Full-Spectrum Spatial–Temporal Dynamic Thermal Analysis for Nanometer-Scale Integrated Circuits

Zyad Hassan, *Student Member, IEEE,* Nicholas Allec, *Student Member, IEEE,* Fan Yang, *Member, IEEE,* Li Shang, *Member, IEEE,* Robert P. Dick, *Member, IEEE,* and Xuan Zeng, *Member, IEEE*

***Abstract*—This article presents NanoHeat, a multi-resolution full-chip dynamic IC thermal analysis solution, that is accurate down to the scale of individual gates and transistors. NanoHeat unifies nanoscale and macroscale dynamic thermal physics models, for accurate characterization of heat transport from the gate and transistor level up to the chip–package level. A non-homogeneous Arnoldi-based analysis method is proposed for accurate and fast dynamic thermal analysis through a unified adaptive spatial–temporal refinement process. NanoHeat is capable of covering the complete spatial and temporal modeling spectrum of IC thermal analysis. The accuracy and efficiency of NanoHeat are evaluated, and NanoHeat has been applied to a large industry design. The importance of considering fine-grain temperature information is illustrated by using NanoHeat to estimate temperature-dependent negative-bias-temperature-instability (NBTI) effects. NanoHeat has been implemented and publicly released for free academic and personal use.**

***Index Terms*—Integrated circuit thermal factors, thermal modeling, model order reduction, integrated circuit reliability.**

I. Introduction

INTEGRATED circuit (IC) thermal analysis characterizes the heat dissipation process from numerous on-die heat sources, e.g., transistors, through the silicon die and packaging layers, to the ambient environment. With increasing power densities and power-induced design challenges, thermal issues have started receiving growing attention in IC design. Thermal analysis methods have been gradually adopted in commercial IC design flows, to quantify and mitigate temperature-induced timing, power consumption, and aging effects.

IC thermal analysis is a challenging problem. An IC may contain billions of transistors. Accurately modeling numerous nanometer-scale on-die heat sources during chip–package level thermal analysis introduces prohibitively high spatial modeling complexity. Moreover, the dynamic power consumptions hence the thermal profiles of individual transistors change at the nanosecond scale. Chip–package level temperature variations, on the other hand, can take as long as a few seconds, several orders of magnitude longer than transistor-level transient thermal effects.

The accuracy and efficiency of IC thermal analysis is thus determined by the modeling granularity – fine-grain discretization in both space and time yields high modeling accuracy, but also results in high time complexity. This leads to a question – *what is the appropriate modeling granularity for IC thermal analysis?* Clearly, there is no need to increase the modeling granularity indefinitely, as no further *useful* information is gained after reaching a certain limit. This limit is problem-dependent. For IC thermal analysis, a transistor is a basic building block, and the atomic heat source. Accurate IC thermal analysis can potentially require characterizing thermal effects with a spatial granularity as fine as the transistor length scale, and with a temporal granularity as fine as the transistor switching speeds.

In the past, device-level thermal effects have had little impact on circuit-level performance, power, and reliability. Therefore, existing work has focused on efficient chip–package level thermal analysis with accuracy at the scale of individual functional unit (10–100 µm length scale). Numerical analysis techniques, using the finite element or the finite difference methods, have been widely used for compact IC thermal modeling. IC temperature is approximated via functional-unit level time–space discretization.

IC design has now entered the nanometer regime. With device feature sizes reaching the nanometer scale, phonons (lattice vibrations), which are the main mechanism for heat transfer in semiconductors, travel ballistically, and create hotspots near the drain terminal region. Such device-level thermal effects have started to show increasingly significant impact on carrier mobility, leakage power consumption, and aging effects [1], [2], [3]. However, unless the spatial–temporal modeling granularity reaches device length scale, device-level thermal effects would be missed. Clearly, IC thermal analysis with the device-level modeling granularity in both space and time domains is challenging due to the huge computation and storage requirements. Furthermore, existing macroscale thermal modeling methods, e.g., the Fourier method, cannot accurately model device-level thermal effects such as ballistic phonon transport. Nanoscale thermal physics models, on the other hand, are computationally expensive.

Both steady-state and dynamic analysis methods have been developed in the past for full-chip IC thermal analysis. Much work has focused on steady-state thermal analysis [4], [5], [6]. Compared with steady-state thermal analysis, dynamic thermal analysis is much more challenging. Skadron et al. developed HotSpot, a dynamic chip–package level thermal analysis tool using both fixed-step and adaptive time-domain methods [7]. Smy et al. modeled IC transient heat flow using a 3-D transmission line matrix [8]. Liu et al. developed a frequency-domain moment matching technique for characterizing the

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-0954157, CCF-0702761 and CNS-0347941, and in part by the NSERC fellowship program.

Z. Hassan and L. Shang are with the Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, CO 80309, U.S.A.

N. Allec is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

F. Yang and X. Zeng are with the State Key Lab of ASIC & System, Fudan University, Shanghai, 200433, China.

R. Dick is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, U.S.A.

architectural-level dynamic temperature profile [9]. Wang et al. proposed a 3-D transient thermal simulator based on the alternating direction implicit (ADI) method [10]. Yang et al. developed ISAC, an adaptive chip–package level dynamic thermal analysis method [11]. Existing dynamic chip–package level thermal analysis methods, however, are unable to support nanometer-scale device-level spatial and temporal modeling granularities. Furthermore, they rely on the Fourier thermal physics model. The Fourier model with fixed material thermal conductivities fails at length scales comparable to the phonon mean free path (the average distance phonons travel before scattering events), and at time scales on which phonon scattering events occur [12]. Current device sizes and switching speeds have already reached those limits. This leaves the Fourier model inadequate for modeling device-level thermal effects.

Several methods have been used to model heat transport in a device. These include molecular dynamics and the Boltzmann transport equation (BTE) [13]. Although characterized with high accuracy, molecular dynamics are extremely computationally expensive, and thus have been only used to model heat transport in a few layers of atoms. The BTE method is much more efficient than molecular dynamics, and is able to accurately characterize phonon transport within a device.

Allec et al. [14] proposed a multi-scale IC thermal analysis solution that can characterize static device-level thermal effects, producing IC thermal profiles with transistor-level spatial resolution. This solution, however, only considered steady-state analysis. As shown in Section II, dynamic thermal effects can significantly influence circuit performance, reliability, and power consumption. In this work, we propose a dynamic IC thermal analysis technique that can handle spatial resolutions spanning from chip–package to nanometer scale, and temporal resolutions spanning from nanoseconds to seconds. We refer to this multiple spatial and temporal resolution solution as a *full-spectrum* thermal analysis framework. This work makes the following contributions:

1) it leverages macroscale and nanoscale dynamic thermal physics models, namely the Fourier heat equation and the BTE method, to accurately capture the dynamic thermal effects from chip–package level to individual transistors. This is done by deriving compact thermal models using the BTE method, which are consulted during full-chip analysis;
2) it proposes a unified spatial–temporal multi-resolution refinement algorithm that enables characterization of the dynamic thermal effects ranging from transistor-level to chip–package level spatial and temporal modeling granularities; and
3) it describes an accurate and numerically stable model order reduction (MOR) based dynamic thermal analysis method that employs a Non-Homogeneous Arnoldi (NHAR) process for generating the projection matrix. Unlike traditional moment-matching methods, our method can very efficiently match up to hundreds of moments of the frequency-domain response. It enables accurate dynamic analysis covering transistor, gate, functional-unit, chip, and package time scales.

II. Motivation

Thermal analysis has become increasingly important for reliable, power-efficient IC design. Fast and accurate thermal analysis allows detailed characterization of temperature-induced performance, power, and reliability effects. In this section, we illustrate the importance of multi-scale analysis including the device length and time scales during full-chip IC thermal analysis.

IC thermal characteristics span wide spatial and temporal scales. Existing work shows that the thermal time constant of a centimeter-scale cooling package is in the range of seconds. Within the silicon die, 10–100 µm functional-unit level hotspots are often observed, with thermal time constants of 100 µs–1 ms [15]. Hotspots with thermal time constants of 1–10 µs were observed in emerging three-dimensional ICs [16].

With continued technology scaling, nanometer-scale device-level hot phonon effects become increasingly significant. In semiconductors, lattice vibrations, or phonons are responsible for determining a device's temperature, with high temperatures corresponding to high phonon energy densities. In CMOS circuits, driven by the strong electric field across the device channel, free carriers travel at high speeds towards the drain, eventually interacting with the lattice, causing it to vibrate (or equivalently creating phonons), and consequently raising the temperature of the drain region. Eventually, these phonons lose their energy by scattering with "cold" phonons. However, since the average distance that phonons travel before suffering scattering events (the mean free path) is in the order of 100 nm [17], and because current device sizes are well below the phonon mean free path, phonons travel ballistically, resulting in a decreased energy loss by the "hot" phonons [17]. As a result, the device peak temperature significantly increases during switching. Despite the short period, e.g., sub-ns, in which a peak phonon density occurs, its effect on the device characteristics can be significant [1]. The non-equilibrium state (and thus a higher temperature) is aggravated by smaller device sizes, faster switching speeds, and the move to new technologies (SOI and FinFET).

Device-level thermal effects are starting to show significant impact on circuit performance, power consumption, and reliability. Lai and Majumdar [2] demonstrated that an increase in device temperature results in a significant reduction in the drive current, which is due to increased electron scattering near the high phonon density drain region. This results in a higher potential barrier seen by electrons in the source, which consequently leads to a decreased source injection rate [1]. Since the propagation delay of a CMOS circuit is determined by the drive current during switching, accurate characterization of delay requires accurate temperature prediction during switching. For power consumption, the short-circuit current has been predicted to increase due to the elevated equivalent temperature during the switching period [1]. Reliability is another major concern. Wang et al. showed that negative-bias-temperature-instability (NBTI) is greatly increased by ballistic phonon effects [3]. Therefore, the ability to identify hot transistors on the chip as well as accurately characterize their temperature is critical for accurately estimating temperature-dependent power consumption, timing, and reliability effects.

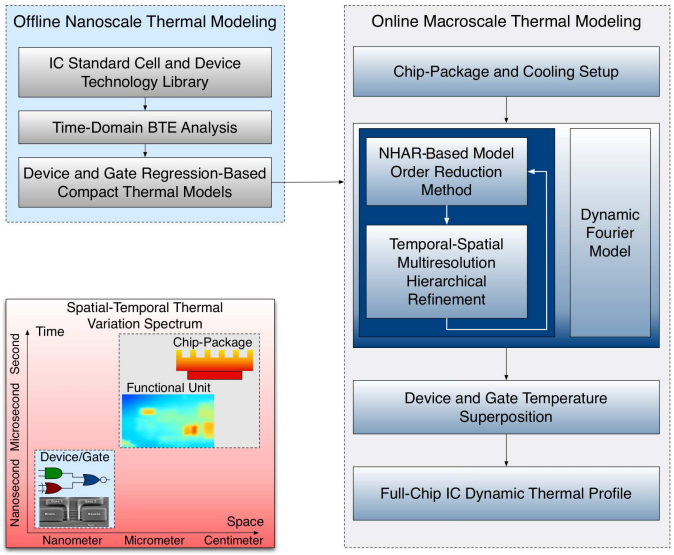


Fig. 1. Dynamic full-spectrum thermal analysis flow.

In summary, accurate characterization of nanometer-scale ICs requires detailed thermal analysis across the entire temporal and spatial spectrum, covering the chip–package level to individual transistors.

III. NanoHeat Overview

This section presents an overview of NanoHeat, the proposed dynamic thermal analysis solution. The aim of this work is to enable accurate and efficient characterization of the dynamic thermal effects of nanometer-scale ICs.

As described in Section II, this work is motivated by the increasing importance of nanoscale thermal effects in nanometer-scale ICs. Chip–package level IC thermal analysis should take gate and transistor-level thermal effects into consideration. The run-time thermal effect of an individual gate or transistor is influenced by its own heat as well as that of other on-chip components and devices. Therefore, accurate gate and transistor nanoscale thermal characterization must be integrated into the full-chip IC thermal analysis process. To this end, the following design challenges must be addressed.

- *Nanoscale thermal modeling challenge:* Accurate characterization of the nanometer-scale thermal effects of logic gates and devices requires computationally intensive thermal physics models.
- *Macroscale thermal modeling challenge:* Accurate analysis of the wide spectrum of spatial–temporal thermal characteristics of billion-transistor ICs has high computational complexity.

We propose NanoHeat, an efficient hierarchical thermal analysis approach to address these challenges. The solution flow is depicted in Figure 1. The flow consists of *(1) Offline nanoscale thermal modeling*, which uses an accurate but time-consuming dynamic BTE analysis to construct efficient gate and transistor compact thermal models based on the targeted IC technology library. These compact models are built once and used repeatedly during on-line IC thermal analysis. *(2) Online macroscale thermal modeling*, which conducts full-chip IC dynamic thermal analysis for the targeted IC designs in a hierarchical fashion using an efficient NHAR-based frequency-domain technique. NanoHeat offers accurate and efficient full-chip dynamic IC thermal analysis with nanometer and nanosecond scale accuracy.

Offline nanoscale thermal modeling: This characterizes the dynamic thermal effects of individual nanometer-scale gates and transistors. Since nanoscale thermal effects cannot be captured using conventional thermal models, such as the Fourier model, NanoHeat uses time-domain dynamic BTE analysis to accurately capture device-level phonon behavior. Dynamic BTE analysis has high computational complexity; performing it for every gate and transistor of a billion-transistor IC is infeasible.

In NanoHeat, nanoscale thermal modeling is an offline process. Given an IC standard cell and device technology library, device-level dynamic BTE simulations are carried out for each type of device and gate. The input power consumption waveforms are obtained using SPICE simulations. For each type of device, the BTE dynamic thermal simulation takes approximately 5 hours. Since each technology library only contains limited types of devices, and each device type only needs to be simulated once, model generation time is reasonable.

Based on the BTE analysis results, a regression-based compact model is constructed for each device and gate type, which models the thermal effects as a function of the device structure, dynamic power profile, and time. Note that offline nanoscale thermal modeling only needs to consider the device self-heating effects, i.e., the transistor and gate thermal effects due to its own power consumption.

In NanoHeat, the constructed regression-based compact thermal models are organized as look-up tables, which are integrated with the proposed macroscale thermal model for full spatial–temporal scale IC thermal analysis. More specifically, during full-chip IC dynamic thermal analysis, incorporating device and gate level dynamic thermal effects only requires device–gate thermal model table lookup and thermal effect superposition over time.

Online macroscale thermal modeling: This characterizes the dynamic thermal effects from IC chip–package level down to gate and transistor level length and time scales. It uses the Fourier heat flow model to characterize macroscale thermal effects. NanoHeat unifies temporal–spatial multi-resolution hierarchical refinement with fast and numerically stable Non-Homogeneous Arnoldi (NHAR) based model order reduction (MOR) method. Together, combined with the compact device and gate nanoscale thermal model library, the proposed macroscale thermal modeling technology enables fast and accurate characterization of the dynamic thermal effects on spatial and temporal scales that vary by several orders of magnitude.

Billion-transistor IC dynamic thermal analysis is a daunting task. A full flat device-level implementation of IC thermal analysis is computationally intractable. On the other hand, an IC on-die thermal profile has spatial and temporal correlation. An on-die hotspot is a result of its hot subcomponents, e.g., logic gates and transistors. The proposed solution leverages strong IC on-die spatial–temporal thermal correlation, and conducts multi-resolution hierarchical dynamic thermal analysis to

efficiently and accurately identify and characterize the on-die hotspots, from chip-package level, to functional unit level, to logic gate and device level. At each level of the refinement hierarchy, the dynamic thermal profile is characterized using the efficient NHAR-based frequency-domain thermal analysis method.

More specifically, the hierarchical analysis method starts with the chip–package level at which functional-unit level spatial and temporal modeling granularities are used. Therefore, full-chip dynamic power profiles can be obtained using efficient architectural power simulation. The functional units with high peak temperature and/or spatial-temporal temperature variation are then identified. Multi-scale hierarchical spatial refinement is applied to these elements, and dynamic thermal analysis is conducted at finer temporal and spatial granularities to identify thermal hotspots and spatial-temporal temperature variations. Low-level power analysis methods, e.g., gate-level, are then needed. For each element of interest, this refinement process stops when no further significant temperature change is identified, or when the logic gate level is reached. In the final stage, the compact device and gate thermal models constructed offline are used. The temperatures of the gates within the hotspots are obtained and superimposed on macroscale thermal analysis results, yielding the full-chip IC dynamic thermal profile.

The proposed hierarchical approach effectively improves the dynamic IC thermal analysis efficiency. Since an IC typically has only a few hotspots, finer-grained dynamic thermal and power analysis only needs to be used for a small part of the chip. As shown in Section VI, given the industrial IC design with over 150 million transistors, chip–package level thermal analysis finishes in less than 20 seconds. 4 functional-unit level hotspots are observed at the chip–package level. Hierarchical refinement and analysis is then applied, which identifies 5–6 hotspots inside each of these four functional units. Further refinement and analysis result in no significant thermal variations, and thus gate and device-level thermal superposition is applied at these hotspot regions to produce the final IC dynamic thermal profile. Fewer than 100,000 hotspot transistors need to be considered. The overall analysis process takes less than an hour.

IV. Offline Nanoscale Thermal Modeling

This section presents the proposed nanoscale dynamic thermal modeling method. In this work, we have developed a time-domain dynamic BTE solver for device-level thermal modeling. The phonon BTE is a semi-classical equation that describes the transport of a distribution function of phonons in non-metallic solids [18], [19]. Here we use the gray phonon BTE under the relaxation time approximation. Although we use the gray BTE model, other approximations for solving the BTE, such as those used in the semi-gray model could be applied. The gray BTE model assumes a single group velocity and relaxation time for phonons, which are independent of their frequency and polarization. It does not take into account details of the scattering mechanisms of phonons or the phonon dispersion curves. The relaxation time approximation used for the gray BTE is valid when the length scales are larger than the heat carrying phonon wavelengths [20], and allows the phonon scattering processes to be taken into account as a deviation from the equilibrium distribution. Using these approximations the BTE can be mathematically expressed as [12]:

$$\frac{\partial e''}{\partial t} + \nabla \cdot (\mathbf{v}_g e'') = \frac{e_0 - e''}{\tau_{eff}} + q_{vol}, \tag{1}$$

where e'' is the energy density per unit solid angle of the phonons, $\mathbf{v}_g$ is the phonon group velocity vector, e_0 is the equilibrium energy density, τ_{eff} is the relaxation time (i.e., the time between independent scattering events), and q_{vol} is the volumetric heat source. The left side of the equation describes the heat transfer due to the group velocity vector of the phonons. The right side describes the rate of change in the phonon distribution due to scattering and particle creation.

The equilibrium energy is given by [18]

$$e^0 = \frac{1}{4\pi}\int_{4\pi} e'' d\Omega = \frac{1}{4\pi} C(T_L - T_{ref}), \tag{2}$$

where Ω is the angular discretization, C is the specific heat, T_L is the lattice temperature, and T_{ref} is the specific heat reference temperature. The lattice temperature, T_L, can be calculated using Equation 2 given the equilibrium energy density.

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

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

where k is the thermal conductivity.

The electron–phonon interactions inside devices are modeled by heat sources, which are denoted by the term q_{vol} in Equation 1. Its time dependent value can be derived from device power consumption information, which can be obtained using circuit simulation.

To obtain the dynamic thermal profile of a device, first e'' is determined by solving Equation 1, using the dynamic power profile of the device (represented by q_{vol}). The equilibrium energy density, and thus thermal profile, can then be obtained using Equation 2. It should be noted that there is a feedback loop between device temperature and power consumption. Thus, it is necessary to iteratively compute the temperature and the power consumption until convergence.

NanoHeat uses a time-domain method for performing device-level thermal analysis. As explained in Section II, a device's temperature rise due to its own power consumption occurs momentarily, and the phonon density drops within a few picoseconds. Because of the absence of any accumulation of phonons between different switching events [17], it is sufficient to simulate a device's temperature variation during a single clock cycle. With simulation times restricted to such a small time interval, the time-domain methods are more accurate and efficient than frequency-domain methods. The core of the dynamic BTE solver is a time-domain, fourth-order Runge-Kutta [21] solver, which iteratively advances in time until the required simulation time is reached. At each time step, the dynamic BTE is solved for each angle.

Dynamic BTE analysis is too slow for direct use in full-chip thermal analysis. However, although an IC contains hundreds of millions to billions of transistors, in reality, there are a limited number of distinct types of gates and transistors

in a given IC standard cell and device technology library. Since identical gates and transistors exhibit the same thermal response, we only need to characterize the response once for each of the them. This observation eliminates the need to perform dynamic BTE analysis for each gate and device. NanoHeat is thus equipped with a device and gate regression-based model for fast full-chip dynamic thermal analysis. Each device and gate is simulated using the dynamic BTE solver, and a regression model is constructed. The regression model is a look-up table that models the thermal effects as a function of the device dynamic power profile, geometric structure, and time. Each row in the table contains the thermal response for a specific device or gate. The thermal response is given in the form of the temperature of the device as a function of time.

During full-chip IC thermal analysis, the temperature of an individual device or gate is the superposition of the results from the macroscale analysis and the device-level analysis. With the relatively small time-scales at the device level, device-level effects can be considered as a perturbation to thermal effects at higher levels, and thus can be superimposed on the inter-device effects.

V. Macroscale Thermal Modeling

This section presents the proposed macroscale thermal modeling technology. Section V-A describes the thermal physics models. Section V-B details the NHAR-based frequency-domain dynamic thermal analysis method with unified adaptive spatial–temporal refinement.

V.A. Thermal Physics Models

Despite the fact that existing macroscopic methods cannot accurately model nanometer-scale device-level thermal effects, past work has shown that they are fast and sufficiently accurate from chip–package level down to inter-device scale. In this work, the Fourier method is selected for macroscale thermal modeling.

The equation governing heat diffusion according to the Fourier model can be mathematically expressed as

$$C\frac{\partial T}{\partial t} = \nabla \cdot k\nabla T + p, \tag{4}$$

where C is the volumetric specific heat, T is the temperature, t is time, k is the thermal conductivity, and p is the heat source power density.

To use the Fourier heat diffusion model, the chip is spatially discretized into numerous thermal elements. The resulting discretized equation can be written in matrix form:

$$C\dot{T}(t) + GT(t) = P(t), \tag{5}$$

where C and G are the thermal capacitance and conductance matrices, and $T(t)$ and $P(t)$ are the time-varying temperature and power consumption vectors. For a chip discretized into N thermal elements, $C, G \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$, and $T(t), P(t) \in \mathbb{R}^{\mathrm{N}\times 1}$.

V.B. NHAR-Based Frequency-Domain Dynamic Thermal Analysis with Unified Adaptive Spatial–Temporal Refinement

This section presents an accurate and numerically stable model order reduction-based dynamic thermal analysis method, which provides fast and accurate characterization of dynamic IC thermal effects from second to nanosecond time scales. It is integrated with a hierarchical multi-resolution spatial refinement method. The selection of appropriate spatial modeling granularity (i.e., the minimal feature size) and temporal modeling resolution (i.e., the required number of frequency moments) are conducted in unison, thereby enabling fast and accurate multi-resolution IC thermal characterization.

V.B.1) Time-Domain Analysis vs. Frequency-Domain Analysis: Dynamic analysis can be carried out using time-domain or frequency-domain methods. Time-domain methods, such as the Euler method and the Runge-Kutta methods, use step-by-step numerical integration to estimate the transient thermal profile. Time-domain methods are fast and accurate for short-time scales, however the execution time and estimation error increase as time scales increase.

In contrast with time-domain methods, frequency-domain methods, such as moment matching, derive an analytic approximation function to directly compute the dynamic thermal profile as a function of run-time power. Frequency-domain methods are fast and accurate over long time scales. However, existing explicit moment matching techniques such as asymptotic waveform evaluation (AWE) [22] rely on Padé approximation to generate a reduced-order transfer function. AWE suffers from a computationally expensive setup phase for deriving the analytical function. In addition, the accuracy of frequency-domain methods depend on the number of moments. The more matched moments available, the more natural frequencies of the system are captured, and therefore, the higher the thermal analysis accuracy. It has been shown that Padé approximation-based explicit moment matching methods suffer instability and ill-conditioned matrices when the number of moments is increased [23]. This renders them unsuitable for use for short time-scales, where they tend to have low accuracy.

In summary, due to the high setup cost required for frequency-domain methods, time-domain methods have superior performance for short simulation times. Therefore, in NanoHeat, transistor-level cycle-by-cycle transient thermal analysis is conducted using a time-domain method (see Section IV). On the other hand, NanoHeat uses a new NHAR-based frequency-domain method for efficient long-time scale chip–package level thermal analysis that retains high accuracy down to the nanosecond-scale gate-level analysis.

V.B.2) NHAR-Based Frequency-Domain Thermal Analysis: This section describes the proposed frequency-domain method to rapidly and accurately characterize dynamic IC thermal effects from the second-scale chip–package level, the microsecond-scale functional-unit level, to the nanosecond-scale gate and transistor levels.

Several frequency-domain-based transient thermal analysis solutions have been proposed in the past. An Arnoldi-based MOR technique was proposed by Codecasa, D'Amore, and Maffezzoni [24] for reducing the order of thermal networks. This technique however, only computes the thermal profiles of individual circuits consisting of a few transistors, and does not handle full-chip analysis. MOR-based full-chip thermal analysis solutions have also been proposed [25], [26], [9]. Tsai and Kang [25] use a hybrid frequency-domain/time-domain method to achieve significant reductions in the order of the system. Wang and Chen [26] describe an improved extended

Krylov subspace method. Liu et al. [9] use the periodicity of the power consumption at the architectural-level to simplify moment matching. However, the accuracy of the proposed solutions is limited to functional-unit granularity, and finer-grained thermal variations are ignored, leading to inaccuracies in estimating the chip performance, power consumption and reliability.

Most MOR methods are projection-based [27]. They generate a projection matrix to transform the system matrices into lower order matrices, such that the frequency-domain transfer function of the reduced system matches that of the original system up to the nth moment. Krylov subspace methods, such as Arnoldi [28] and Lanczos [29], are computationally efficient for generating the projection matrix, and are numerically stable with an increasing number of moments [30]. The wide application of Krylov space MOR-based methods, however, has been hindered by their increased computational cost for systems with a large number of ports. This issue has been addressed by using input-dependent MOR methods in which the specific input waveforms are considered in the moment matching process. Input-dependent methods such as the extended Krylov subspace (EKS) [31], and the improved extended Krylov subspace (IEKS) [32] methods have been proposed. Although efficient, the higher-order terms resulting from taking the input waveforms into account prevent the direct use of numerically stable orthogonalization procedures, such as the Arnoldi technique. Instead EKS/IEKS rely on an incremental orthogonalization procedure, which involves a power iteration process that results in loss of information of the higher order moments as the computed vectors rapidly converge to the eigenvector corresponding to the largest eigenvalue of the matrix. This leads to the failure of these methods to steadily improve the accuracy of the reduced-order system by increasing the reduced order.

Our proposed dynamic thermal analysis method uses an input-dependent MOR technique, which employs an NHAR method for generating the projection matrix. This NHAR-based MOR method uses an Arnoldi-like orthogonalization procedure that guarantees stability. The use of the numerically stable Arnoldi-like orthogonalization procedure is enabled by initially applying a linearization scheme. The NHAR-based MOR method can thus achieve high accuracy by matching hundreds of moments of the original system response, making it accurate for very short time-scales, e.g., nanoseconds. Combined with device-level dynamic thermal analysis (see Section IV), the proposed method is thus able to capture IC temperature variations on all relevant time scales.

The formulation of the proposed dynamic thermal analysis input-dependent MOR technique will now be given. An IC dynamic thermal profile is a function of on-chip power consumption as well as the packaging and cooling solution. In other words, on-chip devices are thermally correlated. Therefore, device X's temperature can be expressed in the frequency domain as

$$T_X(s) = \sum_{j=1}^{N} H_{X,j}(s) P_j(s), \tag{6}$$

where $T_X(s)$ is device X's temperature, $P_j(s)$ is device j's power consumption, and $H_{X,j}(s)$ is the impulse response relation describing the thermal impact of device j's power consumption on device X. N is the total number of on-chip devices, or heat sources.

Characterizing the inter-device thermal impact boils down to deriving a relation for the right side of Equation 6, i.e., the temperature response to a specific input vector $P(s)$. This relation can be expressed as

$$T_X(s) = T_0^x + T_1^x s + T_2^x s^2 + \dots. \tag{7}$$

Moment matching MOR techniques derive an approximate relation for $T_X(s)$ by building a reduced order model, such that the first n moments of the reduced system's response and the original response are exactly matched. To apply MOR to all on-chip devices, Equation 5 is first transformed into the frequency domain via Laplace transform, which yields

$$sCT(s) + GT(s) = P(s). \tag{8}$$

The input-dependent MOR algorithm then proceeds as follows. First, $P(s)$ is expanded in an infinite Taylor series around zero frequency, and truncated to l terms:

$$sCT(s) + GT(s) = u_0 + u_1 s + u_2 s^2 + \ldots + u_{l-1} s^{l-1}. \tag{9}$$

To avoid overflow in the Taylor expansion of inputs, a scaling method [31] can be employed. The scaling method adaptively selects a factor by which the original frequency variable s is scaled. This allows scaling the Taylor expansion coefficients so that they fall within an acceptable range, thus allowing the generation of hundreds of power input moments. This is necessary to capture the high frequency components of the inputs, while preventing numerical instability resulting from overflow in the expansion. By further expanding $T(s)$ in Taylor series around the zero frequency, we have

$$(sC+G)(T_0+T_1 s+T_2 s^2+\cdots) = u_0+\ldots+u_{l-1}s^{l-1}, \tag{10}$$

where the coefficient of the i-th term in the Taylor series, $T_i \in \mathbb{R}^{N\times 1}$, is known as the i-th moment of $T(s)$. Next, the projection matrix is obtained via the numerically stable NHAR method. The projection matrix $Q \in \mathbb{R}^{\mathrm{N}\times\mathrm{n}}$ spans the moment subspace of $T(s)$, i.e., span$\{T_0, T_1, \cdots, T_n\}$. Here N is the order of the original system, n is the order of the reduced system, and $n \ll N$. The NHAR process avoids explicit moment calculation for generating the moment subspace span$\{T_0, T_1, \cdots, T_n\}$, because this often leads to numerical instability. Instead, it relies on the linearization of Equation 9 with high-order terms in the right side, i.e.,

$$\begin{aligned} sCT(s) + GT(s) &= u_0 + s\sum_{i=1}^{l-1} u_i s^{i-1} \\ &= u_0 + sJz(s), \end{aligned} \tag{11}$$

where

$$J = [u_1 u_2 \ldots u_l],\ z(s) = \begin{bmatrix} 1 & s & \ldots & s^{l-1} \end{bmatrix}^T.$$

It can be verified that $z(s)$ satisfies the following relation:

$$z(s) - sFz(s) = e_1, \tag{12}$$

where e_1 is the first column of the identity matrix $I_{l\times l}$ and

$$F = [f_{i,j}], f_{i,j} = \begin{cases} 1, & i = j+1 \text{ and} \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

By combining Equation 11 and Equation 12, the linearization form is expressed as

$$\left(\begin{bmatrix} G & 0 \\ 0 & I \end{bmatrix} - s \begin{bmatrix} -C & J \\ 0 & F \end{bmatrix} \right) \begin{bmatrix} T(s) \\ z(s) \end{bmatrix} = \begin{bmatrix} u_0 \\ e_1 \end{bmatrix}, \quad (14)$$

which can be rewritten as

$$(I - sA) \begin{bmatrix} T(s) \\ z(s) \end{bmatrix} = \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}, \quad (15)$$

where

$$A = \begin{bmatrix} -G^{-1}C & G^{-1}J \\ 0 & F \end{bmatrix}, \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} = \begin{bmatrix} G^{-1}u_0 \\ e_1 \end{bmatrix}.$$

The system in Equation 15 is a linearized form of the original system in Equation 9. The i-th moment of $\begin{bmatrix} T(s) \\ z(s) \end{bmatrix}$ in Equation 15 is exactly the i-th moment of $T(s)$ in Equation 9. Using this approach, the numerical instability associated with explicit moment calculation is avoided. The price paid is a slight increase in the order of the system from N in Equation 9 to $N + l$ in the linearized form of Equation 15. This price is small since $l \ll N$.

From the linearized system in Equation 15, generating the projection matrix Q is straightforward. First, the numerically stable Arnoldi process can be employed to generate the orthonormal basis $\bar{Q}$ of the n-th order Krylov subspace $\mathcal{K}_n \left\{ A, \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} \right\}$, which spans the moment subspace of $\begin{bmatrix} T(s) \\ z(s) \end{bmatrix}$. One way to obtain the projection matrix Q is by orthonormalizing the first N rows of $\bar{Q}$. Alternatively, we propose a more efficient approach described in Algorithm 1.

Algorithm 1 NHAR process

Input: order of the reduced system n, A, ϕ_0, ϕ_1
Output: the projection matrix Q

1: $\begin{bmatrix} q \\ p \end{bmatrix} = \frac{1}{||\phi_0||} \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}$, $P = [p]$, $Q = [q]$
2: **for** $i = 1 : n-1$ **do**
3: $\quad \begin{bmatrix} q \\ p \end{bmatrix} = A \begin{bmatrix} q \\ p \end{bmatrix}$, $h = Q^T q$
4: $\quad \begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} q \\ p \end{bmatrix} - \begin{bmatrix} Q \\ P \end{bmatrix} h$, $\alpha = ||q||$
5: $\quad$ **if** $\alpha \approx 0$ **then**
6: $\quad\quad$ stop(breakdown)
7: $\quad$ **end if**
8: $\quad \begin{bmatrix} q \\ p \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} q \\ p \end{bmatrix}$, $P = [P, p]$, $Q = [Q, q]$
9: **end for**

Step 1 of Algorithm 1 is for initializing the orthonormalization process. The $n-1$ times orthonormalization steps are carried out in steps 2–9. The NHAR process is similar to the classical Arnoldi procedure [28]. The difference is that in the NHAR process, the vectors to be orthonormalized are now partitioned into upper and lower parts and only the upper parts, which correspond to the columns of the projection matrix Q, are orthonormalized during the process. The result of the NHAR process is the projection matrix Q, which is the accurate orthonormal basis of the moment subspace span$\{T_0, T_1, \cdots, T_n\}$. Due to the Arnoldi-like orthonormalization process, the NHAR process is numerically stable and is capable of generating the projection matrix Q spanning hundreds of moments of $T(s)$. Once the projection matrix Q has been obtained, the system matrices are projected, yielding a reduced order system whose moments match the first n moments of the original system.

V.B.3) Unified Adaptive Spatial–Temporal Refinement: The accuracy and efficiency of dynamic IC thermal analysis depends on the spatial and temporal modeling granularities, i.e., the geometric feature size and the number of moments of the reduced order system. In other words, the isothermal assumption is made within each discretized thermal element to minimize runtime complexity, but intra-element spatial and temporal thermal features are also ignored. For instance, existing chip–package level thermal analysis with the modeling granularity of individual functional units ignores device-level thermal effects. In addition, spatial and temporal modeling granularities are interdependent. Smaller thermal elements with lower heat capacities and higher run-time power variations exhibit larger transient thermal variations, thus requiring more thermal elements for accurate short time-scale analysis. More specifically, from coarse-grained functional units to nanometer-scale transistors, the thermal time constant varies from milliseconds to nanoseconds; the required number of moments differs significantly. Furthermore, inter-device thermal correlation is spatially and temporally heterogeneous. The chip, package, and cooling solution serve as low-pass filters for heat transfer. As inter-device distance increases, inter-device thermal correlation, especially high-frequency transient thermal interaction, decreases. Consider the example shown in Figure 2, which characterizes the thermal impact of devices A, B, C, and D on device X. Devices A and B are close to X, and therefore have significant and heterogeneous transient thermal impact on X. Fine-grained spatial and short time scale temporal thermal modeling is thus required. On the other hand, C and D are far away from X. Due to the low-pass filtering effect of chip and cooling package, the thermal impacts of C and D on X have long time scales and are spatially uniform. Coarse-grained modeling in both space and time can then be applied to optimize modeling efficiency.

The aforementioned observations indicate that spatial and temporal modeling granularities must be carefully decided during thermal analysis. Fast and efficient multi-resolution IC thermal analysis calls for an adaptive spatial–temporal modeling method.

We propose a unified adaptive spatial and temporal refinement approach, in which the spatial granularity is hierarchically, adaptively refined from chip–package level, functional-unit level, to gate-level scales, and the number of moments is adaptively determined based on the temporal characteristics of the thermal elements at the corresponding spatial granularity.

The idea of hierarchical spatial refinement is depicted in Figure 3, where two levels of spatial resolution are shown.

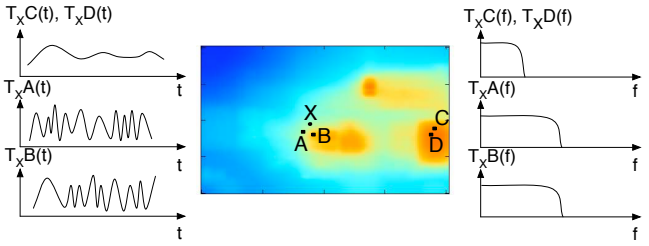


Fig. 2. Inter-device thermal impact.

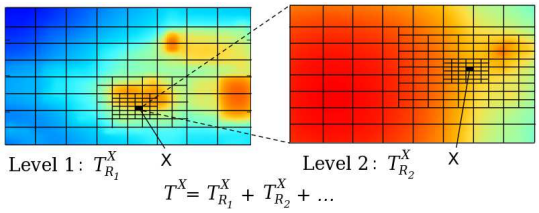


Fig. 3. Hierarchical multi-resolution spatial refinement.

Transistor X for which we are interested in computing the temperature, lies inside the block colored in black at each level, and the response is calculated at this point. At each level, starting with the coarsest (level 1), adaptive spatial refinement is applied with a fine grid close to the point of interest, and a coarser grid further away. The thermal response of this block due to other heat sources is characterized using the NHAR MOR method. The number of moments to generate using the NHAR method is adaptively chosen according to the thermal time constants at this level which, as explained earlier, depends on the element sizes. Next, the block is further partitioned as shown in level 2. The thermal response for the block containing transistor X in level 2 is then characterized. This process is repeated until no further significant thermal variations are observed, or the transistor-level granularity is reached. The thermal response for transistor X is computed as the sum of responses from all levels.

More specifically, to estimate the temperature of device X, as heat sources located in nearby regions have more heterogeneous, short time-scale thermal impact, fine spatial granularities with a high number of moments are required. Coarser spatial granularities and fewer moments are used as we move farther from the targeted device. Together, device X's temperature is the superposition of the responses from on-chip heat sources modeled with heterogeneous spatial–temporal modeling granularities, i.e.,

$$T_X(s) = \sum_{j=1}^{m} T^X_{R_j}(s), \tag{16}$$

where $T^X_{R_j}(s)$ is device X's temperature response to the heat sources in region j, and m is the total number of regions. The response from region j, $T^X_{R_j}$ can be written as follows:

$$T^X_{R_j}(s) = T^j_0 + T^j_1 s + T^j_2 s^2 + \ldots + T^j_{n_j} s^{n_j}, \tag{17}$$

where n_j is the number of moments to be considered in region j. The number of grid elements in region j is denoted by N_j. We determine the thermal impact of each region starting with those farthest from X and ending with those nearest. Thus, the algorithm first handles the farthest region, in which we use coarse-grained grid elements and a small number of moments, i.e., small N_j and n_j. The response $T^X_{R_j}$ is derived using the NHAR-based MOR method described in Section V-B.2. As we proceed to closer regions, spatial refinement is adaptively applied based on the region's distance to X and the number of moments is adaptively increased based on the temporal thermal characteristics at the current spatial granularity. The NHAR algorithm is applied to characterize the response of each region. The appropriate increase in moments is determined by computing the relative error resulting from using n vs. $n+1$ moments.

To understand the potential of this hierarchical approach, we contrast it to the non-hierarchical one. Without adaptive spatial–temporal refinement, the whole chip is considered as one region, N is determined by the finest spatial granularity needed, and n is the minimum number of moments that enables capturing the fastest thermal variations. Both N and n are large, and they influence the run time of the NHAR algorithm. On the other hand, using the hierarchical approach, $\sum_{j=1}^{m} N_j \ll N$, and $n_j \ll n$, except for the nearest region to device X, where $n_j = n$. For example, for a 1 cm×1 cm chip, a non-adaptive approach will require $N \sim 10^{14}$ elements, and $n > 100$ moments. For the adaptive approach, N_j doesn't exceed 10^6 elements for any region j, with the number of moments usually falling below 100 moments. The complexity of determining the responses is thus significantly reduced.

We observed that once the temperature of device X is determined, characterizing the temperatures of many other devices on chip becomes simpler. For instance, the responses of devices A and B in Figure 2 to power consumed in devices C and D is similar to device X's response to the power consumed in devices C and D. Since this response has already been determined, it can be reused when computing the temperatures of devices A and B. This is true for all devices in close proximity to X, where the closer the device is to X, the more responses are shared, and the less computation is necessary. This sharing is not possible for the non-hierarchical approach since the chip is considered to be one region. Consequently, the response has to be rederived for each device.

NanoHeat combines the techniques discussed in this section to accurately and efficiently characterize the dynamic thermal effects from the IC chip–package level down to the gate and transistor level length and time scales.

VI. Results

In this section, we evaluate and demonstrate the use of NanoHeat. NanoHeat has been implemented as a software package and has been publicly released for free academic and personal use. Section VI-A evaluates NanoHeat, and Section VI-B describes NanoHeat's library interfaces and functionality, and explains its use. Section VI-C describes its use for full spatial–temporal spectrum dynamic thermal analysis of an industry IC design containing over 150 million transistors. Finally, Section VI-D demonstrates its use in temperature-dependent reliability analysis.

VI.A. Evaluating NanoHeat

Since existing thermal analysis solutions can only support either chip–package level or device-level dynamic thermal

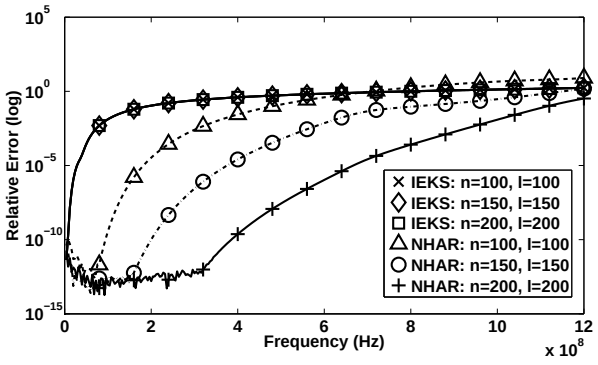


Fig. 4. The numerical stability of NHAR MOR compared against IEKS.

analysis, we evaluate NanoHeat's macroscale thermal modeling and device-level thermal analysis techniques individually. Section VI-A.1 compares the numerical stability of the NHAR-based MOR method against the improved extended Krylov subspace (IEKS) MOR method [32]. Section VI-A.2 evaluates NanoHeat's NHAR-based dynamic thermal analysis method. Section VI-A.3 evaluates NanoHeat's device-level nanoscale thermal model against published results.

VI.A.1) Numerical Stability of NHAR MOR vs. IEKS: In this section, we evaluate the numerical stability of the proposed NHAR MOR method, compared to IEKS, a widely used input-dependent MOR method.

The test setup involves an RC-network with an order of 49,600, which contains 1,025 piece-wise linear (PWL) current sources. We use both methods to reduce the order of the circuit to 100, 150, and 200. The errors of the output response of the reduced-order models are computed relative to the output response of the original circuit.

Figure 4 shows the results, where n indicates the reduced order, and l indicates the truncation order of the Taylor series expansion of the inputs. It can be seen from the results that the relative errors for IEKS remain constant when the order of the reduced system increases from 100 to 150 to 200. In contrast, the NHAR-reduced systems can match the response more accurately, and in a wider frequency range, as the order of the reduced system increases. The NHAR method stably matches up to hundreds of moments of the frequency-domain response, and steadily increase the accuracy by increasing the order of the reduced system.

VI.A.2) NHAR-Based Dynamic Thermal Analysis Method: It is known that frequency-domain methods have better accuracy and efficiency than time-domain methods for long time scale simulation. However, traditional frequency-domain methods, such as the AWE method, have low accuracy for short time scale analysis. This is due to the instability they suffer for high-order moment calculation. A limited number of low-order moments cannot provide sufficient accuracy for short time scale analysis. The NHAR-based method can efficiently derive hundreds of moments of the frequency-domain response. It therefore produces accurate results for short time scales. In this section, we evaluate the accuracy of NanoHeat's NHAR-based frequency-domain method for short time scale thermal analysis compared to a time-domain, globally adaptive fourth-order Runge-Kutta (GARK4) solver. We also show the efficiency of our proposed NHAR-based dynamic thermal analysis technique, by comparing its performance with an AWE-based dynamic thermal analysis technique.

TABLE I
ACCURACY OF NHAR-BASED DYNAMIC THERMAL ANALYSIS METHOD

Simulation Time	AWE		NHAR	
	# of moments	e_{avg} (%)	# of moments	e_{avg} (%)
10 ns	Failed	-	140	0.49
10 µs	Failed	-	100	0.09
10 ms	9	1.71	12	0.37

A quad-core chip-multiprocessor chip–package setup is considered in this experiment. Each core is a 2 GHz Alpha 21264 like core, containing 15 functional units. The silicon die is 9.88 mm×9.88 mm, and 50 µm thick. The cooling setup contains a 6.9 mm thick copper heat sink using forced air cooling. Cycle-by-cycle power profile is generated using the M5 full-system simulator [33] with a Wattch-based EV6 power model [34], by running 12 different multithreaded benchmarks on the cores from the SPEC2000 [35], MediaBench [36], and ALPBench [37] benchmark suites. The chip is partitioned into 2,376 3-D thermal elements, and dynamic thermal analysis is conducted using the NHAR-based, GARK4, and AWE methods, for 10 ns, 10 µs, and 10 ms.

The AWE-based frequency-domain method is also considered to demonstrate the need of a large number of moments for accurate short time-scale dynamic thermal analysis. The difference metric used is average error relative to the GARK4 method, $e_{avg} = 1/E \sum_{e \in E} |T_e - T'_e|/|T_e - T_a|$, where T_a is the ambient temperature and E is the set of points in the active layer at which the temperature is evaluated. Subtracting T_a from the denominator is necessary to evaluate ambient-temperature-independent errors. For the AWE-based and the NHAR-based frequency-domain techniques, we select the minimum required number of moments that yields an error of less than 1%, if possible, for all the benchmarks compared to the time-domain GARK4 method.

Table I shows the experimental results. The "# of moments columns" show the required number of moments to achieve an error of less than 1%, whenever feasible. An entry marked as "Failed" indicates that no moment count value could achieve the desired accuracy, due to the instability problem. The "e_{avg}" columns show the average error among the 12 benchmarks using the above error metric.

This study demonstrates that the proposed NHAR-based method can achieve high accuracy for short time scale thermal analysis. As frequency-domain based techniques are inherently accurate for long time scales, the NHAR-based method is suitable for use across the full range of dynamic thermal analysis time scales. The AWE-based frequency-domain technique, on the other hand, failed to produce accurate results for very short time scales, i.e., the 10 ns and the 10 µs cases.

TABLE II
EFFICIENCY OF NHAR-BASED DYNAMIC THERMAL ANALYSIS METHOD

GARK4	AWE		NHAR		
CPU time (s)	CPU time (s)	Speedup vs. GARK4 (×)	CPU time (s)	Speedup vs. GARK4 (×)	Speedup vs. AWE (×)
55,273.7	15,593.0	3.5	23.7	2335.2	658.8

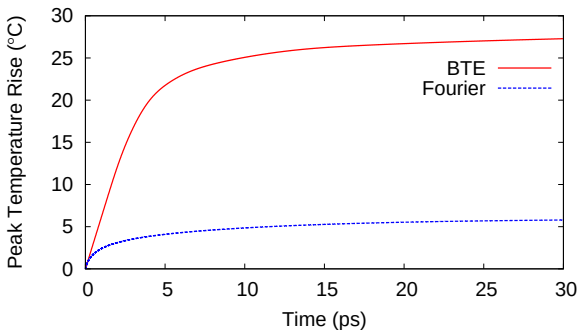


Fig. 5. Peak temperature rise predicted by the BTE.

Next, we evaluate the efficiency of our proposed NHAR-based dynamic thermal analysis technique. The same test setup described above is used, and the chip is simulated for 10 s using the three methods. The results are shown in Table II.

The speed advantage of the NHAR-based technique over the time-domain method is apparent for long time-scales (over 2,000 ×). Since accurate characterization of the dynamic thermal effects of a modern IC design requires at least tens of milliseconds, seconds, or even minutes of simulation, the performance advantage of NanoHeat over existing time-domain methods is significant. On the other hand, for short-time scale analysis, time-domain methods are faster. Therefore, in NanoHeat, short time scale device-level thermal analysis is conducted using a time-domain method. Furthermore, although the AWE-based method is able to produce accurate results for long time scales, the NHAR technique shows superior performance.

VI.A.3) Device-Level Model Validation: NanoHeat supports macroscale (via Fourier heat flow modeling) and nanoscale (via the dynamic BTE method) thermal modeling. The Fourier heat flow model has been validated by existing chip–package thermal analysis methods. In this section, we focus on evaluating the dynamic BTE solver by comparing it with prior experimental results [38].

The experiment conducted by Yang et al. [38] involves a heat source embedded in a silicon substrate. This setup resembles the nanoscale heat generation and transport scenario in MOSFET devices, with the device power consumption represented by the heat source. The device dimensions are 100 nm×50 nm×1 µm. Initially, the entire device is at ambient temperature. At time $t = 0$, the heat source is turned on. The simulation is carried out for 30 ps. The heat flux in the y-direction at the centerline of the device is evaluated at t =1.5 ps and 10 ps, and the spatial peak temperature of the device is observed as a function of time.

The peak temperature rise as a function of time is plotted in Figure 5. The results are in excellent agreement with those of Yang et al. [38] with a relative error of less than 1%. This shows the reliability of our BTE solver's results.

It is important to note that the Fourier model fails to accurately characterize the temperature rise in the device. For this experiment [38], as shown in Figure 5, the Fourier model only predicts a 5 °C increase in temperature, compared to the actual 25 °C increase predicted by the BTE.

VI.B. NanoHeat's Library Interface

We have implemented NanoHeat as a C++ software package and publicly released the package for free academic and personal use. NanoHeat's functionalities have been implemented within the chip–package thermal analysis tool, ISAC [11]. NanoHeat is available for download at http://eces.colorado.edu/~hassanz/NanoHeat. In this section, we describe the library interfaces and explain how each of them is used. Please refer to [11] for an overview of ISAC's basic functionalities.

VI.B.1) Chip–Package Level Down to Gate–Transistor Level Thermal Analysis Interfaces: The static and dynamic thermal profiles of the chip can be generated using *solve_static(·)* and *step_dynamic(·)* methods. The *zoom_and_solve_static(·)* method zooms in on a region of the chip and performs steady-state thermal analysis. This method takes as input the four edges of the desired region, as well as the chip power profile. The corresponding method for performing dynamic thermal analysis is *zoom_and_step_dynamic(·)*. In addition to the inputs of *zoom_and_solve_static(·)*, it takes an extra input, which is the duration of the simulation. Both methods return an STL vector of thermal element temperatures. The interfaces provided allow observing the chip thermal profile at different granularities starting from the chip–package level spatial and temporal granularities down to the gate–transistor level spatial and temporal granularities. NanoHeat also provides interfaces for visualizing the output thermal profiles using the plotting utility, `gnuplot`.

The *report_device_temperatures(·)* computes temperatures of individual devices. This method takes as input the thermal profile computed using the static or dynamic analysis methods, and a rectangle. The method computes the temperatures of all devices in the given rectangle based on (1) the temperatures in the provided thermal profile, and (2) the device BTE temperatures provided from the look-up table, which is generated using the device BTE solver. The use of the device BTE solver is explained in Section VI-B.2. This method returns a vector of devices containing their temperatures. Computed device temperatures can be used to predict performance, power consumption, and reliability.

VI.B.2) Gate and Transistor Level Thermal Analysis Interfaces: NanoHeat provides the *Device_structure* class to handle complex device structures. A device structure is described using an input file that contains the device geometry, materials, and BTE solver options. The input file can be read via the *input(·)* method, and can be solved using the *solve(·)* method. Lastly, the look-up table to be consulted during full-chip analysis can be generated using the *generateLUT(·)* method.

The device-level thermal modeling interfaces can also be used to study the effect of device geometry and power consumption on a device's thermal profile. NanoHeat also provides interfaces for visualizing the device thermal profile generated from the BTE solver.

VI.C. Full Spatial–Temporal Spectrum IC Thermal Analysis

This section demonstrates the use of NanoHeat for full spatial–temporal dynamic IC thermal analysis using a 65 nm industry chip design. The chip contains over 150 million transistors on a 16 mm×16 mm silicon die. The cooling setup

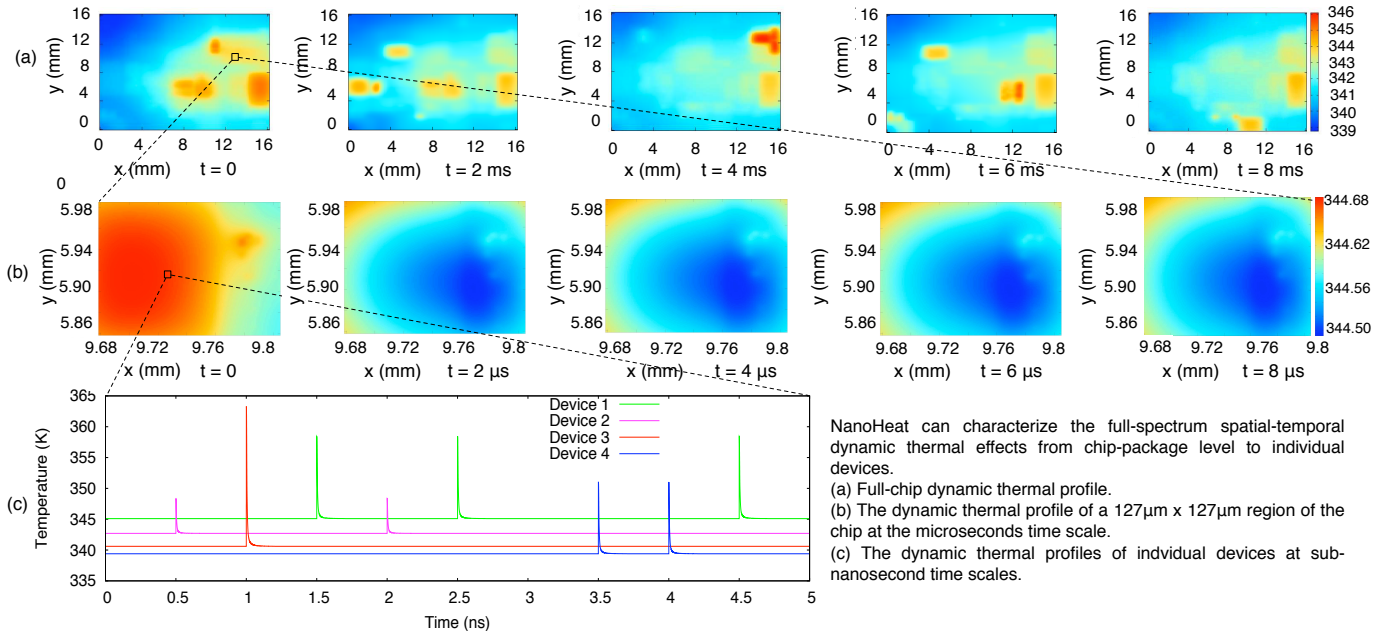


Fig. 6. Full-spectrum spatial–temporal dynamic IC thermal analysis.

consists of a 34 mm×34 mm aluminum heat sink with forced air cooling.

NanoHeat is capable of conducting full-spectrum spatial–temporal dynamic thermal analysis with reasonable runtime and memory usage. This is essential to determine the complete dynamic thermal characteristics of nanometer-scale ICs. In this study, the simulation is carried out for 8 ms. The results are shown in Figure 6, which shows the chip-package level dynamic power profile at $t = 0$, 2, 4, 6, and 8 ms, and an enlarged 127 µm×127 µm region of the chip, in which the thermal profile is observed at a finer time-scale. The snapshots shown are for $t = 0$, 2, 4, 6, and 8 µs. Figure 6 also shows the peak temperature run-time profiles of four transistors.

NanoHeat makes it possible, for the first time, to observe dynamic thermal effects in nanometer-scale transistors and millimeter-scale IC across the complete range of relevant time scales. As shown in Figure 6, significant spatial and temporal temperature variations are observed at the chip–package level over long time scales which is due to significant spatial and temporal power variations. On the other hand, the temperatures within a small region of a functional unit are spatially smooth, which is due to the fact that the operations of the devices within such small region are highly correlated. The average switching activities of these devices over a time duration comparable to the functional unit level thermal time constant are uniform. This leads to uniform spatial thermal profile. In addition, this study also shows that the temporal thermal variation within a functional unit is small at the microsecond scale. As shown in Figure 6, significant spatial and temporal temperature variations are observed within individual devices.

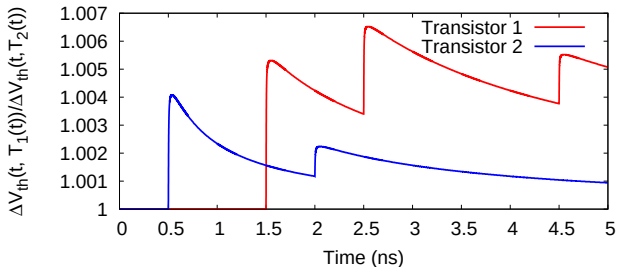
The importance of considering transistor-level temperature variations can be observed from the results. The actual chip temperature variations that happen on short time scales and small length scales are completely ignored when large time steps and element sizes are used in the analysis. As discussed in Section II, and demonstrated in Section VI-D, this can lead to inaccuracies in estimating temperature dependent factors, such as leakage power, reliability and circuit performance.

VI.D. Temperature-Dependent NBTI Analysis

As mentioned in the earlier discussion in Section II, the hotspots appearing at the transistor level can have severe impact on its characteristics. Elevated transistor temperatures have been shown to cause a substantial drop in the drive current [2]. Although the temperature peak occurs momentarily, its characterization is important since it takes place during the transistor switching period, which determines its propagation delay. Short-circuit currents are aggravated by the higher temperature during transistor switching [1], leading to higher chip power consumption. Chip aging effects are strong functions of temperature, thus high temperatures cause reliability degradation.

Existing chip–package thermal analysis tools are able to identify the hotspots at the functional-unit level. However, identifying the hotspots at this level is not enough for finding the actual transistor temperature; gate and device analysis is also needed. Next, we illustrate the importance of considering fine-grain thermal effects when characterizing chip performance, power consumption, and reliability by comparing the results of using coarse-grained, and fine-grained thermal analysis when estimating NBTI-induced threshold voltage shifting.

The NBTI effect has received great interest in the past decade. It is one of the dominant chip aging processes [39]. NBTI degrades circuit performance by shifting PMOS threshold voltages. NBTI is strongly dependent on temperature; a higher transistor temperature experienced during the stress phase severely accelerates the threshold voltage shifting [40], [41]. In addition, it has been demonstrated that neglecting temperature variation by assuming a constant transistor temperature during the ON and OFF phases could result in an error as high as 52.6 % [42]. Wang et al. [3] provided experimental evidence that shows that the NBTI effect is exacerbated by



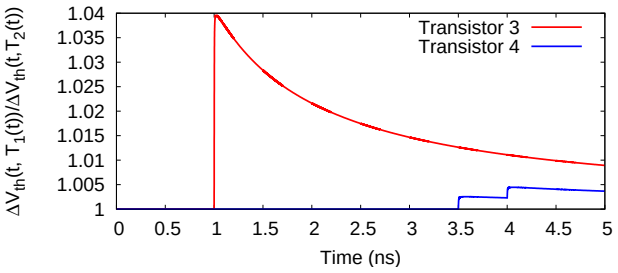


Fig. 7. Short-term NBTI effect.

the hotspot located in the transistor drain region.

In order to characterize the NBTI effect, we carry out NBTI analysis using the temperature results obtained in Section VI-C. We utilize the NBTI model developed by Zhang and Orshansky [43] to show how temperature affects NBTI in the short term (cycle-by-cycle) and at longer time scales. Zhang and Orshansky's model [43] can characterize NBTI under arbitrary dynamic temperature variation. According to this model, the shift in threshold voltage as a function of time and temperature history can be written as

$$\Delta V_{th}\left[t, T(t)\right] = c\left[\int_0^t e^{(\frac{E_a}{k}[1/T_{ref}-1/T(t^\dagger)])}dt^\dagger\right]^{1/6}, \quad (18)$$

where c is a constant that depends on the initial Si–H bond density, the diffusivity of H_2, and the transistor properties; E_a is the activation energy; k is the Boltzmann constant; and T_{ref} is a constant reference temperature. The accepted range of NBTI degradation with temperature is about 2.15X every 50°C [44]. Thus, we choose the values for the parameters in Equation 18 such that the degradation follows the same trend.

We start with short-term temperature-dependent NBTI analysis. To show the error resulting from using spatially and temporally coarse-grained thermal analysis, we compute $\Delta V_{th}[t, T_1(t)]$ normalized to $\Delta V_{th}[t, T_2(t)]$, where $T_1(t)$ is the dynamic temperature obtained using fine-grained spatial and temporal resolutions, and $T_2(t)$ is that obtained using coarse-grained resolutions. Figure 7 shows the results for the four transistors whose thermal profiles are shown in Figure 6.

As observed from the results, the threshold voltage estimation is affected by the resolution with which the temperature information is provided. The temporary temperature peaks that appear during transistor switching affect the temporal threshold voltage shifting. The actual threshold voltage shift is higher than that estimated using coarse-grained temperatures. Although the difference might not be significant at such short-time scales, as will be shown next, the error accumulated over time leads to a larger deviation at long-time scales.

The long-time scale results are shown in Figure 8, where $\Delta V_{th}/c$ is plotted over time. As we can see, the estimation error due to ignoring fine-grained thermal effects increases over time. This study shows the importance of considering nanoscale thermal effects during NBTI analysis. Since the temperature-dependent NBTI effect has significant impact on IC performance and lifetime reliability, NanoHeat can be used to facilitate NBTI analysis and optimization. Recent studies have proposed architectural and circuit-level techniques to address the NBTI effects. For instance, one of our collaborative

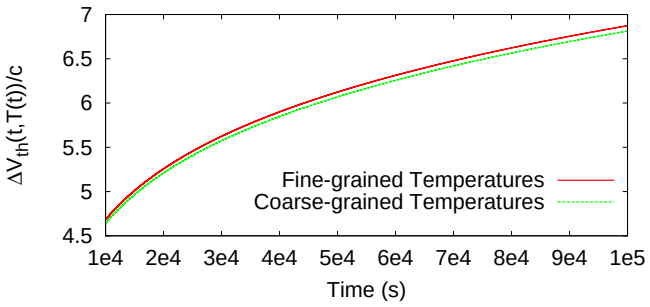


Fig. 8. Long-term NBTI effect.

work focuses on statistical NBTI optimization using circuit-level gate resizing [45]. The integration of NanoHeat and statistical NBTI analysis will provide more accurate guidance to statistical gate resizing for NBTI optimization.

We illustrate the use of NanoHeat for NBTI analysis only to provide an example. It is our hope that designers also apply this multi-scale thermal analysis framework to other problems. Note that the results of this analysis strongly depend on the NBTI parameters as well as on the input power profile waveforms used, both of which carry uncertainties. Thus the accuracy of NBTI estimates is inherently limited by the accuracy of those inputs. However, our proposed framework eliminates another source of inaccuracy. To obtain transistor temperatures, designers previously carried out chip–package level thermal analysis with functional-unit accuracy, and assumed transistor temperatures to be the same as the temperature of the functional-unit they belong to. However, as explained in Section II, nanoscale thermal effects have become more pronounced due to transistor scaling so this assumption is no longer valid. Thus, utilizing inaccurately computed transistor temperatures in NBTI analysis (or in the estimation of any other IC design metric) leads to inaccurate results. The proposed framework eliminates this inaccuracy by providing an efficient hierarchical method by which IC thermal profiles can be computed with accuracy down to the transistor level, thus enabling more accurate estimation of IC temperature-dependent effects.

VII. Conclusions

Thermal analysis is important for reliable, power-efficient IC design. With technology scaling, nanometer-scale device-level thermal effects can no longer be ignored during IC thermal analysis. However, full-chip thermal analysis yielding accurate results for individual devices is challenging.

In this paper, we have presented a dynamic thermal analysis solution capable of handling the complete IC spatial and tem-

poral thermal spectrum, from chip–package level, functional-unit level down to individual gates and transistors; from centimeters and seconds down to nanometers and nanoseconds. We have evaluated the accuracy of our proposed solution and demonstrated its use for multi-resolution spatial–temporal dynamic thermal analysis of a large industry IC design.

Performing full-chip IC thermal analysis with transistor-level accuracy can help obtain better estimates of temperature-dependent effects, such as IC performance, power consumption, and reliability. A temperature-dependent NBTI analysis example is used to illustrate the importance of high-resolution thermal information in achieving accurate aging estimates. In addition, being able to observe the chip-wide temperature variations with transistor-level spatial and temporal scales will enable designers to better understand how design decisions affect the thermal behavior of ICs.

REFERENCES

[1] J. Rowlette and K. Goodson, "Fully coupled nonequilibrium electronphonon transport in nanometer-scale silicon FETs," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 220–232, Jan. 2008.

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

[3] Y. Wang, H. Luo, K. He, R. Luo, H. Yang, and Y. Xie, "Temperature-aware NBTI modeling and the impact of input vector control on performance degradation," in *Proc. Design, Automation & Test in Europe Conf.*, Apr. 2007, pp. 546–551.

[4] P. Li, L. T. Pileggi, M. Ashghi, and R. Chandra, "Efficient full-chip thermal modeling and analysis," in *Proc. Int. Conf. Computer-Aided Design*, Nov. 2004, pp. 319–326.

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

[6] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proc. Int. Symp. Low Power Electronics & Design*, Aug. 2003, pp. 78–83.

[7] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "HotSpot: a compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. VLSI Systems*, vol. 14, no. 5, pp. 501–524, May 2006.

[8] T. Smy, D. Walkey, and S. Dew, "Transient 3D heat flow analysis for integrated circuit devices using the transmission line matrix method on a quad tree mesh," *Solid-State Electronics*, vol. 45, no. 7, pp. 1137–1148, July 2001.

[9] 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.

[10] 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.

[11] Y. Yang, C. Zhu, Z. P. Gu, L. Shang, and R. P. Dick, "Adaptive multi-domain thermal modeling and analysis for integrated circuit synthesis and design," in *Proc. Int. Conf. Computer-Aided Design*, Nov. 2006.

[12] 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.

[13] S. Sinha and K. E. Goodson, "Review: Multiscale thermal modeling in nanoelectronics," *Int. J. for Multiscale Computational Engineering*, vol. 3, pp. 107–133, 2005.

[14] N. Allec, Z. Hassan, L. Shang, R. P. Dick, and R. Yang, "ThermalScope: multi-scale thermal analysis for nanometer-scale integrated circuits," in *Proc. Int. Conf. Computer-Aided Design*, Nov. 2008, pp. 603–610.

[15] 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.

[16] C. Zhu, Z. P. Gu, L. Shang, R. P. Dick, and R. Joseph, "Three-dimensional chip-multiprocessor run-time thermal management," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 8, pp. 1479–1492, Aug. 2008.

[17] S. Sinha, E. Pop, R. E. Dutton, and K. E. Goodson, "Non-equilibrium phonon distributions in sub-100nm silicon transistors," *ASME J. of Heat Transfer*, pp. 638–647, July 2006.

[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] 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.

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

[21] W. H. Press, B. P. F. S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[22] L. Pillage and R. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 9, no. 4, pp. 352–366, Apr 1990.

[23] D. F. Anastasakis, N. Gopal, S. Y. Kim, and L. T. Pillage, "On the stability of moment-matching approximations in asymptotic waveform evaluation," in *Proc. Design Automation Conf.*, 1992, pp. 207–212.

[24] 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.

[25] C.-H. Tsai and S.-M. Kang, "Fast temperature calculation for transient electrothermal simulation by mixed frequency/time domain thermal model reduction," in *Proc. Design Automation Conf.*, 2000, pp. 750–755.

[26] T.-Y. Wang and C. C.-P. Chen, "SPICE-compatible thermal simulation with lumped circuit modeling for thermal reliability analysis based on modeling order reduction," in *Proc. Int. Symp. Quality Electronic Design*, Mar. 2004, pp. 357–362.

[27] J. Wang, C.-C. Chu, Q. Yu, and E. Kuh, "On projection-based algorithms for model-order reduction of interconnects," *IEEE Trans. Circuits and Systems I*, vol. 49, no. 11, pp. 1563–1585, Nov 2002.

[28] L. M. Silveira, M. Kamon, I. Elfadel, and J. White, "A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits," in *Proc. Int. Conf. Computer-Aided Design*, Nov. 1996, pp. 288–294.

[29] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by Padé approximation via the Lanczos process," in *Proc. European Design Automation Conf.*, Sept. 1994, pp. 170–175.

[30] R. W. Freund, "Reduced-order modeling techniques based on Krylov subspaces and their use in circuit simulation," in *Applied and Computational Control, Signals, and Circuits*, 1999, pp. 435–498.

[31] J. M. Wang and T. V. Nguyen, "Extended Krylov subspace method for reduced order analysis of linear circuits with multiple sources," in *Proc. Design Automation Conf.*, 2000, pp. 247–252.

[32] Y. Cao, Y.-M. Lee, T.-H. Chen, and C. C.-P. Chen, "HiPRIME: hierarchical and passivity reserved interconnect macromodeling engine for RLKC power delivery," in *Proc. Design Automation Conf.*, 2002, pp. 379–384.
[33] 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, 2006.
[34] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proc. Int. Symp. Computer Architecture*, June 2000, pp. 83–94.
[35] J. L. Henning, "SPEC CPU2000: Measuring CPU performance in the new millennium," *Computer*, pp. 28–35, July 2000.
[36] C. Lee, M. Potkonjak, and W. H. M. Smith, "Mediabench: A tool for evaluating and synthesizing multimedia and communications systems," http://cares.icsl.ucla.edu/MediaBench.
[37] M.-L. Li, R. Sasanka, S. V. Adve, Y.-K. Chen, and E. Debes, "The ALPbench benchmark suite for complex multimedia applications," in *Proc. Int. Symp. Workload Characterization*, Oct. 2005, pp. 34–35.
[38] 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.
[39] S. Chakravarthi, A. Krishnan, V. Reddy, C. Machala, and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," in *IEEE Int. Reliability Physics Symposium Proc.*, April 2004, pp. 273–282.
[40] M. Alam and S. Mahapatra, "A comprehensive model of PMOS NBTI degradation," *Microelectronics Reliability*, vol. 45, no. 1, pp. 71–81, Jan. 2005.
[41] V. Huard, M. Denais, and C. Parthasarathy, "NBTI degradation: From physical mechanisms to modeling," *Microelectronics Reliability*, vol. 46, no. 1, pp. 1–23, Jan. 2006.
[42] H. Luo, Y. Wang, K. He, R. Luo, H. Yang, and Y. Xie, "Modeling of PMOS NBTI effect considering temperature variation," in *Proc. Int. Symp. Quality of Electronic Design*, Mar. 2007, pp. 139–144.
[43] B. Zhang and M. Orshansky, "Modeling of NBTI-induced PMOS degradation under arbitrary dynamic temperature variation," in *Proc. Int. Symp. Quality of Electronic Design*, Mar. 2008, pp. 774–779.
[44] M. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, "A comprehensive model for PMOS NBTI degradation: Recent progress," *Microelectronics Reliability*, vol. 47, no. 6, pp. 853–862, 2007.
[45] Y. Lu, L. Shang, H. Zhou, H. Zhu, F. Yang, and X. Zeng, "Statistical reliability analysis under process variation and aging effects," in *Proc. Design Automation Conf.*, 2009, pp. 514–519.

Zyad Hassan (S'08) received his B.Sc. degree in Electronics and Electrical Communications from Cairo University, Cairo, Egypt in 2006 and his M.Sc. degree in Electrical and Computer Engineering from the University of Colorado at Boulder. He is currently pursuing his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Colorado at Boulder. His research interests include computer-aided design of integrated circuits and formal verification.

Nicholas Allec (S'02) received his B.E. from Lakehead University, Thunder Bay, Canada and his M.Sc. from Queen's University, Kingston, Canada. He is currently pursuing his Ph.D. at the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. His research interests include thermal, device, circuit, and x-ray detector modeling and simulation.

Fan Yang (M'08) received the B.E. degree in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2003. He received the Ph.D degree in microelectronics from Fudan University, Shanghai, China, in 2008. He is now an assistant professor of the microelectronics department, Fudan University. His research interests include model order reduction and circuit simulation.

Li Shang (S'99-M'04) is an Assistant Professor of the Department of Electrical, Computer, and Energy Engineering, University Colorado at Boulder. He received his Ph.D. degree from Princeton University in 2004 and his B.S. degree with honors from Tsinghua University. His work was nominated for the Best Paper Award at ISLPED 2010, ICCAD 2008, DAC 2007, and ASP-DAC 2006. His work was selected for publication in MICRO Top Picks 2006, and won the Best Paper Award at PDCS 2002. He serves as an Associate Editor of IEEE Transactions on VLSI Systems and serves on the technical program committees of several embedded systems, CAD/VLSI, and computer architecture conferences. He won his department's Best Teaching Award in 2006. He received an NSF CAREER award in 2010.

Robert P. Dick (S'95-M'02) is an Associate Professor of Electrical Engineering and Computer Science at the University of Michigan. He received his Ph.D. degree from Princeton University in 2002 and his B.S. degree from Clarkson University in 1996. He worked as a Visiting Professor at Tsinghua University's Department of Electronic Engineering in 2002, as a Visiting Researcher at NEC Labs America in 1999, and was on the faculty of Northwestern University from 2003–2008. Robert received an NSF CAREER award and won his department's Best Teacher of the Year award in 2004. His technology won a Computerworld Horizon Award and his paper was selected as one of the 30 most influential appearing in DATE during the past 10 years in 2007. He is an Associate Editor of IEEE Trans. on VLSI Systems, a Guest Editor for ACM Trans. on Embedded Computing Systems, and serves on the technical program committees of several embedded systems and CAD/VLSI conferences.

Xuan Zeng (M'97) received the B.Sc. and Ph.D. degrees in electrical engineering from Fudan University, Shanghai, China, in 1991 and 1997.

She is currently a Full Professor with the Microelectronics Department and serves as the Director of State Key Laboratory of ASIC and Systems, Fudan University. She was a Visiting Professor with the Electrical Engineering Department, Texas A&M University, College Station, and Microelectronics Department, Technische Universiteit Delft, Delft, The Netherlands, in 2002 and 2003. Her research interests include design for manufacturability, high-speed interconnect analysis and optimization, analog behavioral modeling, circuit simulation, and ASIC design.

Dr. Zeng received the first-class Award of Electronic Information Science and Technology from the Chinese Institute of Electronics in 2005. She received the second-class Award of Science and Technology Advancement and the Cross-Century Outstanding Scholar Award from the Ministry of Education of China in 2006 and 2002. She received the award of "IT Top 10" in Shanghai in 2003. She served on the Technical Program Committee of the IEEE/Association for Computing Machinery Asia and South Pacific Design Automation Conference in 2000 and 2005.