ThermalScope: Multi-Scale Thermal Analysis For Nanometer-Scale Integrated Circuits

Nicholas Allec*, Zyad Hassan†, Li Shang†, Robert P. Dick‡, and Ronggui Yang§

* ECE Department
Queen's University
Kingston, ON K7L 3N6, Canada
nicholas.allec@ece.queensu.ca

† ECE Department
University of Colorado at Boulder
Boulder, CO 80309, U.S.A
{zyad.hassan, li.shang}@colorado.edu

‡ EECS Department
Northwestern University
Evanston, IL 60208, U.S.A
dickrp@northwestern.edu

§ ME Department
University of Colorado at Boulder
Boulder, CO 80309, U.S.A
ronggui.yang@colorado.edu

***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 nanoscale fabrication processes is increasing the importance of quantum thermal phenomena such as ballistic phonon transport. Accurate thermal analysis of nanoscale ICs containing hundreds of millions of devices requires characterization of thermal effects on length scales that vary by several orders of magnitude, from nanoscale quantum thermal effects to centimeter-scale cooling package impact. Existing chip–package thermal analysis methods based on classical Fourier heat transfer cannot capture nanoscale quantum 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 work, we propose and develop ThermalScope, a multi-scale thermal analysis method for nanoscale IC design. It unifies microscopic and macroscopic thermal physics modeling methods, i.e., the Fourier and Boltzmann transport modeling methods. Moreover, it supports adaptive multi-resolution modeling. Together, these ideas enable efficient and accurate characterization of nanoscale quantum heat transport as well as chip–package level heat flow. ThermalScope is designed for full-chip thermal analysis of billion-transistor nanoscale IC designs, with accuracy at the scale of individual devices. ThermalScope enables accurate characterization of temperature-related effects, such as variation in leakage power and delay. ThermalScope has been implemented in software and used for full-chip thermal analysis and temperature-dependent leakage analysis of an IC design with more than 150 million transistors. It will be publicly released for free academic and personal use.

I. Introduction

Process scaling and increasing device density increase power density and thermal effects. Increased integrated circuit (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 sub-threshold current [3]), and cooling cost. IC thermal analysis is thus critical because it is possible to improve performance, reliability, and power consumption via run-time thermal management techniques as well as by considering thermal issues during the design process.

CMOS technology is fast approaching the nanometer-scale regime. 45 nm CMOS fabrication technology is entering mainstream use. In the coming five years and beyond, ultra-thin body device structures, such as multi-gate MOSFET (FinFET) and silicon-on-insulator (SOI), will be used for mainstream ICs. Quantum thermal effects will become prominent in nanoscale devices. When the mean free path of phonons (lattice vibrations) approaches the device feature scale, ballistic phonon transport serves as the main mechanism of heat transfer. Heat transport within nanoscale devices is strongly affected by interface scattering and reflection effects. IC thermal analysis thus requires accurate modeling of heat transport across multiple scales, from nanoscale on-chip devices, through millimeter-scale silicon chip and centimeter-scale cooling package, to the ambient environment.

This work was supported in part by the SRC under awards 2007-HJ-1593 and 2007-TJ-1589, in part by the NSF under awards CCF-0702761 and CNS-0347941, and in part by the NSERC fellowship program.

Conventional chip–package thermal analysis techniques have been so slow that evaluating numerous design alternatives was prohibitively expensive during IC design [4], [5], [6]. As a result, most thermal optimization was done during packaging and cooling solution design. Unfortunately, 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], [8], [9], [10], [11], [12], [13], [14], [15]. Using these methods, heat transfer through chip and cooling package is modeled using the classical Fourier transport model. IC chip and cooling packages are virtually partitioned into discrete three-dimensional thermal elements. Compact heat transfer equations are then derived and solved using numerical methods to characterize the thermal profile of IC chip and cooling package. Although some of these techniques are fast enough for use during IC design and within run-time thermal management techniques, they are all based on the Fourier heat flow model. This model cannot capture phonon quantum thermal effects and yields inaccurate results when used at length scales on the order of phonon mean free path [16]. These observations are supported by the data presented in Section IV-A.

Techniques with different fidelities and efficiencies have been developed to model nanoscale device-level phonon heat transport, including molecular dynamics methods, Boltzmann transport equation (BTE), and ballistic-diffusion model. Computational complexity has been the primary challenge of adopting nanoscale heat transfer methods for large-scale IC chip–package thermal analysis. Molecular dynamics methods model heat transfer by directly simulating interatomic interactions [17]. Approaches implemented using this method are highly accurate. However, they are extremely computationally expensive. The BTE method and its variants model heat transfer by simulating the transport of phonons [18]. It can accurately approximate ballistic phonon transport. BTE methods are much more efficient than molecular dynamics methods. However, their computational complexity remains prohibitive, and their use has been restricted to device-level analysis. The ballistic-diffusion based thermal analysis method is an approximation of the Boltzmann transport method [19]. Although the ballistic-diffusion model is the most efficient of these, it is still much more computationally-demanding than the Fourier model. In addition, results from the ballistic-diffusion model tend to have low fidelity [19].

In summary, there is a gap between the efficiency and accuracy of nanoscale and chip-package thermal analysis techniques that must be closed if high-quality temperature-aware design techniques and run-time thermal management algorithms are to be developed for ICs composed of nanoscale devices. Our goal is to close this gap. We propose and develop a multi-scale solution, named ThermalScope, for unified device–chip–package thermal analysis targeting billion-transistor nanoscale ICs. ThermalScope is a multi-scale solver that integrates *microscopic and macroscopic thermal physics modeling methods* (enabling characterization of nanoscale quantum heat transport as well as chip–package level heat flow), *detailed and compact numerical analysis techniques* (allowing the usage of computationally-intensive non-classical device-level modeling within full-chip thermal characterization), and *multi-resolution adaptive*

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

modeling granularities (permitting modeling on length scales ranging from nanoscale devices to centimeter-scale packaging and cooling structures). The proposed solution overcomes the limitations of existing chip–package level and device-level thermal analysis methods. It provides a unified modeling infrastructure for IC heat flow analysis from nanoscale devices to billion-device IC chips. We have applied ThermalScope to an IC design containing over 150 million transistors.

The rest of this article is organized as follows. Section II introduces the nanoscale IC thermal analysis problem and highlights the challenges of efficient and accurate thermal analysis of nanoscale ICs. Section III describes the proposed multi-scale thermal analysis method. Section IV evaluates and demonstrates the use of ThermalScope. We conclude in Section V.

II. Challenges

This section gives an overview of the IC thermal analysis problem, and discusses the challenges for accurate thermal analysis of nanoscale ICs.

IC thermal analysis is the process of characterizing the three-dimensional thermal profile of an IC chip and cooling package. An IC thermal profile is a complex function of its design, fabrication technology, cooling and package configuration, power consumption, and ambient environment. The thermal profile of a nanoscale IC depends on power consumption variation at multiple scales. Hotspots in the active layer are often caused by high power density functional units, e.g., a floating point unit. Inside a transistor, a hotspot often occurs near the drain terminal region, mainly due to the accumulation of slow-moving (optical) phonons (which are quanta of vibrational energy, i.e., heat particles). IC thermal analysis thus requires accurate modeling of heat transport across multiple scales, from nanoscale on-chip devices through millimeter-scale silicon chip and centimeter-scale cooling package to the ambient environment.

The Fourier heat diffusion model has been widely used in recently-developed IC chip–package thermal analysis packages [7], [8], [9], [10], [11], [12], [13], [14], [15]. In the classical Fourier model, the temperature distribution is governed by the Fourier heat conduction equation. This model incorrectly implies that thermal effects will not worsen as device dimensions are scaled down if power dissipation per unit length remains constant, as prescribed in the ITRS roadmap [20]. However, the conventional diffusive treatment of heat transfer is no longer valid at length scales less than the phonon mean free path in silicon, i.e., 200–300 nm. This is analogous to the failure of the drift and diffusion model for describing electron transport in nanoscale MOSFETs, for which the critical dimension is only a few electron mean free paths. Ballistic phonon transport implies reduced effective thermal conductivity in proportion to the ratio of the hotspot size to the phonon mean free path. It is expected that heat conduction in 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 IV-A. In addition, the microelectronics industry is fast approaching the scaling limits of bulk CMOS. 32 nm or 22 nm features appear to be the end of the road. As a result, there is extensive research on thin-film SOI, FinFET, and other novel device structures. Many of these unconventional structures will introduce new thermal problems that did not exist for bulk silicon. In the case of thin-film SOI and FinFET, the presence of boundary scattering at the many material interfaces and the thick insulating films can raise thermal resistance substantially. Figures 1 and 3 show the top view of thermal profiles of 65 nm bulk silicon and FinFET devices simulated using ThermalScope, the proposed multi-scale thermal analysis solution. Figures 2 and 4 show the corresponding results for a Fourier-only based solver. These figures demonstrate that the Fourier-only solver results not only deviate in the peak temperature reported by ThermalScope, but also deviate in the thermal profile itself. For FinFETs, the difference between the results reported by the two methods is more significant, mainly due to boundary scattering at the interface with the oxide layer surrounding the device, which cannot be captured by the Fourier heat flow model.

In summary, thermal analysis for nanoscale ICs raises the following challenges:

1) The major challenges of numerical thermal analysis of nanoscale devices and ICs are high computational complexity and memory usage. Accurate thermal analysis requires the use of detailed numerical analysis methods with fine-grain models. For nanoscale ICs, from nanometer-scale transistors to centimeter-scale cooling package, the modeling granularities 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 overheads.

2) Accurate thermal analysis requires unified heat transport modeling from nanoscale devices to the chip–package level. However, chip–package level thermal analysis and device-level thermal analysis 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 quantum thermal effects. Device-level modeling techniques, such as molecular dynamics and BTE, model nanoscale quantum thermal effects. However, their use has been limited to individual devices due to their high computational complexities.

III. ThermalScope: A multi-scale thermal analysis infrastructure

In this section, we present ThermalScope, the proposed multi-scale thermal analysis solution for nanoscale ICs.

III.A. Overview

Figure 5 illustrates the flow of ThermalScope. ThermalScope is a multi-scale solution that integrates microscopic and macroscopic thermal physics modeling methods, as well as multi-resolution macromodeling techniques. In contrast with existing Fourier-based chip–package thermal analysis methods, ThermalScope uses accurate BTE analysis to capture quantum thermal effects that are common at nanometer length scales. BTE analysis can be extremely time-consuming, even for a single device, which is why a hybrid Fourier/BTE method using adaptive spatial discretization is used for thermal analysis at the device level. This accelerates thermal analysis by orders of magnitude compared to BTE, while maintaining accuracy. However, it is still too slow for full–chip thermal analysis. A multi-scale macromodeling method was developed that enables fast full-chip thermal analysis with accuracy even at the scales of individual devices. The macromodel contains thermal impact coefficients that efficiently characterize thermal interactions among thermal modeling elements from the chip–package to the inter-device level. Hierarchical multi-scale partitioning and clustering techniques are used to partition the IC into modeling elements. In addition, the macromodel contains a look-up table, constructed using hybrid Fourier/BTE analysis, for accurate and efficient device-level thermal modeling. In summary, ThermalScope provides a unified modeling infrastructure for IC heat flow analysis from nanoscale devices to billion-device IC chips.

The rest of this section details the techniques and algorithms of the proposed thermal analysis infrastructure. Section III-B first describes the microscopic and macroscopic thermal physics modeling methods developed in ThermalScope. Next, Section III-C describes the hybrid Fourier/BTE analysis method. Finally, Section III-D describes the design of the multi-scale macromodeling method.

III.B. Modeling

ThermalScope uses both Fourier and BTE modeling methods to characterize the thermal effects from nanometer-scale devices to centimeter-scale chip and package. This section describes thermal physics models and explains their use in ThermalScope.

III.B.1) Fourier Model: The steady–state classical Fourier model is characterized by the following equation [18]:

$$\nabla \cdot (K \nabla T) + q_{vol} = 0 \quad (1)$$

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

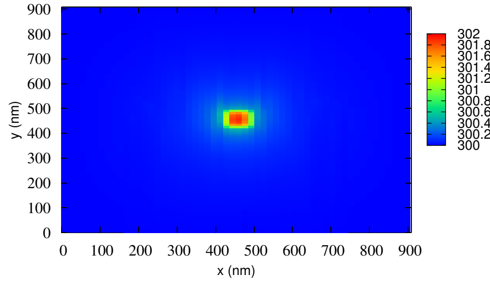

Fig. 1. Bulk silicon device simulated using ThermalScope.

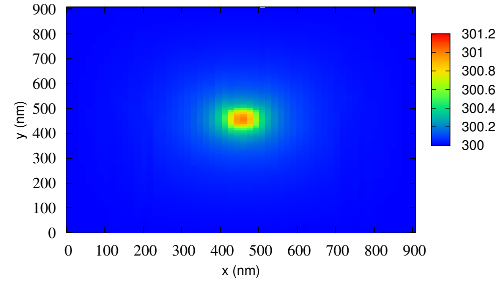

Fig. 2. Bulk silicon device simulated using the Fourier model.

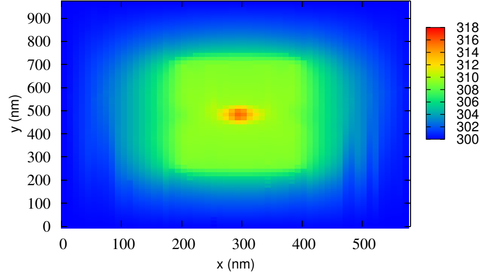

Fig. 3. FinFET device simulated using ThermalScope.

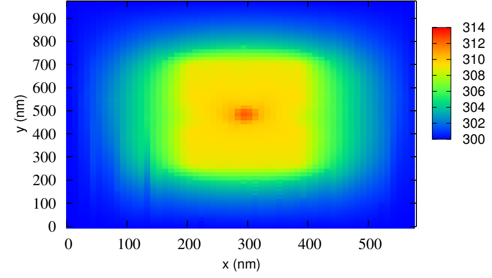

Fig. 4. FinFET device simulated using the Fourier model.

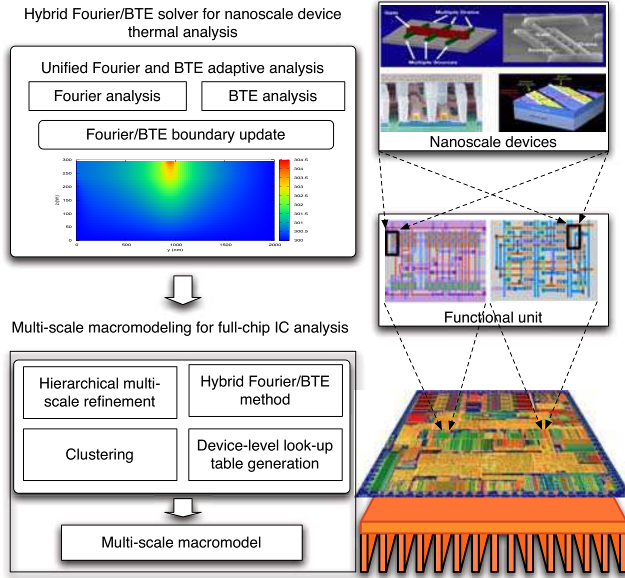

Fig. 5. ThermalScope multi-scale thermal analysis infrastructure.

where K is thermal conductivity, T is temperature, and q_{vol} is a volumetric heat source.

In contrast with the BTE method, the Fourier model is capable of efficiently (and accurately) modeling the thermal effects at feature length scales much longer than the mean free path of phonons. 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 used only at the device level.

III.B.2) BTE Model: As described in Section I, models that can capture the nanometer-scale quantum thermal effects include Molecular dynamics, the BTE model, and the Ballistic Diffusive model. Molecular dynamics models are not suitable for multi-scale thermal analysis because of their extremely high computational complexity, while the Ballistic Diffusive model provides insufficient accuracy. ThermalScope uses the Gray phonon BTE under the relaxation time approximation to model the device regions. The Gray BTE model employs a phonon distribution function, e'', and neglects the wave-like behavior of phonons. It also assumes a single group velocity and relaxation time for phonons, which are independent of their frequency and polarization. The use of the relaxation time approximation allows the 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 follows [18]:

$$\nabla \cdot (\vec{s} v_g e'') = \frac{e^0 - e''}{\tau_{\mathit{eff}}} + q_{vol} \tag{2}$$

where $\vec{s}$ is the normalized 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 follows [18]:

$$e^0 = \frac{1}{4\pi} \int_{4\pi} e'' d\Omega = \frac{1}{4\pi} C(T_L - T_{ref}) \tag{3}$$

where Ω is the angular discretization, C is the specific heat, T_L is the lattice temperature, and T_{ref} is the reference temperature at the specific heat. The lattice temperature, T_L, can be calculated using Equation 3 once the equilibrium energy density is known.

The relaxation time, τ_{eff}, can be found using the bulk material equation:

$$k = \frac{1}{3} C v_g^2 \tau_{\mathit{eff}} \tag{4}$$

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 Equation 2. Its value can be derived from device power consumption, which can be obtained using circuit simulation. To obtain the thermal profile of a device, e'' is determined by solving Equation 2, using the power profile of the device (represented by q_{vol}). The equilibrium energy density, and thus thermal profile, can then be obtained using Equation 3.

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

III.C. The Hybrid Fourier/BTE analysis method

The main drawback of the BTE model is its high computational complexity. To overcome this problem, we propose a hybrid Fourier/BTE method that combines the best of both models. Compared to the BTE method, the Fourier method is orders of magnitude faster, but less accurate. However, its accuracy is sufficient if used to model IC regions farther than the mean free path of phonons from heat sources, providing significant simulation time savings. The flow of the hybrid approach is described next. Its accuracy and efficiency are evaluated in Section IV-A.

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 $\eta \times v_g \tau_{eff}$, surrounding the devices is chosen as the BTE region, where η is a constant. $v_g \tau_{eff}$ is the mean free path of a phonon, where v_g is the phonon group velocity and τ_{eff} is the effective relaxation time. Varying the constant η changes the number of elements in the BTE region, and thus the physical area modeled using the BTE solver. The effect of changing that constant is evaluated in Section IV-A. The rest of the structure outside of this region is evaluated with the Fourier solver.

In the hybrid solver, the Fourier solver and the BTE solver are invoked iteratively. The boundary temperatures of the BTE/Fourier 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.

To solve Equations 1 and 2 for the BTE and Fourier models, we use the tri-diagonal matrix algorithm (TDMA) method. TDMA is a frequently-used numerical analysis technique in structured meshes [21]. It exploits the tri-diagonal form of the coefficients matrix, allowing the system to be solved in $O(N)$ operations. The TDMA algorithm improves storage efficiency by only storing the non-zero elements of the coefficient matrices. In order to obtain a tri-diagonal matrix form, the line-by-line TDMA (LBL-TDMA) method is used. As its name indicates, this method transforms the system along each dimension, thereby transforming from a three-dimensional to a one-dimensional system so that the coefficients of each element are non-zero only for neighboring elements [21].

The discretized equations have the form:

$$a_p e''_p = \sum_{nb} a_{nb} e''_{nb} + b_p \tag{5}$$

where p refers to the self element and nb refers to six nearest neighboring elements, e, w, n, s, t, b. The pseudo code for solving a line along the z-direction using LBL-TDMA is shown in Algorithm 1. The algorithm goes over all the elements in the system and forms a one-dimensional system (whose coefficients are denoted by the $1D$ subscript) from the three-dimensional system. When solving for elements along the z-direction, only the energy densities (or temperatures in the case of the Fourier equation) along this line are assumed to be unknowns, while guessed values (obtained from the previous iteration) are used for energy densities (or temperatures) of neighboring elements along the other two directions. After forming the one-dimensional system, the TDMA algorithm is called (lines 13 to 24).

III.D. Multi-scale macromodeling method

This section describes the proposed multi-scale macromodeling method. As shown in Figure 5, the macromodeling method is built using hierarchical multi-scale partitioning and clustering techniques, and device-level compact hybrid Fourier/BTE method. Our goal is to build an efficient compact macromodeling method for full-chip IC thermal analysis with accuracy at the scale of an individual device.

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

$$T_i = f_i(P_1, \ldots, P_N) = r_{i,1} \times P_1 + \cdots + r_{i,N} \times P_N \tag{6}$$

Algorithm 1 LBL-TDMA algorithm.

1: **for** each element i in x-direction **do**
2: **for** each element j in y-direction **do**
3: **for** each element k in z-direction **do**
4: Get the system of equations for this line:
5: $a_{p1D}(k) = a_p(i,j,k)$
6: $a_{t1D}(k) = -a_t(i,j,k)$
7: $a_{b1D}(k) = -a_b(i,j,k)$
8: $b_{1D}(k) = b(i,j,k)$
9: Use guessed values for neighboring elements and add to the b term:
10: $b_{1D}(k) = b_{1D}(k) + a_e * e''(i+1,j,k) +$
11: $a_w * e''(i-1,j,k) + a_n * e''(i,j+1,k) + a_s * e''(i,j-1,k)$
12: Add the effect of the two boundary elements and the heat source term to the b term
13: **end for**
14: Execute the TDMA algorithm:
15: **for** elements k along the diagonal, starting from second row **do**
16: $r = a_{b1D}(k)/a_{p1D}(k)$
17: $a_p(k) = a_p(k) - r * a_t(k-1)$
18: $b_{1D}(k) = b_{1D}(k) - r * b_{1D}(k-1)$
19: **end for**
20: Back Substitution:
21: $e''(i,j,last\ element) = b_{1D}(last\ element)/a_p(last\ element)$
22: **for** elements k along the diagonal starting from the one before last **do**
23: $e''(i,j,k) = (b_{1D}(k) - a_t(k) * e''(i,j,k+1))/a_p(k)$
24: **end for**
25: End TDMA algorithm
26: **end for**
27: **end for**

where T_i is the temperature of device i, P_j is the power consumption 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 of 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 temperature and power vectors of on-chip devices. $\mathbf{R}$ is called the thermal impact coefficient matrix, which can be obtained by calculating the inverse of the thermal conductance matrix, $\mathbf{K}$ (see Equation 1). 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 sub-matrix of $\mathbf{K}^{-1}$, containing the coefficients corresponding to the device elements. To accurately model the device-level quantum thermal effect, 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 centimeter-scale IC design containing hundreds of millions of nanoscale devices, matrix $\mathbf{K}$ will contain a massive number of elements, i.e., M is an extremely large number. Computing matrix $\mathbf{K}^{-1}$ is a daunting task.

We propose the following two techniques to tackle this problem: hierarchical multi-scale spatial partitioning and thermal impact clustering.

We first describe *hierarchical multi-scale spatial partitioning*. Since on-chip inter-device distances vary widely, for any device of interest, it is important to differentiate between local devices (those in close proximity to it), and remote devices (those far away from it). Devices located in a small neighboring region will have widely varying thermal impact coefficients on the device of interest. However, distant devices will have similar thermal impacts on the device of interest. We use this observation to simplify Equation 6 by using adaptive spatial partitioning. Remote devices can be characterized using a single dependency, and thus a coarse-grained partition can be used to identify these dependencies. Since the local thermal impacts cannot be described by a small number of dependencies, fine-grained partitioning is required in these regions. The temperature equation for device i follows.

$$T_i = \zeta_{i,1} \times \xi_1 + \cdots + \zeta_{i,L} \times \xi_L \tag{7}$$

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

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

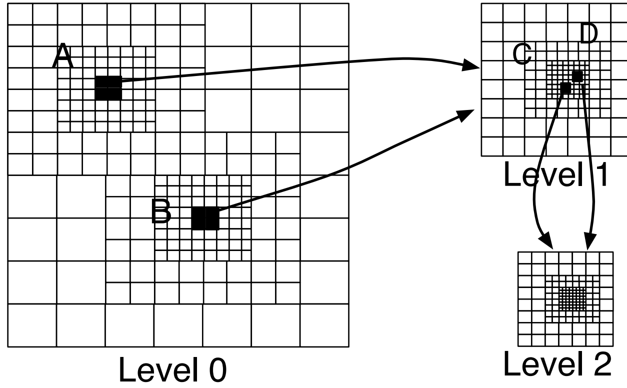

Fig. 6. The hierarchical multi-resolution partitioning.

device i's temperature. As shown in Figure 6, since coarse-grained partitioning can be used in most locations, i.e., $L \ll N$, Equation 6 is greatly simplified. Therefore, adaptive spatial partitioning can reduce the modeling complexity. However, as the device-level quantum thermal effect becomes increasingly significant, nanoscale fine-grain modeling in the device neighborhood is required in order to accurately estimate the temperature of each device. Therefore, L in Equation 7 is still a large number. Given N devices, the total memory usage is proportional to $N \times L$.

The idea of hierarchical spatial partitioning is to further reduce the memory usage by sharing common partitions hierarchically when computing temperatures of different devices. This concept is depicted in Figure 6, which shows a three-level hierarchical adaptive partitioning. At the chip–package level (Level 0), if we are interested in the temperatures of two devices i and j located at the centers of regions **A** and **B** respectively, we can use coarse-grained partitioning for regions far from **A** and **B**. **A** and **B** will share similar local dependencies, and thus can be characterized by the same fine-grained partitioning (Level 1). In addition, heterogeneous partitioning is used for the surrounding areas/devices of interest. Assuming 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 device i and j are simplified as follows.

$$\begin{aligned} T_i &= f_{i,\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{\mathrm{Level}_2}(\dots) \\ T_j &= f_{j,\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{\mathrm{Level}_2}(\dots) \end{aligned} \quad (8)$$

i.e., these two temperature equations share the same level 1 and level 2 expressions. On the other hand, if we are interested in the temperatures of two devices i and j both in regions **C**, the temperature equations for device i and j are simplified as follows.

$$\begin{aligned} T_i &= f_{\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{i,\mathrm{Level}_2}(\dots) \\ T_j &= f_{\mathrm{Level}_0}(\dots) + f_{\mathrm{Level}_1}(\dots) + f_{j,\mathrm{Level}_2}(\dots) \end{aligned} \quad (9)$$

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

The second technique we propose, *thermal impact clustering*, clusters redundant dependencies. 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, we have devised a clustering scheme that reduces the amount of redundant information by grouping equivalent thermal impacts into a single representative thermal impact.

A hierarchical clustering technique is used for each row of the thermal impact coefficient matrix. Each row corresponds to the thermal impact coefficients of all elements for a single element. The clustering algorithm works as follows. The thermal impact coefficient matrix is subdivided into single row vectors. Different 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 equal to 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.

In order to obtain accurate thermal profiles at the finest granularity level, we must consider nanoscale thermal effects. Using the hybrid Fourier/BTE solver, a compact model in look-up table form is derived to model the device-level quantum thermal effect. This compact model contains device-level temperature information for different device geometries, power consumptions, and technologies. By combining the device-level look-up table with the clustered thermal impact coefficients at different levels of granularity, from inter-device level to chip–package level, to form the multi-scale macromodel, the macromodel can be used for accurate and efficient thermal analysis from the device level to chip–package level.

IV. Results

In this section we evaluate ThermalScope, the proposed multi-scale thermal analysis method. ThermalScope unifies Fourier and BTE modeling techniques as well as a multi-scale macromodeling method. In Section IV-A and Section IV-B, we first evaluate the proposed hybrid Fourier/BTE analysis method. In ThermalScope, hybrid Fourier/BTE analysis is responsible for characterization of device-scale quantum heat transport and functional unit length scale thermal effects. Using this hybrid method, intra-device thermal effects are characterized using BTE analysis, and inter-device thermal effects are characterized using Fourier analysis. In Section IV-A, we evaluate this hybrid analysis method using device-level thermal analysis. Next, in Section IV-B, we report results that indicate that inter-device thermal interaction can be accurately modeled using Fourier thermal analysis. In Section IV-C, we evaluate the multi-scale macromodeling method. ThermalScope is developed to target billion-transistor nanoscale 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.

IV.A. Device-level thermal modeling using hybrid Fourier/BTE analysis

In this section, we show that the BTE method is necessary for accurate computation of device-level thermal profiles. We then evaluate the accuracy and speedup of the hybrid Fourier/BTE method.
Accuracy of the BTE method: ThermalScope is capable of using a BTE solver and a hybrid BTE/Fourier solver for device-level thermal analysis. To evaluate the accuracy of our BTE solver for length scales below the mean free path of phonons, we have modeled the Heaslet and Warming problem [22]. In this problem, a block of material has two opposing walls held at different temperatures, while the other walls are insulating. The distance between the two fixed-temperature walls is varied, and the resulting thermal gradients are observed. Our results were in excellent agreement with those of Rutily et al. [23].
The BTE method vs. Fourier analysis: Here we show the inaccuracy of the Fourier model in capturing device-level thermal effects by comparing it to the BTE model. We simulate a 910 nm× 910 nm×500 nm region containing a bulk silicon device for a range of different process technologies (65 nm, 45 nm, and 32 nm). The simulation reports that, compared to the BTE method, the Fourier method introduces 34.0%, 44.8%, and 54.1% error for 65 nm, 45 nm, 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. If used alone, the Fourier method is unable to model the quantum thermal effects of nanoscale structures.

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

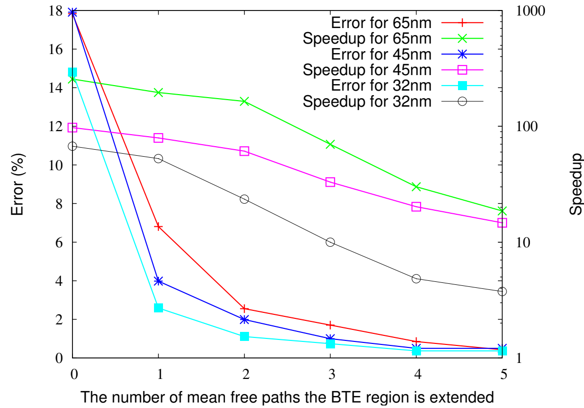

Fig. 7. Accuracy and efficiency of the hybrid solver.

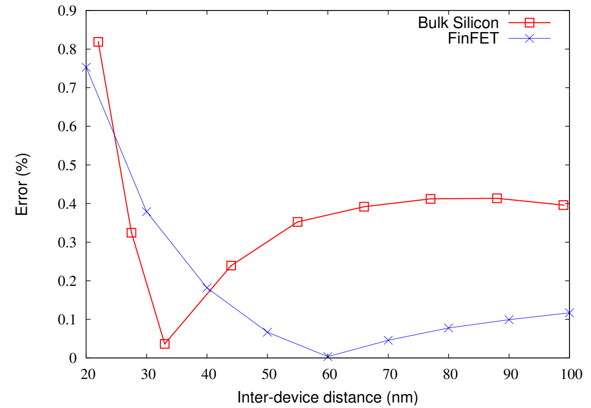

Fig. 8. BTE inter-device correlation for bulk silicon/FinFET devices.

The hybrid method vs. 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 above. This material is partitioned into 343,128 thermal elements. We first apply BTE analysis to the whole material. The overall simulation time was 16.3 hours. Next, we use the hybrid approach. We vary the number of elements solved using the Fourier method by changing, η (defined in Section III-C), the number of mean free paths the BTE region is extended away from the heat source. We report the relative temperature differences and the speedups compared to the BTE-only based method. The test setup is repeated for 45 nm and 32 nm technologies. Figure 7 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× with an error of 4%, and a 10× to 70× 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 and 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 these material layers. Therefore, the hybrid method can greatly accelerate the simulation process.

IV.B. Inter-device thermal effect modeling using Fourier analysis

The goal of this analysis is to demonstrate that the Fourier method is sufficient to accurately model the thermal interaction between neighboring devices. This allows 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. At the device level, only the device of interest need be characterized using the BTE model. In addition, as described in Section III, the hybrid Fourier/BTE method is too expensive for use in full-chip thermal analysis. A multi-scale macromodeling method is therefore developed in ThermalScope using the hybrid Fourier/BTE method. BTE analysis is only needed for intra-device thermal analysis. This greatly simplifies the development process of the full-chip macro-modeling method.

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

Figure 8 shows the analysis error of Fourier-based inter-device thermal correlation analysis as a function of inter-device distance for both bulk silicon and FinFET devices. The analysis error is estimated using $\frac{T_{BTE}-T_{Fourier}}{T_{BTE}-T_a}$, 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. The results show that the estimation error decreases as the inter-device distance increases and quantum thermal effects become less significant. Compared to the BTE method, the Fourier method can accurately estimate inter-device thermal effects with less than 1% error even when the inter-device distance is as low as 20 nm, which suggests that the Fourier method can provide sufficient accuracy for inter-device thermal correlation analysis, and that only individual devices of interest must have their thermal profiles computed using the BTE model.

IV.C. Multi-scale simulation

ThermalScope is designed for thermal analysis of billion-transistor nanoscale 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 mm×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 mm×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.

IV.C.1) Adaptive spatial granularity: First we will evaluate the potential simulation time and memory storage savings of the proposed technique. To maintain accuracy during device-level thermal analysis, we require modeling elements to be much smaller than the heat source. Assuming the heat source is the size of the device and the process technology is 65 nm, then we require the element size to be a few nanometers along each dimension. At the other side of the spectrum, the sizes of the chip and cooling package are in the range of centimeters. Using the industry design used in this analysis, if one were to use homogeneous partitioning, the memory requirements would be approximately on the order of 10^{18} bytes. The computations required to evaluate the temperature of a single device would be 10^{12} additions and 10^{12} multiplications. A modern IC may contain hundreds of millions of transistors for which the temperature must be evaluated to allow accurate thermal-dependent leakage and timing analysis. From 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 computation time. Hierarchical adaptive modeling

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

granularities are used from the chip level down to the inter-device level. This adaptive modeling reduces the problem size to requiring storage on the order of 10^8 bytes for the thermal impact coefficient matrices for the same problem as described above. For comparison, the input power profile of the industry design itself requires more than 7×10^8 bytes of storage. The 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 clustering. The simulation run-time and memory usage results for the device-level temperature evaluation (after obtaining the coefficient matrices) for clustering are shown in Table I. The chip evaluated contained over 150 million devices and was evaluated for both a bulk silicon design and a FinFET design. The results show that although the memory usage may not be significantly reduced using a clustering technique, significant speed-up 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.

TABLE I
EFFICIENCY EVALUATION.

	Bulk silicon		FinFET	
	Clustering	No clustering	Clustering	No clustering
t_{CPU} (min)	167	485	179	607
Memory (MB)	548	604	620	604

IV.C.2) 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 using the large industry design.

Since the leakage power of the chip is strongly affected by temperature, it is necessary to include the leakage power estimation in the thermal analysis simulation flow. To determine thermal profile 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. The results of this simulation are then used to update the leakage power of the chip. This iterative process continues until convergence is reached for the simulated thermal and the power profiles. In this study, the temperature-leakage power dependency is obtained using curve fitting of a measured 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 multi-scale macromodeling method using the described iterative analysis process. During thermal analysis, the temperature of every 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 Figures 9 through 11 for bulk silicon technology, and Figures 12 through 14 for FinFET technology. Figures 10 and 13 are enlarged fine-grained thermal profiles of a hotspot on the chip. Figures 11 and 14 illustrate a further enlargement of the area, showing the device-level information for two devices out of the hundreds of millions for which temperature is calculated. Although ThermalScope evaluates the temperature of every device, it is also capable of coarse-grained thermal analysis. The thermal profiles demonstrate the capability of ThermalScope to evaluate the thermal profile at different scales, which vary through six orders of magnitude. The profiles also indicate the inaccuracy resulting from using coarse-grained estimates of device temperatures.

Figures 9 through 14 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. For the bulk silicon designs, at the intermediate level (255μm×255μm) 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 device temperatures. The strength of ThermalScope is that it reports the temperature of each device allowing detailed full-chip thermal analysis. In coarse-grained thermal analysis, the features occurring at device level are not considered, which leads to inaccurate estimation of device temperatures, as shown in Figures 9–14.

In addition to thermal analysis, ThermalScope can also be used to estimate the leakage power. The leakage power is determined by the same iterative process described earlier. We have compared 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 at an 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 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 II.

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 demonstrates 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 iterative full-chip thermal analysis with device-level modeling granularity technique is shown in Figure 15.

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 this information given reasonable time and storage.

TABLE II
LEAKAGE POWER ESTIMATION.

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

V. CONCLUSIONS

Thermal analysis and optimization are now critical in nanoscale IC design. The goal of this work is 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 multi-scale thermal analysis solution. It unifies microscopic and macroscopic thermal physics modeling methods and multi-resolution adaptive macromodeling methods, permitting accurate thermal modeling on length scales ranging from nanoscale 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 thermal analysis and characterization of thermal-related effects for billion-transistor nanoscale IC designs.

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.

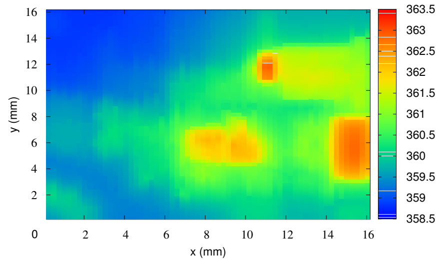

Fig. 9. Bulk full-chip.

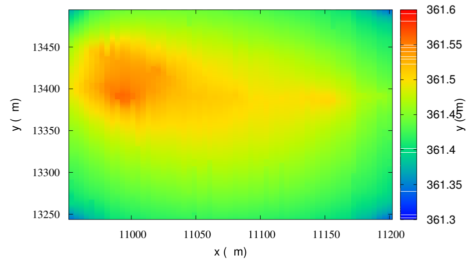

Fig. 10. Bulk 255μm×255μm.

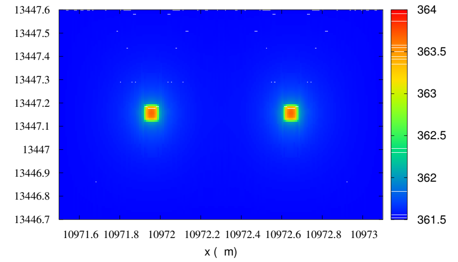

Fig. 11. Bulk 1.6μm×1.6μm.

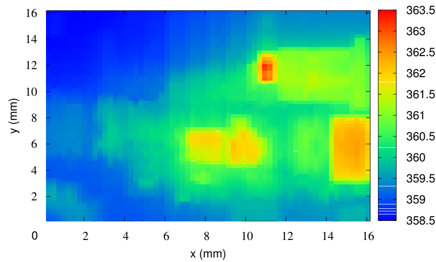

Fig. 12. FinFET full-chip.

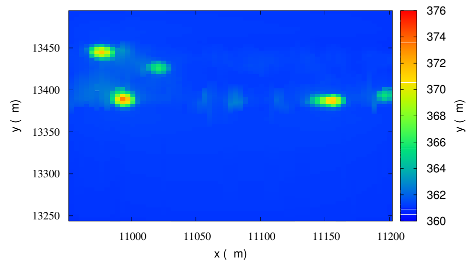

Fig. 13. FinFET 255μm×255μm.

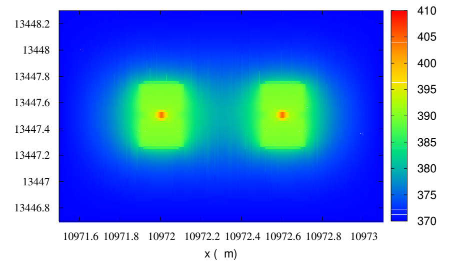

Fig. 14. FinFET 1.6μm×1.6μm.

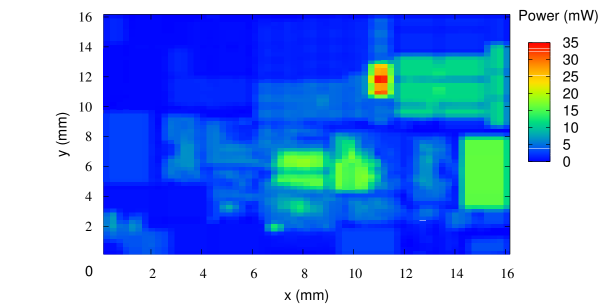

Fig. 15. Leakage power profile of the industry design.

The study also indicates that incorporating device–level quantum thermal effects is especially important for future technologies, such as FinFETs.

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, pp. 2842–2850, Dec. 2001.

[2] J. R. Black, "Electromigration failure modes in aluminum metallization for semiconductor devices," *Proc. IEEE*, vol. 57, pp. 1587–1594, Sept. 1969.

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

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

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

[6] ANSYS. 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. Computer Architecture*, June 2003, pp. 2–13.

[8] 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. Computer-Aided Design of Integrated Circuits and Systems*, Jan. 2007.

[9] Y. Zhan and S. S. Sapatnekar, "A high efficiency full-chip thermal simulation algorithm," in *Proc. Int. Conf. Computer-Aided Design*, Oct. 2005.

[10] P. Liu, Z. Qi, H. Li, L. Jin, W. Wu, S. Tan, and J. Yang, "Fast thermal simulation for architecture level dynamic thermal management," in *Proc. Int. Conf. Computer-Aided Design*, Oct. 2005.

[11] T. Wang and C. Chen, "3-D thermal-ADI: A linear-time chip level transient thermal simulator," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 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," *Trans. Components and Packaging Technologies*, 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 Design Automation Conf.*, Jan. 2005.

[14] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, "IC thermal simulation and modeling via efficient multigrid-based techniques," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 9, pp. 1763–1766, Sept. 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. Quality of Electronic Design*, Mar. 2000, pp. 145–149.

[16] A. Majumdar, *Microscale energy transport in solids*. Taylor & Francis, 1998, ch. 1.

[17] D. G. Cahill, W. K. Ford, K. E. Goodson, G. D. Mahan, A. Majumdar, H. J. Maris, R. Merlin, and S. R. Phillpot, "Nanoscale thermal transport," *J. Applied Physics*, vol. 93, pp. 793–818, Jan. 2003.

[18] S. V. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Boltzmann transport equation-based thermal modeling approaches for hotspots in microelectronics," *Heat Mass Transfer*, vol. 42, pp. 478–491, 2006.

[19] 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," *Heat Transfer*, vol. 127, pp. 298–306, Mar. 2005.

[20] "International Technology Roadmap for Semiconductors," 2006, http://public.itrs.net/.

[21] J. Y. Murthy, S. V. J. Narumanchi, J. A. Pascual-Gutierrez, T. Wang, C. Ni, and S. R. Mathur, "Review of multi-scale simulation in sub-micron heat transfer," *Int. J. for Multiscale Computational Engineering*, vol. 3, pp. 5–32, 2005.

[22] M. A. Heaslet and R. F. Warming, "Radiative transport and wall temperature slip in an absorbing planar medium," *Int. J. Heat Mass Transfer*, vol. 8, pp. 979–994, 1965.

[23] B. Rutily, L. Chevallier, and J. Pelkowski, "K. Schwarzchild's problem in radiation transfer theory," *Journal of Quantitative Spectroscopy & Radiative Transfer*, vol. 98, pp. 290–307, 2006.

Authorized licensed use limited to: University of Michigan Library. Downloaded on February 8, 2009 at 14:07 from IEEE Xplore. Restrictions apply.