. Section III describes the proposed multiscale thermal analysis infrastructure, ThermalScope. Sections IV and V describe the major components of ThermalScope, namely, the proposed hybrid analysis method employed for efficient device-level thermal analysis and the proposed multiscale techniques for interdevice and chip–package-level thermal analysis. Section VI evaluates and demonstrates the use of ThermalScope. Finally, we conclude in Section VII.

II. BACKGROUND

The problem of subcontinuum heat conduction in transistors has received much attention, particularly in the last decade or so [23], [24]. This section gives a brief overview of the current understanding of heat transport and the different methods used to model it in semiconductors.

Heat transport is governed by phonons, i.e., lattice vibrations. These phonons exhibit the wave–particle duality. There are different types of thermal effects that can exist as the dimensions of structures are decreased. One of these is the classical size effect, for which the particle description of phonons is sufficient. The other is the wave effect, where the phase of the wave nature of phonons must be taken into account [17].

To determine when these effects need to be considered, several length scales can be used. These include the mean free path, phonon wavelength, and the phase coherence length. The

phonon mean free path is the average distance the phonon will cover between interactions. For the wave aspects of phonons, the phase and wavelength are of particular importance. The phase coherence length can be treated as having the same order of magnitude as the mean free path [17]. The wavelength of phonons in silicon (at room temperature) is approximately 1×10^{-10} m [17]. When the length scales of a structure are this small, nanoscale thermal effects must be modeled. Due to the small size of the wavelength, treating the phonons as particles is sufficient for the thermal analysis of current CMOS and FinFET devices. Only at very low temperatures will the wavelength become long and need to be taken into account.

At length scales smaller than the mean free path, phonons travel ballistically, i.e., they travel without scattering. The ballistic transport creates a nonequilibrium situation in which there are no scattering events between the hot phonons from the heat source region and the cold phonons from the region surrounding the heat source, leading to ineffective heat transfer from the heat source in a device. This, in turn, leads to a high temperature in the heat source region because the temperature in this region is representative of the energy of the hot phonons (which have not lost energy to scattering events with cold phonons). In the Fourier model, it is assumed that localized regions reach equilibrium, and the heat can be effectively transferred between these localized regions through scattering within the medium. However, this does not hold when the number of scattering events is negligible, which occurs when the phonon mean free path is larger than the device feature size. The assumption that a local equilibrium is reached implies that there are sufficient scattering events to reduce the energy of the hot phonons, and thus, the Fourier model overpredicts the thermal conductivity [18]. It should be noted, however, that phonon scattering at the device boundaries can reduce the phonon mean free path, making the mean free path dependent on the device geometry [25].

There have been modeling methods with different fidelities to capture heat transfer, including molecular dynamics, the BTE, the ballistic-diffusion model, and the Fourier model. These are briefly discussed here.

Molecular dynamics methods model heat transfer by directly simulating interatomic interactions (by solving atomistic equations of motion) [20], [26], [27]. Approaches using this method are accurate. However, they are known to be extremely computationally expensive and are only suitable for systems having a few atomic layers [28] or several thousands of atoms [29]; thus, these methods are not suitable for even device-level thermal analysis.

The BTE is another method used to model heat transport. In this method, the particle description of phonons is used to emulate the transport of phonons in nonmetallic solids [20], [21], making its use valid when the length scale of the structure is larger than the phonon wavelength. Using this model, the thermal effects due to the ballistic transport of phonons are taken into account. The phonon BTE is given by [27]

$$\frac{\partial f}{\partial t} + \nabla \cdot (\mathbf{v}_g f) + \alpha \cdot \nabla_{\mathbf{v}_g} f = S_{\text{scattering}} \qquad (1)$$

where f is the phonon distribution function, $\mathbf{v}_g$ is the phonon group velocity vector, $S_{\text{scattering}}$ accounts for scatter-

ing processes, and α is the acceleration $d\mathbf{v}_g/dt$ and accounts for external forces. The left-hand side of the equation describes the heat transfer due to the group velocity vector of the phonons. The right-hand side describes the rate of change in the phonon distribution due to scattering and phonon generation. Solutions to the BTE assume a classical definition of temperature in which a local temperature exists [20]. This temperature describes the localized energy of phonons [27].

Solving the BTE for a realistic system is nontrivial, and reasonable assumptions must be applied to make the problem tractable. Several models have been proposed, each with a different level of complexity, such as the gray BTE [21], [24], [27], [30], [31] and the semigray BTE [21], [27], [30]. These models differ in how they treat phonons. The gray model is the simpler of the two, as it treats all phonons as the same (no consideration for polarization or frequency), having a single group velocity. In the semigray model, the phonons are separated into propagating and reservoir modes. Phonons in the reservoir mode are stationary, whereas phonons in the propagating mode are responsible for energy transport. The different modes model the different types of phonons in the system at the cost of additional complexity. In the past, the propagating mode has been used to model longitudinal acoustic phonons (as they are fast moving, having velocities of 7000–8000 m/s in silicon [32]), and the reservoir mode has been used to model transverse acoustic and optical phonons (as they are slow moving, having group velocities of 1000–2000 m/s for longitudinal optical phonons in silicon [32]) [21].

The BTE has been solved using Monte Carlo-based methods [26], [33]–[35] and finite-volume methods [28], [30], [31]. In Monte Carlo-based methods, the system is discretized into elements, each having an initial temperature. These initial temperatures determine the number of phonons in each element and their properties, such as initial velocity. During the simulation, the phonons scatter randomly, with a probability that is dependent on the estimated lifetime of the phonon [33]. In finite-volume methods, the system is discretized into spatial and angular partitions to model phonon transport, and discretized equations are used to solve the BTE. This method is described in detail in Section IV-A. Direct numerical solutions of the BTE are preferred [17] and are much more efficient than molecular dynamics. In addition, solutions to the BTE have been shown to agree well with experimental data [20]. However, they still have high computational complexity, and their usage has been restricted to device-level analysis [36]. In the acoustically thick regime (i.e., when the phonon mean free path is much smaller than the system of interest), solutions to the BTE agree with those of the Fourier model [27].

The ballistic-diffusion-based thermal analysis method is an approximation of the BTE method [21], [22], [27], [36]. In this method, the energy carriers are divided into ballistic and diffusive components. Although the ballistic-diffusion model is more efficient than a solely BTE-based method, it is still much more computationally demanding than the Fourier model. In addition, results from the ballistic-diffusion model tend to have low fidelity [22]. There has also been other work where hybrid solutions of diffusive- and BTE-based models are used, e.g., [24] and [37]. In [37], BTE and diffusion components are

used to model heat transport, which leads to thermal resistances being the sum of BTE and diffusive resistances. In this model, the BTE is solved in a region that is comparable to the phonon mean free path to determine the component of heat transport due to emitted phonons not yet thermalized. In [24], the BTE is used to solve heat transport in the silicon layer while the diffusion equation is used in the oxide layer since the phonon mean free path in this layer is much smaller than its thickness. These studies take advantage of the simplicity of the diffusion-based description of heat transport when appropriate. However, they focus on single device analysis and have limited flexibility to model different device geometries and/or materials. In contrast to these methods, we present a complete system description that provides thermal analysis from the chip–package level down to the device level, while combining diffusive- and BTE-based models.

In the Fourier (diffusion) model, heat transport is governed by the Fourier law [17]

$$q = -k\Delta T \tag{2}$$

where q is the heat flux (rate of heat transfer per unit cross-sectional area), k is the thermal conductivity, and ΔT is the temperature gradient. This law states that the heat flux is proportional to the (negative) gradient of the temperature. Solutions to the Fourier model can be found efficiently and have been shown to agree with experimental results [38]. However, in this model, the thermal conductivity is not dependent on material dimensions, which leads to erroneous predictions of the temperature when the structure dimensions are smaller than the phonon mean free path.

III. THERMALSCOPE: A MULTISCALE THERMAL ANALYSIS INFRASTRUCTURE

In this section, we present an overview of ThermalScope, the proposed multiscale thermal analysis solution for nanometer-scale ICs.

Fig. 2 shows the flow of ThermalScope. ThermalScope is a multiscale solution that integrates microscopic and macroscopic thermal physics modeling methods, as well as multiresolution macromodeling techniques. In contrast with existing Fourier-based chip–package thermal analysis methods, ThermalScope uses accurate BTE analysis to capture nanometer-scale thermal effects that are common at nanometer length scales. Due to the computational complexity of BTE analysis, ThermalScope uses hybrid Fourier/BTE analysis for modeling device thermal analysis. This accelerates thermal analysis by orders of magnitude compared with BTE, while maintaining accuracy. ThermalScope is also equipped with multiresolution modeling methods, including hierarchical multiscale partitioning, clustering, and thermal-gradient-based spatial resolution adaptation, enabling accurate and efficient characterization of thermal effects from chip–package, functional-unit, to device level.

ThermalScope takes as input the power profile, device structure and technology, as well as the dimensions and material information of the chip–package. Using this information, ThermalScope conducts full-chip thermal analysis and reports
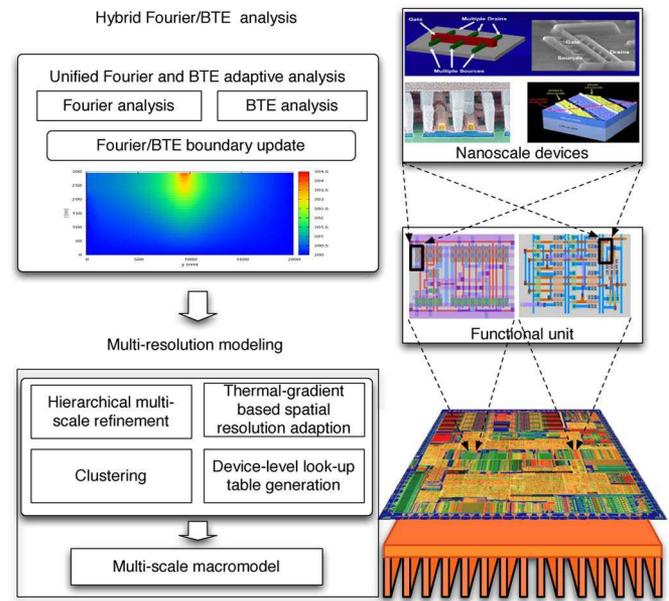
Fig. 2. ThermalScope: The proposed multiscale thermal analysis infrastructure.

the temperatures of all on-chip devices. Using ThermalScope, it is also possible to obtain the leakage power using an iterative approach, as described in Section VI-D. In summary, Thermal-Scope provides a unified modeling infrastructure for IC heat flow analysis from nanometer-scale devices to billion-device IC chips. The functionality and importance of each component used within ThermalScope are described in Sections IV and V.

IV. UNIFIED FOURIER/BTE ADAPTIVE ANALYSIS

This section presents the microscopic and macroscopic thermal physics modeling methods used in ThermalScope. Section IV-A describes the thermal physics models. Section IV-B details the proposed unified Fourier/BTE analysis method.

A. Thermal Physics Models

ThermalScope uses both the Fourier and BTE modeling methods to characterize the thermal effects from nanometer-scale devices to centimeter-scale chips and packages. This section details the thermal physics models and how they are used in ThermalScope.

1) Fourier model: The steady-state classical Fourier model is characterized by the following:

$$\nabla \cdot (K\nabla T) + q_{\text{vol}} = 0 \tag{3}$$

where K is the thermal conductivity, T is the temperature, and q_{vol} is the volumetric heat source. The finite-volume method is typically used to solve this equation by partitioning the domain into numerous discrete elements and transforming it into a set of discretized equations [39] as follows:

$$-2(K_x + K_y + K_z)T_p + K_x(T_e + T_w)$$
$$+ K_y(T_n + T_s) + K_z(T_t + T_b) + q_{\text{vol}}\Delta x\Delta y\Delta z = 0 \tag{4}$$

where K_x, K_y, and K_z are the thermal conductances that element p shares with its neighbors in the x, y, and z directions

and Δx, Δy, and Δz are the element sizes in the x, y, and z directions, and the subscripts e, w, n, s, t, and b refer to the east, west, north, south, top, and bottom directions relative to element p. Equation (4) can be rewritten as

$$a_p T_p = \sum_{nb} a_{nb} T_{nb} + b_p \qquad (5)$$

where nb represents the neighbors of element p, the coefficients a_p and a_{nb} are functions of K_x, K_y, and K_z, and $b_p = q_{\mathrm{vol}} \Delta x \Delta y \Delta z$.

The Fourier model is capable of accurately modeling the thermal effects only at feature length scales much longer than the mean free path of phonons [27]. Since it is much more efficient than the BTE method, ThermalScope uses the Fourier method to model the thermal effects from the chip–package level down to the functional-unit level. The computationally expensive BTE model is only used at the device level.

2) BTE model: As described in Section I, models that can capture the nanometer-scale thermal effects include molecular dynamics, the BTE model, and the ballistic-diffusion model. Molecular dynamics methods are not suitable for the multi-scale thermal analysis problem because of their extremely high computational complexity, while the ballistic-diffusion model provides insufficient accuracy. ThermalScope uses the gray phonon BTE under the relaxation time approximation to model the device regions. The gray BTE is a widely used model which does not require complex material parameters or parameter relations that may not be easily obtainable, such as the phonon dispersion curves (which relate the phonon frequency and the wave vector) used for the frequency dependence of phonons, to be known. Although the gray BTE model is used here, other approximations for solving the BTE, such as those used in the semigray model, could be applied. However, it should be noted that, when using the semigray model, the choice for determining which types of phonons are responsible for heat transport is critical [27], [29]. The gray BTE model employs a phonon distribution function e'' and assumes a single group velocity and relaxation time for phonons, which are independent of their frequency and polarization. Although it is possible to account for the interactions of phonons with differing frequencies (e.g., for 1-D systems [24]), it increases the computational complexity significantly, making thermal analysis prohibitively expensive. In addition, the relaxation times for the phonon–phonon interactions for different phonon modes are not well understood [40], and the approximations in theories used to obtain the frequency dependence of relaxation times have led to large uncertainties [17]. The relaxation time approximation used for the gray BTE is valid when the length scales are larger than the heat carrying phonon wavelengths [28] and allows the phonon scattering processes to be taken into account as a deviation from the equilibrium distribution. The steady-state BTE equation using the gray model and relaxation time approximation is as follows [21]:

$$\nabla \cdot (\vec{s} v_g e'') = \frac{e^0 - e''}{\tau_{\mathrm{eff}}} + q_{\mathrm{vol}} \qquad (6)$$

where $\vec{s}$ is the phonon propagation direction, v_g is the group velocity of the phonons, e'' is the energy density per unit solid

angle of the phonons, e^0 is the equilibrium energy density, τ_{eff} is the relaxation time, and q_{vol} is the volumetric heat source. The equilibrium energy is given by [21]

$$e^0 = \frac{1}{4\pi} \int_{4\pi} e'' d\Omega = \frac{1}{4\pi} C(T_L - T_{\mathrm{ref}}) \qquad (7)$$

where Ω is the angular discretization, C is the specific heat, T_L is the lattice temperature, and T_{ref} is the reference temperature for the specific heat. The lattice temperature T_L can be calculated using (7), once the equilibrium energy density is known.

The relaxation time τ_{eff} is the time between independent scattering events and can be found using the bulk material equation

$$k = \frac{1}{3} C v_g^2 \tau_{\mathrm{eff}} \qquad (8)$$

where k is the thermal conductivity.

The electron–phonon interactions that occur inside devices are modeled by heat sources, which are denoted by the term q_{vol} in (6). Its value can be derived from the power consumption information of the device, which can be obtained using circuit simulation.

Different types of boundary conditions are modeled in our solver, such as specular and diffuse boundaries [30]. In addition, the model for the Fourier/BTE interface contains a reflection coefficient parameter, which can be varied according to the material properties at the two sides of the interface [30]. The reflection coefficient allows the accurate modeling of phonon interactions at the interfaces of materials.

Applying the finite-volume method to discretize (6) and noting that the angular domain is also discretized to account for different phonon directions, we obtain

$$v_g \Delta y \Delta z S_x \left(e''_{i,e} - e''_{i,w} \right) + v_g \Delta x \Delta z S_y \left(e''_{i,n} - e''_{i,s} \right)$$
$$+ v_g \Delta x \Delta y S_z \left(e''_{i,t} - e''_{i,b} \right) = \left(\frac{e^0_p - e''_{i,p}}{\tau_{\mathrm{eff}}} + q_{\mathrm{vol}} \right) \Delta x \Delta y \Delta z \Delta \Omega \qquad (9)$$

where $\Delta \Omega$ is the angle extent resulting from angular discretization, i is the index for the angle of propagation, and S_x, S_y, and S_z are the x, y, and z components obtained from the following integration:

$$\vec{S} = \int_{\phi - \Delta\phi/2}^{\phi + \Delta\phi/2} \int_{\theta - \Delta\theta/2}^{\theta + \Delta\theta/2} \vec{s_i} \sin \theta_i d\theta d\phi \qquad (10)$$

where θ and ϕ are the polar and azimuthal angles, respectively. The rest of the terms in (9) are the same as defined previously. Equation (9) can be rewritten as follows:

$$a_p e''_{i,p} = \sum_{nb} a_{nb} e''_{i,nb} + b_p \qquad (11)$$

where a_{nb} are the coefficients that relate the energy density of element p to its neighboring elements and b_p is a function of e^0_p, τ_{eff}, Δx, Δy, Δz, and $\Delta \Omega$. Note that (11) is only for a single angular direction $\vec{s_i}$.

Obtaining the temperature of an element using the BTE involves three steps. First, (11) is solved for different angular directions. Second, the solution is substituted in the first equality of (7) to obtain the equilibrium energy density, and lastly, the second equality is used to evaluate the temperature of the element.

As can be noted from the previous derivations, solving the BTE equation is much more complex than the Fourier equation because of the following: 1) All the angles need to be considered for each element, and 2) (11) is, in fact, implicit, since the term b_p is a function of the equilibrium energy density e_p^0, which slows down convergence of the iterative solver.

The technique we use for solving (5) and (11) in our hybrid Fourier/BTE approach is the tridiagonal matrix algorithm [41].

B. Hybrid Fourier/BTE Analysis Method

As indicated in Section IV-A, the main difficulty associated with the BTE model is its high computational complexity. To overcome this difficulty, we propose a hybrid Fourier/BTE method which combines the best of both models. Compared with the BTE method, the Fourier method is orders of magnitude faster, but can be inaccurate when used for analysis of nanoscale devices. However, its accuracy is sufficient if used to model the regions beyond the mean free path of phonons, providing significant simulation time savings. The flow of the hybrid approach is described next, while its accuracy and efficiency are evaluated in Section VI-A.

Fig. 2 shows the hybrid Fourier/BTE flow, which consists of the following stages.

1) Unified Fourier and BTE adaptive solver: The hybrid solver leverages both Fourier and BTE models to offer accurate and efficient thermal analysis. The appropriate modeling technique is selected based on a distance measure. The distance $\eta \times v_g \tau_{\text{eff}}$ surrounding the devices is chosen as the BTE region, where η is a constant; $v_g \tau_{\text{eff}}$ is defined as the mean free path of a phonon, where v_g is the phonon group velocity and τ_{eff} is the effective relaxation time. Here, we assume that the phonon mean free path is independent of the device size and structure. Varying the constant η changes the number of elements in the BTE region and, thus, the physical area modeled using BTE as well. The effect of changing the constant η is evaluated in Section VI-A. The rest of the structure outside of this region is analyzed using the Fourier solver.

2) Fourier/BTE boundary update: In the hybrid solver, the Fourier and BTE solvers are invoked iteratively. The boundary temperatures of the Fourier/BTE region interfaces and the heat flow into the Fourier region are updated after each iteration. Once convergence is reached, the thermal profile of the entire structure is reported.

3) Device-level lookup-table generation: In order to obtain accurate temperature profiles at the finest granularity level, we must consider nanoscale thermal effects. Due to the prohibitively expensive simulation time of a BTE-based (and even a Fourier/BTE based) solver, it is not feasible to simulate the thermal profiles for each device in a billion-transistor chip. To address this issue, we construct a compact modeling method to enable accurate and efficient device-level thermal analysis.

Using the hybrid Fourier/BTE solver, a compact model, i.e., a lookup table, is derived to model the device-level thermal effect. This compact model contains device-level temperature information for different device geometries, power consumptions, and technologies. During full–chip thermal-profile evaluation, the lookup table is consulted to obtain intradevice thermal effects.

V. MULTISCALE THERMAL ANALYSIS

This section describes the proposed multiscale modeling techniques which optimize the efficiency of thermal analysis while capturing the diverse thermal behavior at all scales from chip–package level to device level.

Modern ICs contain hundreds of millions of devices. The temperature of a device i is influenced by the power consumptions of all the on-chip devices as follows:

$$T_i = f_i(P_1, \ldots, P_N) = r_{i,1} \times P_1 + \cdots + r_{i,N} \times P_N \quad (12)$$

where T_i is the temperature of device i, P_j is the power of device j, and N is the total number of devices. $r_{i,j}$ is defined as the thermal impact coefficient, which indicates the impact of a unit power consumption of device j on the temperature of device i. Given N devices, $\mathbf{T}_{N \times 1} = \mathbf{R}_{N \times N} \times \mathbf{P}_{N \times 1}$, where $\mathbf{T}$ and $\mathbf{P}$ are the vectors of temperatures and powers of on-chip devices. $\mathbf{R}$ is called the thermal impact coefficient matrix. It is the inverse of the thermal conductance matrix $\mathbf{K}$ [see (3)]. Note that $\mathbf{K}$ is an $M \times M$ matrix, where M is the total number of elements of the whole chip and package partition and $M > N$. In other words, $\mathbf{R}$ is a submatrix of $\mathbf{K}^{-1}$, containing the coefficients corresponding to the device elements. To model the device-level thermal effect accurately, each device needs to be partitioned into a large number of elements. The size of each element is in the nanometer scale. Given a modern IC design containing hundreds of millions of nanometer-scale devices with a package on the centimeter length scale, matrix $\mathbf{K}$ will contain a massive number of elements, i.e., M is an extremely large number. Computing (12) is a daunting task.

ThermalScope uses the following techniques to optimize thermal modeling efficiency: 1) hierarchical multiscale spatial partitioning (Section V-A); 2) thermal impact clustering (Section V-B); and 3) thermal-gradient-based spatial resolution adaptation (Section V-C).

A. Hierarchical Multiscale Spatial Partitioning

As shown in (12), to characterize device i's temperature T_i, we need to consider the thermal impact of all on-chip devices. To simplify the thermal analysis process, we propose a hierarchical multiscale spatial partitioning method. The basic idea of this technique is as follows. Interdevice thermal interaction, characterized by interdevice thermal impact coefficients, is strongly influenced by the distances between devices. Starting from the local neighborhood of the device of interest, i.e., device i, the thermal impact coefficient $r_{i,j}$ (which characterizes the thermal impact of a unit of power consumption of device j on device i) decreases significantly with the increase of the distance between devices i and j. Outside of the local
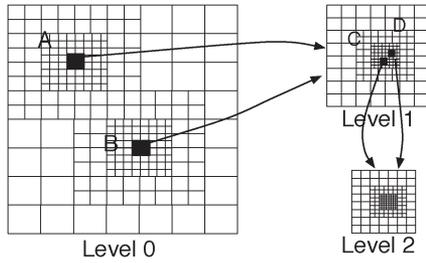
Fig. 3. Hierarchical multiresolution partitioning.

neighborhood, the thermal impact coefficient $r_{i,j}$ decreases slowly with the increase of interdevice distance. The proposed spatial partitioning method leverages such characteristics by adaptively adjusting the spatial partitioning, i.e., fine-grain partition is applied to the local neighborhood of the device of interest to characterize the heterogeneous thermal impacts of neighboring devices, and coarse-grained partitions are used to characterize the thermal effects of remote devices. The temperature equation for device i is as follows:

$$T_i = \zeta_{i,1} \times \xi_1 + \cdots + \zeta_{i,L} \times \xi_L \qquad (13)$$

where $\zeta_{i,j}$ is the thermal impact coefficient of partition j and ξ_j is the total power consumption of the devices inside partition j. L is the total number of partitions required to accurately model device i's temperature. As shown in Fig. 3, since coarse-grained partitioning can be used in most locations, i.e., $L \ll N$, (12) is greatly simplified. Therefore, adaptive spatial partitioning can reduce the modeling complexity. However, as the device-level thermal effect becomes increasingly significant, nanometer-scale fine-grain modeling in the device neighborhood is required to accurately estimate the temperature of each device. Therefore, L in (13) is still a large number. Given N devices, the total memory usage is thus proportional to $N \times L$.

The proposed hierarchical multiscale spatial partitioning technique further improves thermal analysis efficiency and reduces memory usage by sharing common partitions hierarchically when computing the temperatures of different devices. This concept is shown in Fig. 3, which shows a three-level hierarchical adaptive partitioning. At the functional-unit level (level 0), if we are interested in the temperatures of two devices i and j located at the center of regions **A** and **B**, respectively, we can use a coarse-grained partition for regions far away from **A** and **B**. For the fine-grained partition, **A** and **B** will share similar local dependencies and, thus, can be characterized by the same fine-grained partitioning (level 1). These fine-grained partitions also contain heterogeneous partitioning surrounding areas/devices of interest. Assuming that we are interested in devices in regions **C** and **D**, which are located in **A** and **B**, respectively, we can further reuse the fine-grained partitioning at level 2. Therefore, the temperature equations for devices i and j are simplified as follows:

$$T_i = f_{i,\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{\mathrm{Level}_2}(\dots)$$
$$T_j = f_{j,\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{\mathrm{Level}_2}(\dots) \qquad (14)$$

i.e., these two temperature equations share the same level 1 and 2 equations. On the other hand, if we are interested in

the temperatures of two devices i and j both in region **C**, the temperature equations for devices i and j are simplified as follows:

$$T_i = f_{\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{i,\mathrm{Level}_2}(\dots)$$
$$T_j = f_{\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{j,\mathrm{Level}_2}(\dots) \qquad (15)$$

i.e., these two temperature equations share the same level 0 and 1 equations. Therefore, this hierarchical multiscale spatial partitioning method can reduce the modeling complexity significantly.

B. Clustering

Due to symmetry, distance, or material properties, the thermal impact of devices in different regions of the spatial partition may be equivalent. This leads to the storage of redundant information and inefficiencies in the total number of computations. To increase the efficiency of the compact model, ThermalScope uses clustering to minimize the amount of redundant information by grouping equivalent thermal elements into a single representative element.

A hierarchical clustering technique is used to simplify each row of the thermal impact coefficient matrix. A single row of the thermal impact coefficient matrix contains all coefficients required to describe the thermal impact of every element on a given level X to element i. By reducing the total number of thermal impact coefficients and, thus, thermal elements, clustering looks to simplify the following:

$$f_{i,\mathrm{Level}_X}(P_1, \dots, P_Q) = r_{i,1} \times P_1 + \cdots + r_{i,Q} \times P_Q \qquad (16)$$

where Q is the number of elements on level X.

The clustering algorithm flow is as follows. The thermal impact coefficient matrix is subdivided into single row vectors. Elements in the row are sorted to increase the efficiency of clustering. The elements are then clustered using a hierarchical clustering algorithm where elements having similar thermal impacts within a certain threshold are grouped together and assigned a representative thermal impact value taken as the average of all elements within the cluster. Clustering elements of each row vector together leads to significant reduction in the coefficient matrix size. This clustering technique is applied to the thermal impact coefficient matrix of each granularity level.

C. Thermal-Gradient-Based Spatial Resolution Adaptation

Due to the highly varying power profile at the chip–package level and the existence of multiple material layers, including the heat sink, an efficient thermal analysis method with the ability to model fine-grained features is required. To handle this problem, ThermalScope uses an adaptive Fourier solver to model heat flow from the functional-unit level to the chip–package level. This solver contains an adaptive spatial refinement scheme that allows the chip to be partitioned heterogeneously, providing fine-grain modeling elements where there are significant thermal fluctuations and coarse-grained elements where the temperature is relatively constant.

Using the power profile and chip–package material properties and dimensions, a Fourier solver is invoked. Spatial refinement is carried out, and the thermal-gradient conditions are verified. If the threshold thermal-gradient condition has not been met, the Fourier solver is invoked again. This process continues until all elements have satisfied the thermal-gradient condition, i.e., all the neighboring elements' temperature differences are below the threshold. Assuming that the temperature of thermal element i is T_i and S is the threshold, the new number of elements Q will be

$$Q = \left\lceil \log_2 \frac{T_i - T_j}{S} \right\rceil, \qquad (17)$$

which specifies that, based on a threshold S and the temperature difference between element i and its neighbors, the solver decides if element i will be further split into a number of elements that is a power of two (i.e., 2, 4, 8,...) or will not be split at all. S determines the sensitivity of partitioning to the temperature difference, with a lower value forcing more partitioning and vice versa. This condition is tested in all three dimensions to determine the final partition of element i.

Once the heterogeneous partition is obtained for the chip–package level to the functional-unit level, it can be used in conjunction with the compact model to obtain the full-chip thermal profile. The temperatures of all thermal elements, which are obtained during the chip–package thermal analysis stage, are used in (13) to describe the chip–package-level temperature effect on devices as follows:

$$T_i = \zeta_{i,1} \times \xi_1 + \cdots + \zeta_{i,K} \times \xi_K + T_{\text{sre}} \qquad (18)$$

where T_i is the temperature of device i, $K < L$, where, as before, L is the total number of partitions required to model the temperature of device i and T_{sre} is the temperature of the chip–package-level spatially resolved element, i.e., the temperature of the coarse-grained element that device i belongs to (whose self-thermal impact has been removed to assure that it is not accounted for twice). The thermal impact due to partitions 1 to K are handled by the compact model, which accounts for levels between (and including) the functional-unit and device levels.

VI. Results

In this section, we evaluate ThermalScope, the proposed multiscale thermal analysis method. ThermalScope unifies Fourier and BTE modeling techniques as well as a multiscale macromodeling method. Since there are no available tools for the direct comparison of the proposed multiscale thermal analysis method, due to its ability to simultaneously handle thermal analysis at scales ranging from chip–package to device level, we evaluate the proposed methods at different levels of granularity.

1) In Section VI-A, we evaluate the hybrid analysis method using device-level thermal analysis.
2) In Section VI-B, we demonstrate that interdevice thermal interaction can be accurately modeled using Fourier thermal analysis.

TABLE I
ACCURACY EVALUATION FOR BTE METHOD FOR
VARIOUS ACOUSTIC THICKNESSES

n	0.01	0.1	0.5	1	10	100
e_{avg} (%)	0.64	1.33	0.70	0.96	0.56	0.87

3) In Section VI-C, we examine the full-chip thermal modeling capability and evaluate the chip–package- and functional-unit-level modeling accuracy by comparing it with COMSOL, a commercial physics modeling package [4].
4) ThermalScope is developed to target billion-transistor nanometer-scale IC designs. We report our experience using ThermalScope for thermal analysis and temperature-dependent leakage analysis of an industry IC design with over 150 million transistors in Section VI-D.

A. Device-Level Thermal Modeling Using Hybrid Fourier/BTE Analysis

In this section, we show that the BTE method is necessary for the accurate computation of device-level thermal profiles. We then evaluate the accuracy and speedup of the hybrid Fourier/BTE method.

1) Accuracy of the BTE method: ThermalScope uses a hybrid BTE/Fourier solver to model the thermal profile at the device level. To evaluate the accuracy of the BTE component for length scales below and above the mean free path of phonons, we have modeled the Heaslet and Warming problem [42]. In this problem, a block of material has two opposing walls, separated by distance L, held at different temperatures, while all other walls are insulating. The distance between the two fixed-temperature walls is varied, and the resulting thermal gradients are observed. The number of mean free paths n is used to characterize the distance between the walls. For example, $n = 1$ refers to the length of the structure equaling the distance of one mean free path.

The two types of meshes, namely, structured and unstructured meshes, have been used in numerical analysis methods. In our solver, we use a structured mesh with a large number of elements to guarantee numerical accuracy. The accuracy of our solver was verified via comparison with the results of Rutily *et al.* [43]. Table I shows the error obtained for structures with various lengths. In the simulations, a total of 4000 elements were used in the direction of the temperature gradient. Using this granularity, all results had differences of less than 0.36% when the number of elements was double from 2000 to 4000. The error was determined by using the following:

$$e_{\text{avg}} = 1/|E| \sum_{e \in E} |T_i - T_i'| / T_i \qquad (19)$$

where E is the set of points used in [43], at which the temperatures are evaluated, T_i is the temperature at location i along the structure [43], and T_i' is the temperature at location i along the structure obtained using ThermalScope. As can be seen in Table I, the results of ThermalScope are in excellent agreement with those of Rutily *et al.* [43].

2) BTE method versus Fourier analysis: Here, we show the inaccuracy of the Fourier model in capturing the device-level
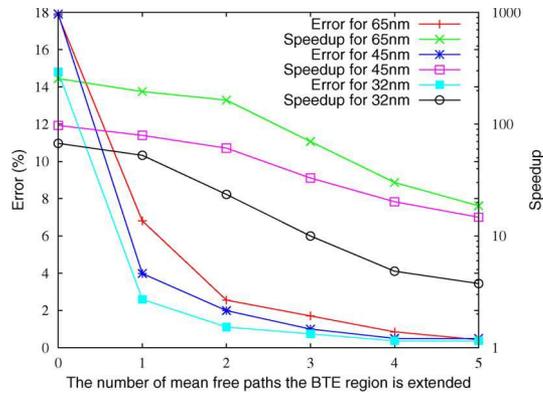
Fig. 4.   Accuracy and efficiency of the hybrid solver.

Fig. 5.   Interdevice thermal correlation analysis for bulk silicon/FinFET devices.

thermal effects by comparing it to the hybrid Fourier/BTE method. We simulate a $910 \times 910 \times 500$-nm region containing a bulk silicon device at different technology nodes (65, 45, and 32 nm). We evaluate the error of the Fourier method using $(T_{\mathrm{BTE}} - T_{\mathrm{Fourier}})/(T_{\mathrm{BTE}} - T_a)$, where T_{BTE} is the peak temperature of the device when solved using the BTE solver, T_{Fourier} is the peak temperature of the device when solved using the Fourier solver, and T_a is the ambient temperature. The ambient temperature is subtracted from the denominator to give a more conservative error since it is reported relative to the ambient temperature as opposed to 0 K. Compared with the BTE method, the Fourier method introduces 34.0%, 44.8%, and 54.1% error at 65-, 45-, and 32-nm technologies, respectively. This analysis shows a clear trend that the error of the Fourier method increases as device size decreases, which is expected since the Fourier model becomes less accurate as the length scales approach the mean free path of phonons. Therefore, if used alone, the Fourier method is unable to model the thermal effects of nanometer-scale structures.

3) Hybrid method versus BTE analysis: The idea of the hybrid method is to leverage the advantages of both Fourier and BTE methods. The BTE method is only used when necessary, e.g., for regions within the mean free path of phonons from device heat sources. The Fourier method is applied to other regions to speed up thermal analysis. To test the accuracy of the hybrid method, we use the same setup described previously. This material is partitioned into 343 128 thermal elements. We first apply BTE analysis to the whole system. The overall simulation time was 16.3 h. Next, we use the hybrid approach, and we vary the number of elements solved using the Fourier method by changing η, which is the number of mean free paths that the BTE region extends from the heat source. We report the relative temperature differences and the speedups compared with the BTE-only method. The test setup is repeated for 45- and 32-nm technologies. Fig. 4 shows the results. This study indicates that the hybrid method can accurately model the thermal effect beyond the mean free path of phonons using the Fourier method, with speedups ranging from 23 to over 150 times with an error of less than 4% and a 10–70 times speedup with an error of less than 2%.

Note that this analysis only considers the device and its local neighborhood. The chip–package material outside of the mean free path of phonons, such as silicon substrate, packaging, and
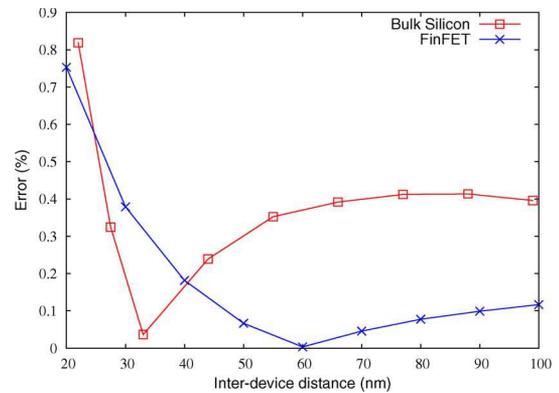
cooling structure, are not considered. These structures account for the vast majority of the analyzed system, and it is known that Fourier analysis is capable of accurately modeling them. From the results in this section, we conclude that the hybrid method greatly accelerates the simulation process of IC full-chip thermal analysis compared with BTE-only approach with only slight degradation in accuracy.

Results from the hybrid method simulations are used to construct a lookup table, which is used during the full-chip thermal-profile evaluation. Thus, the complexity of generating the lookup table is associated with the simulation times. Depending on the device structure and geometry, simulations can take as little as 0.5 up to 7 h. Although slow, those simulations need to be carried out only once for each device with a certain structure and geometry, and thus, we achieve significant time savings in the full-chip thermal analysis by simulating once per device type instead of once per device instance.

B. Interdevice Thermal Effect Modeling Using Fourier Analysis

The goal of this analysis is to demonstrate whether the Fourier method is sufficient to accurately model the thermal interaction between neighboring devices. This would allow us to apply the Fourier model for everything but characterizing individual devices, i.e., from chip-level analysis all the way down to, but not including, device-level analysis (which was described in the previous section). At the device level, only the device of interest needs to have its temperature effect computed using the BTE model.

We evaluate interdevice thermal correlation using both the hybrid Fourier/BTE and BTE-only methods. We report the peak temperature of one of the two devices when the BTE solver is used for both and compare it against the peak temperature of the same device when its neighbor has been solved using the Fourier model. We repeat this simulation for different interdevice distances. This study allows us to determine the accuracy of Fourier-based interdevice thermal correlation analysis, as well as the length scale at which the BTE model becomes necessary.

Fig. 5 shows the error of Fourier-based interdevice thermal correlation analysis as a function of interdevice distance for

both bulk silicon and FinFET devices. The analysis error is defined as $|(T_{\text{BTE}} - T_{\text{Fourier}})/(T_{\text{BTE}} - T_a)|$, where T_{BTE} is the peak temperature of the device when its neighbor is solved using the BTE solver, T_{Fourier} is the peak temperature of the device when its neighbor is solved using the Fourier solver, and T_a is the ambient temperature. As shown in Fig. 5, the error of Fourier-based interdevice thermal correlation analysis decreases as the interdevice distance increases, and thermal effects become less significant. It should be noted that, because we are using the absolute error, the minima appearing in the curves in Fig. 5 are the crossing points between under- and overestimation of the temperature. Compared with the BTE method, the Fourier method can accurately estimate interdevice thermal effects with less than 1% error even when the interdevice distance is as low as 20 nm for both bulk silicon and FinFET devices, which suggests that the Fourier method can provide sufficient accuracy for interdevice thermal correlation analysis, and thus, only individual devices of interest will need to have their thermal profiles computed using the BTE model.

C. Chip–Package- and Functional-Unit-Level Accuracy Evaluation

To evaluate the chip–package- and functional-unit-level modeling accuracy of ThermalScope, we compare it against COMSOL, a commercial physics modeling package [4], using a quad-core chip-multiprocessor design. The chip design contains four Alpha 21 264 cores and an L2 cache. Each core contains 15 functional units. The silicon die is 9.88×9.88 mm, with a 50-μm thickness. There is a 10-μm layer of thermal grease between the heat sink and die, and the extruded copper heat sink is 9.88×9.88 mm with a thickness of 6.9 mm. The functional-unit power profile of the on-chip cores depends on the programs running on the cores. In our evaluation, we consider running 17 different multithreaded and multiprogrammed benchmarks, which are listed in the top row of Table III. The benchmarks are from the SPEC benchmark suite [44], [45]. Each benchmark has different functional-unit requirements and thus generates a different power profile on the multicore chip, for instance, a benchmark containing floating-point programs would highly utilize the floating-point units of the cores, which would lead to a high power consumption in those units. The functional-unit-level power profile (containing static and dynamic power breakdown) of each benchmark was obtained by the M5 full-system simulator [46] with a Wattch-based EV6 power model [47].

The temperature profiles for benchmark *Cholesky* obtained using COMSOL and ThermalScope are shown in Figs. 6 and 7, respectively. Table II reports the results for the functional units in all four cores. The modeling error of ThermalScope, err $= |(T_{\text{COMSOL}} - T_{\text{ThermalScope}})/(T_{\text{COMSOL}} - T_a)|$, for each functional unit is calculated against COMSOL. The results show a maximum of 3.95% error for ThermalScope compared with COMSOL, with an average error of 2.14% for all functional units.

Table III shows the results of the 17 testing cases. The average error e_{avg} from (19) is used for comparing the thermal profile of the entire chip where E is the set of elements of the
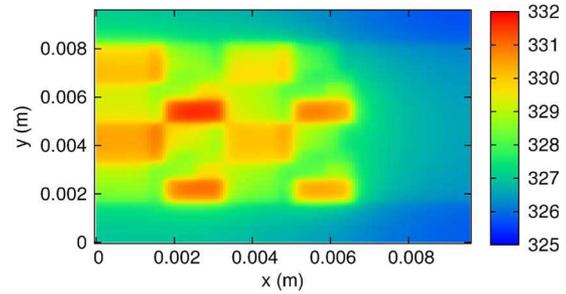
Fig. 6. Temperature profile for benchmark Cholesky using COMSOL.
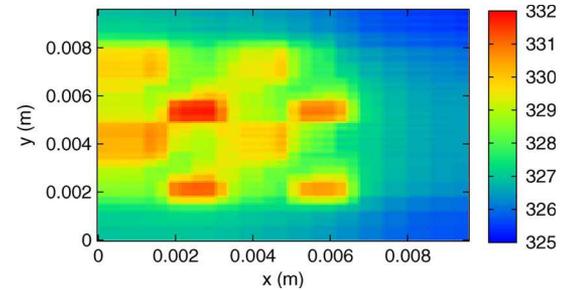
Fig. 7. Temperature profile for benchmark Cholesky using ThermalScope.

active layer, T_i is the temperature of element i obtained using COMSOL, and T_i' is the temperature of element i obtained using ThermalScope. As in Table II, the errors were calculated relative to the ambient temperature. The results show a maximum error of 1.97%, while the average error for all benchmarks is 1.77%.

D. Test Case: Full-Chip Thermal Analysis and Temperature-Dependent Leakage-Power Estimation

ThermalScope is designed for the thermal analysis of billion-transistor nanometer-scale ICs. In this section, we demonstrate the use of ThermalScope in full-chip thermal analysis and temperature-dependent leakage analysis using an industry design containing over 150 million transistors.

The configuration of the chip design considered in this analysis is as follows. The silicon die is 16×16 mm, with a 725-μm thickness for bulk silicon technology and 202-μm thickness (including the oxide layer) for FinFET technology. The aluminum heat sink is 34×34 mm with a 2-mm-thick base and 23-mm fin height. The chip uses flip-chip packaging and a layer of interface material between the silicon die and cooling solution. The air-cooling flow rate is 1.5 m/s.

We will now evaluate the potential simulation time and memory storage savings of the proposed technique. For device-level thermal analysis, we require elements to be much smaller than the heat source. Assuming that the heat source is the size of the device and the process technology is 65 nm, we require the element size to be a few nanometers along each dimension. At the other end of the spectrum, the sizes of the chip and cooling package are in the range of centimeters. To construct a partition of the industry design with over 150 million transistors, the storage requirements would be on the order of 10^{18} B. The computations required to evaluate the temperature of a single device would be 10^{12} additions and 10^{12} multiplications. From

TABLE II
ACCURACY EVALUATION USING BENCHMARK CHOLESKY

	Core	Functional unit													
		Icache	ITB	Bpred	LdStQ	IntMap	FPMap	IntQ	FPQ	IntReg	FPReg	Dcache	DTB	IntExec	FPAdd	FPMul
err (%)	Core 1	1.05	1.59	1.54	1.97	1.24	2.45	1.61	1.99	3.30	1.74	1.05	1.68	2.87	2.23	2.08
	Core 2	2.39	1.60	3.01	2.01	2.51	2.73	2.62	2.31	3.75	3.95	1.55	2.52	2.75	2.58	3.67
	Core 3	1.99	1.74	2.14	2.07	1.73	1.95	1.73	1.84	2.49	1.77	1.37	1.74	2.40	2.72	2.39
	Core 4	1.74	1.91	2.81	1.61	1.58	1.59	2.58	1.19	3.46	2.30	1.67	2.47	2.12	1.66	1.52

TABLE III
ACCURACY EVALUATION USING 17 BENCHMARKS

Benchmarks	cholesky	gzip, mgrid	parser, vpr	gzip, parser	mgrid, vpr	lu	applu, gcc	twolf, mcf	applu, mcf
e_{avg} (%)	1.77	1.83	1.88	1.83	1.81	1.67	1.77	1.80	1.77
Benchmarks	gcc, twolf	jpegdec, gsmenc	gsmenc, g721enc	mpgenc	radix	sphinx3	waternsq	waterspa	
e_{avg} (%)	1.78	1.97	1.97	1.82	1.60	1.92	1.64	1.20	

TABLE IV
EFFICIENCY EVALUATION

	Bulk silicon		FinFET	
	ThermalScope	No clustering	ThermalScope	No clustering
t_{CPU} (m)	167	485	179	607
Mem. (MB)	548	604	620	604

Fig. 8. Bulk full chip.

Fig. 9. Bulk 255×255 μm.

this example, we see that device-level thermal analysis of entire chips is computationally intensive.

ThermalScope uses several methods to reduce the storage requirements and total amount of computation. Hierarchical adaptive modeling granularities are used from the chip level down to the device level. This adaptive modeling reduces the problem size to requiring storage on the order of 10^8 B for the thermal impact coefficient matrices for the same problem as described previously, an improvement of ten orders of magnitude. For comparison purposes, the input power profile of the industry design itself requires more than 7×10^8 B of storage. The number of computations required to evaluate the temperature of a single device would also be reduced to 10^8 additions and multiplications, and the results from the majority of these computations can be reused among devices. The amount of computation is further reduced by ThermalScope's clustering technique. The simulation run-time and memory usage results for the device-level temperature evaluation (after obtaining the coefficient matrices) for our proposed technique with and without clustering are shown in Table IV. The chip evaluated contained over 150 million devices. We evaluated both a bulk silicon design and a FinFET design. The results show that, although the memory usage may not be significantly reduced by clustering, significant speedup can be achieved. For the clustering technique, memory usage for indexing is required in addition to storing the clustered information, which can explain the lack of significant memory reduction.

1) Thermal analysis and temperature-dependent leakage-power estimation: Accurate thermal analysis is critical for evaluation of temperature-dependent effects. ThermalScope is capable of handling large IC designs with device-level accuracy. In this section, we report the use of ThermalScope for full-chip thermal analysis and temperature-dependent IC leakage analysis of a large industry design.

Since the leakage power of the chip is strongly affected by the temperature, it is necessary to include leakage-power estimation in the thermal analysis simulation flow. To determine the thermal profile of the industry chip while taking into account
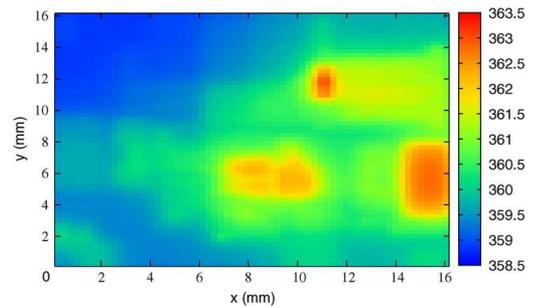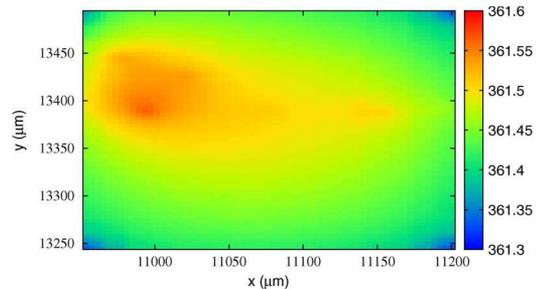
the leakage power, the following iterative process can be used. From the data set of the industry design, the initial dynamic and leakage power are estimated at the ambient temperature of 55 °C. The device-level thermal profile is then evaluated for the given initial power profile. The results of this simulation are then used to update the leakage power of the chip. This is an iterative process that continues until convergence is reached between the simulated temperature and the power. In this study, the temperature-leakage-power dependence is obtained by curve fitting the leakage measurement results of the industrial design data set, which contains power numbers for various temperatures.

We consider both bulk silicon and FinFET technologies. The thermal profile of the IC design is characterized using the multiscale macromodeling method through the described iterative analysis process. During thermal analysis, the temperature of every individual device is evaluated, and the leakage power of each device is adjusted based on its change in temperature. This process is carried out for every single device on the chip. The temperature profiles obtained for three different levels of granularity are shown in Figs. 8–10 for bulk silicon technology

Fig. 10. Bulk 1.6×1.6 μm.

Fig. 11. FinFET full chip.

Fig. 12. FinFET 255×255 μm.

Fig. 13. FinFET 1.6×1.6 μm.

TABLE V
LEAKAGE-POWER ESTIMATION

	P_1	P_2	P_3	P_4
Bulk silicon $P_{leakage}$ (mW)	13277.4	16452.5	16454.3	16563.4
FinFET $P_{leakage}$ (mW)	13277.4	16565.0	16242.6	16983.8

The profiles also indicate, however, the inaccuracies of coarse-grained estimates for device temperatures.

Figs. 8–13 show the information lost when device-level thermal analysis is not considered. Using coarse-grained thermal analysis, large inaccuracies occur due to the assumption that all devices within a single coarse-grained element have the same temperature as the element. For the bulk silicon design, at the intermediate level (255×255 μm), this may be a valid assumption; however, at the device level, we clearly see a significant deviation from the average coarse-grained temperature. This demonstrates that thermal analysis of the entire chip at the intermediate level would not be sufficient to characterize the device temperatures. In contrast, ThermalScope determines the temperature of each device on chip which allows for detailed full-chip thermal analysis. In coarse-grained thermal analysis, the features that occur at the device level are not considered, which leads to inaccurate estimation of the device temperatures, as shown in Figs. 8–13.

Chip power consumption is one of the critical characteristics guiding IC design decisions. With technology scaling, the contribution of leakage-power consumption to total power consumption increases. Thus, it is important to provide IC designers with accurate leakage-power information to help them evaluate the different design tradeoffs. In addition to thermal analysis, ThermalScope can also be used to estimate the leakage power of the chip. The leakage power is determined by the same iterative process described earlier. For comparison purposes, we compare the results of the leakage power obtained using four distinct techniques.

The first leakage-power value to be compared P_1 is the leakage power from the industrial benchmark data set for the ambient temperature of 55 °C. The second leakage power P_2 was obtained by estimating the leakage power after full-chip thermal analysis, using device-level modeling granularity. The third and fourth leakage powers were evaluated using the iterative process. The iterative process was carried out for both the coarse-grained thermal analysis (chip divided into 64×64 elements) for P_3 and full-chip thermal analysis, using the device-level modeling granularity for P_4. By comparing the leakage power obtained with the various techniques, we can gain insight into the importance of device-level information on leakage-power estimation and the significance of iterative solutions. The leakage-power results are presented in Table V.

The results indicate that iterative solutions converge to a significantly higher leakage power, and thus, single iteration evaluation methods are not sufficient for accurate leakage-power estimation. The results also demonstrate the effect of considering device-level thermal behavior during leakage analysis. For both the FinFET and bulk silicon designs, the leakage power reported using multiple iterations of device-level thermal analysis is higher than the other leakage powers reported. The leakage power profile of the industry design, obtained using the

and Figs. 11–13 for FinFET technology. Figs. 9 and 12 show the enlarged fine grain thermal profiles of a hot spot on the chip. Figs. 10 and 13 show a further enlargement of the area, showing the device-level information for two devices out of the hundreds of millions whose temperatures have been reported. Although ThermalScope evaluates the temperature of every device, it also has the capability of coarse-grained thermal analysis. The thermal profiles demonstrate the capability of ThermalScope to handle analysis at scales varying by six orders of magnitude.

Fig. 14.   Leakage-power profile of the industry design.

iterative full-chip thermal analysis using device-level modeling granularity technique, is shown in Fig. 14.

From the results presented in this section, we can conclude that device-level thermal analysis is necessary for both accurate thermal-profile information and leakage-power estimation. By using a compact macromodeling method, ThermalScope is able to obtain such information within reasonable time frames and storage requirements.

VII. CONCLUSION

Thermal analysis and optimization are now critical in nanometer-scale IC design. The goal of this work has been to develop thermal modeling techniques that are accurate at nanometer length scales and also computationally efficient for full-chip thermal analysis. To achieve this goal, we have developed ThermalScope, a multiscale thermal analysis solution. It unifies microscopic and macroscopic thermal physics modeling methods and multiresolution adaptive macromodeling methods, permitting accurate thermal modeling on length scales ranging from nanometer-scale devices to centimeter-scale packaging and cooling structures. We have used ThermalScope in a large IC design consisting of more than 150 million transistors. The study shows that ThermalScope is suitable for the characterization of thermal and thermal-related effects for billion-transistor nanometer-scale IC designs.

REFERENCES

[1] D. Esseni, M. Mastrapasqua, G. K. Celler, C. Fiegna, L. Selmi, and E. Sangiorgi, "Low field electron and hole mobility of SOI transistors fabricated on ultrathin silicon films for deep submicrometer technology application," *IEEE Trans. Electron Devices*, vol. 48, no. 12, pp. 2842–2850, Dec. 2001.

[2] J. R. Black, "Electromigration failure modes in aluminum metallization for semiconductor devices," *Proc. IEEE*, vol. 57, no. 9, pp. 1587–1594, Sep. 1969.

[3] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. Int. Symp. Low Power Electron. Des.*, 1999, pp. 163–168.

[4] *COMSOL Multiphysics*, COMSOL, Inc. [Online]. Available: http://www.comsol.com/products/multiphysics/

[5] FLOMERICS. [Online]. Available: http://www.flomerics.com/

[6] ANSYS. [Online]. Available: http://www.ansys.com/

[7] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *Proc. Int. Symp. Comput. Archit.*, Jun. 2003, pp. 2–13.

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

[9] Y. Zhan and S. S. Sapatnekar, "A high efficiency full-chip thermal simulation algorithm," in *Proc. Int. Conf. Comput.-Aided Des.*, Oct. 2005, pp. 635–638.

[10] P. Liu, Z. Qi, H. Li, L. Jin, W. Wu, S. X.-D. Tan, and J. Yang, "Fast thermal simulation for architecture level dynamic thermal management," in *Proc. Int. Conf. Comput.-Aided Des.*, Oct. 2005, pp. 639–644.

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

[12] 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. 168–192, Mar. 2003.

[13] Y. Zhan and S. S. Sapatnekar, "Fast computation of the temperature distribution in VLSI chips using the discrete cosine transform and table look-up," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2005, pp. 87–92.

[14] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, "IC thermal simulation and modeling via efficient multigrid-based techniques," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 9, pp. 1763–1766, Sep. 2006.

[15] Z. Yu, D. Yergeau, R. W. Dutton, S. Nakagawa, N. Chang, S. Lin, and W. Xie, "Full chip thermal simulation," in *Proc. Int. Symp. Qual. Electron. Des.*, Mar. 2000, pp. 145–149.

[16] A. Majumdar, *Microscale Energy Transport in Solids*. New York: Taylor & Francis, 1998, ch. 1.

[17] R. Yang, "Nanoscale heat conduction with applications in nanoelectronics and thermoelectrics," Ph.D. dissertation, Dept. Mech. Eng., Massachussetts Inst. Technol., Berkeley, CA, Feb. 2006.

[18] G. Chen, "Nonlocal and nonequilibrium heat conduction in the vicinity of nanoparticles," *Trans. ASME, J. Heat Transf.*, vol. 118, no. 3, pp. 539–545, Aug. 1996.

[19] R. Yang, G. Chen, M. Laroche, and Y. Taur, "Multidimensional transient heat conduction at nanoscale using the ballistic-diffusive equations and the Boltzmann equation," *Trans. ASME, J. Heat Transf.*, vol. 127, pp. 298–306, 2005.

[20] D. G. Cahill, W. K. Ford, K. E. Goodson, G. D. Mahan, A. Majumdar, H. J. Maris, P. Merlin, and R. Simon, "Nanoscale thermal transport," *J. Appl. Phys.*, vol. 93, no. 2, pp. 793–818, Jan. 2003.

[21] S. V. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Boltzmann transport equation-based thermal modeling approaches for hotspots in microelectronics," *Heat Mass Transf.*, vol. 42, no. 6, pp. 478–491, Apr. 2006.

[22] R. Yang, G. Chen, M. Laroche, and Y. Taur, "Simulation of nanoscale multidimensional transient heat conduction problems using ballistic-diffusive equations and phonon Boltzmann equation," *Trans. ASME, J. Heat Transf.*, vol. 127, pp. 298–306, Mar. 2005.

[23] J. Lai and A. Majumdar, "Concurrent thermal and electrical modeling of sub-micrometer silicon devices," *J. Appl. Phys.*, vol. 79, no. 9, pp. 7353–7361, May 1996.

[24] P. G. Sverdrup, Y. S. Ju, and K. E. Goodson, "Sub-continuum simulations of heat conduction in silicon-on-insulator transistors," *Trans. ASME, J. Heat Transf.*, vol. 123, no. 1, pp. 130–137, Feb. 2001.

[25] E. Pop, S. Sinha, and K. Goodson, "Heat generation and transport in nanometer-scale transistors," *Proc. IEEE*, vol. 94, no. 8, pp. 1587–1601, Aug. 2006.

[26] A. McConnel and K. Goodson, "Thermal conduction in silicon micro- and nanostructures," *Annu. Rev. Heat Transf.*, vol. 14, no. 14, pp. 129–168, 2005.

[27] J. Y. Murthy, S. V. J. Narumanchi, J. A. Pascual-Gutierrez, T. Wang, C. Ni, and S. R. Mathur, "Review of multiscale simulation in submicron heat transfer," *Int. J. Multiscale Comput. Eng.*, vol. 3, no. 1, pp. 5–32, 2005.

[28] S. V. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Submicron heat transport model in silicon accounting for phonon dispersion and polarization," *Trans. ASME, J. Heat Transf.*, vol. 126, no. 6, pp. 946–955, Dec. 2004.

[29] S. Kakac, L. Vasiliev, and Y. Bayazitoglu, *Microscale Heat Transfer: Fundamentals And Applications*. Berlin, Germany: Springer-Verlag, 2005.

[30] S. V. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Comparison of different phonon transport models for predicting heat conduction in silicon-on-insulator transistors," *Trans. ASME, J. Heat Transf.*, vol. 127, no. 7, pp. 713–723, Jul. 2005.

[31] J. Murthy and S. Mathur, "An improved computational procedure for sub-micron heat conduction," *Trans. ASME, J. Heat Transf.*, vol. 125, no. 5, pp. 904–910, Oct. 2003.

[32] E. Pop, K. Banerjee, P. Sverdrup, R. Dutton, and K. Goodson, "Localized heating effects and scaling of sub-0.18 micron CMOS devices," in *IEDM Tech. Dig.*, Dec. 2001, pp. 677–680.

[33] Y. Chen, D. Li, J. R. Lukes, and M. Arun, "Monte Carlo simulation of silicon nanowire thermal conductivity," *Trans. ASME, J. Heat Transf.*, vol. 127, no. 10, pp. 1129–1137, Oct. 2005.

[34] D. Lacroix, K. Joulain, D. Terris, and D. Lemonnier, "Monte Carlo simulation of phonon confinement in silicon nanostructures: Application to the determination of the thermal conductivity of silicon nanowires," *Appl. Phys. Lett.*, vol. 89, no. 10, p. 103 104, Sep. 2006.

[35] S. Mazumder and A. Majumdar, "Monte Carlo study of phonon transport in solid thin films including dispersion and polarization," *Trans. ASME, J. Heat Transf.*, vol. 123, no. 4, pp. 749–759, Aug. 2001.

[36] G. Chen, "Ballistic-diffusive equations for transient heat conduction from nano to macroscales," *Trans. ASME, J. Heat Transf.*, vol. 124, no. 2, pp. 320–328, Apr. 2002.

[37] S. Sinha, E. Pop, R. W. Dutton, and E. Goodson, "Non-equilibrium phonon distributions in sub-100 nm silicon transistors," *Trans. ASME, J. Heat Transf.*, vol. 128, no. 7, pp. 638–647, Jul. 2006.

[38] S. Velusamy, W. Huang, J. Lach, M. Stan, and K. Skadron, "Monitoring temperature in FPGA based SoCs," in *Proc. IEEE VLSI Comput. Processsors ICCD*, Oct. 2005, pp. 634–637.

[39] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*. Washington, DC: Hemisphere, 1980.

[40] S. Sinha and K. E. Goodson, "Thermal conduction in sub-100 nm transistors," *Microelectron. J.*, vol. 37, no. 11, pp. 1148–1157, Nov. 2006.

[41] W. J. Minkowycz, E. M. Sparrow, and J. Y. Murthy, *Survey of Numerical Methods*. New York: Wiley, 2006, ch. 1.

[42] M. A. Heaslet and R. F. Warming, "Radiative transport and wall temperature slip in an absorbing planar medium," *Int. J. Heat Mass Transf.*, vol. 8, no. 7, pp. 979–994, 1965.

[43] B. Rutily, L. Chevallier, and J. Pelkowski, "K. Schwarzchild's problem in radiation transfer theory," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 98, pp. 290–307, 2006.

[44] J. L. Henning, "SPEC CPU2000: Measuring CPU performance in the new millennium," *Computer*, vol. 33, no. 7, pp. 28–35, Jul. 2000.

[45] *The Standard Performance Evaluation Corporation (SPEC)*. [Online]. Available: http://www.spec.org/

[46] N. L. Binkert, R. G. Dreslinski, L. R. Hsu, K. T. Lim, A. G. Saidi, and S. K. Reinhardt, "The M5 simulator: Modeling networked systems," *IEEE Micro*, vol. 26, no. 4, pp. 52–60, Jul./Aug. 2006.

[47] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proc. Int. Symp. Comput. Archit.*, Jun. 2000, pp. 83–94.

Zyad Hassan (S'08) received the B.Sc. degree in electronics and electrical communications from Cairo University, Cairo, Egypt, in 2006. He is currently working toward the M.Sc. degree in the Department of Electrical and Computer Engineering, University of Colorado, Boulder.

His research interests include computer-aided design of integrated circuits, emerging nanotechnology devices, and embedded system design.

Nicholas Allec (S'02) received the B.E. degree from Lakehead University, Thunder Bay, ON, Canada, and the M.Sc. degree from Queen's University, Kingston, ON. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON.

His research interests include device and circuit modeling and simulation and thermal modeling.

Li Shang (S'99–M'04) received the B.E. degree (with honors) from Tsinghua University, Beijing, China, and the Ph.D. degree from Princeton University, Princeton, NJ.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Colorado, Boulder. Before that, he was with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. He has published in the areas of design automation for embedded systems, design for nano-technologies, distributed computing, and computer architecture, particularly in thermal/reliability modeling, analysis, and optimization.

Dr. Shang currently serves as an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS and serves on the technical program committees of several embedded systems and design automation conferences. He received the Best Paper Award nomination at the International Conference on Computer-Aided Design 2008, Design Automation Conference (DAC) 2007, and the Asia and South Pacific DAC 2006. His work on temperature-aware on-chip networks has been selected for publication in the MICRO Top Picks 2006. He is the recipient of the Best Paper Award at the Parallel and Distributed Computing and Systems 2002 and his department's Best Teaching Award in 2006. He is the Walter Light Scholar.

Robert P. Dick (S'95–M'02) received the B.S. degree from Clarkson University, Potsdam, NY, and the Ph.D. degree from Princeton University, Princeton, NJ.

He was a Visiting Professor at the Department of Electronic Engineering, Tsinghua University, Beijing, China, a Visiting Researcher at NEC Laboratories America, and an Associate Professor with Northwestern University, Evanston, IL. He is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. He has published in the areas of embedded operating systems, data compression, embedded system synthesis, dynamic power management, low-power and temperature-aware integrated-circuit design, wireless sensor networks, human-perception-aware computer design, reliability, embedded system security, and behavioral synthesis.

Dr. Dick is an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS and serves on the technical program committees of several embedded systems and computer-aided design/VLSI conferences. He is the recipient of a National Science Foundation CAREER award and his department's Best Teacher of the Year Award in 2004. His technology won a Computerworld Horizon Award, and his paper was selected by the Design Automation and Test in Europe as one of the 30 most influential in the past ten years in 2007.

Vishak Venkatraman (S'02–M'07) received the B.E. degree in electronics and communication engineering from the University of Madras, Chennai, India, in 2000, the M.S. degree in electrical and computer engineering from the University of Hartford, West Hartford, CT, in 2002, and the Ph.D. degree in electrical engineering from the University of Massachusetts, Amherst, in 2006.

He is currently with Advanced Micro Devices, Sunnyvale, CA. His research interests include full-chip- and transistor-level thermal modeling, simulation, and analysis, reliability analysis, and interconnect signaling for low-power high-performance microprocessors.

Ronggui Yang (M'01) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1996, an M.S. degree from the University of California, Los Angeles, in 2001, and the Ph.D. degree from the Massachusetts Institute of Technology (MIT), Cambridge, in 2006.

He is currently an Assistant Professor with the Department of Mechanical Engineering, and the Sanders Faculty Fellow, University of Colorado, Boulder. His current research interests are on nanoscale and ultrafast thermal sciences and their applications in energy and information technologies.

Dr. Yang's innovative research has won him numerous national and international awards including the 2009 National Science Foundation CAREER Award, the 2008 MIT Technology Review's TR35 Award, the 2008 Defense Advanced Research Projects Agency Young Faculty Award, the 2005 Goldsmid Award from the International Thermoelectrics Society, and a number of Best Paper Awards and nominations from the American Society of Mechanical Engineers and IEEE.