An Efficient Method for Analyzing On-Chip Thermal Reliability Considering Process Variations

YU-MIN LEE, National Chiao Tung University
PEI-YU HUANG, Industrial Technology Research Institute

This work provides an efficient statistical electrothermal simulator for analyzing on-chip thermal reliability under process variations. Using the collocation-based statistical modeling technique, first, the statistical interpolation polynomial for on-chip temperature distribution can be obtained by performing deterministic electrothermal simulation very few times and by utilizing polynomial interpolation. After that, the proposed simulator not only provides the mean and standard deviation profiles of on-chip temperature distribution, but also innovates the concept of thermal yield profile to statistically characterize the on-chip temperature distribution more precisely, and builds an efficient technique for estimating this figure of merit. Moreover, a mixed-mesh strategy is presented to further enhance the efficiency of the developed statistical electrothermal simulator.

Experimental results demonstrate that (1) the developed statistical electrothermal simulator can obtain accurate approximations with orders of magnitude speedup over the Monte Carlo method; (2) comparing with a well-known cumulative distribution function estimation method, APEX [Li et al. 2004], the developed statistical electrothermal simulator can achieve 215× speedup with better accuracy; (3) the developed mixed-mesh strategy can achieve an order of magnitude faster over our baseline algorithm and still maintain an acceptable accuracy level.

Categories and Subject Descriptors: B.7.2 [**Integrated Circuits**]: Design Aids; B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

General Terms: Design, Algorithms, Performance, Reliability

Additional Key Words and Phrases: Electrothermal simulation, thermal analysis, chip temperature, thermal reliability, process variation, simulation

ACM Reference Format:
Lee, Y.-M. and Huang, P.-Y. 2013. An efficient method for analyzing on-chip thermal reliability considering process variations. ACM Trans. Des. Autom. Electron. Syst. 18, 3, Article 41 (July 2013), 32 pages.
DOI: http://dx.doi.org/10.1145/2491477.2491485

1. INTRODUCTION

As technology scales down to the sub-90*nm* node, on-chip power densities increase rapidly. Hence, power dissipation and thermal management have become important issues of VLSI design. High on-chip temperature distribution and thermal gradients

Preliminary versions of this article appeared in *Proceedings of the IEEE International Systems-on-Chip Conference (SoCC'10)* [Chang et al. 2010] and in *Proceedings of the IEEE/ACM International Asia and South Pacific Design Automation Conference (ASP-DAC'12)* [Huang et al. 2012].
This work was supported in part by the National Science Council of Taiwan under grants NSC 99-2220-E-009-035, 100-2221-E-009-074, and 101-2221-E-009-168, and by the Industrial Technology Research Institute, Taiwan.
Authors' addresses: Y.-M. Lee, Department of Electrical and Computer Engineering, National Chiao Tung University, HsinChu, Taiwan; email: yumin@nctu.edu.tw; P.-Y. Huang, Industrial Technology Research Institute, Taiwan.

DOI: http://dx.doi.org/10.1145/2491477.2491485

drastically degrade the circuit performance and design reliability, operating temperatures seriously affect gate delays [Kumar and Kursun 2006], and nonuniform on-chip temperature distribution induces timing faults [Bota et al. 2004]. Since on-chip power consumption is proportional to operating temperatures, thermal runaway might occur if thermal-related issues are not carefully considered in package design [Vassighi and Sachdev 2006]. To ensure design qualities, such as performance, power consumption, and reliability, researchers have been devoted to dealing with thermal-related issues in physical design [Tsai et al. 2006; Liu et al. 2008]. To provide the related thermal cost for optimization engines of physical design, several efficient deterministic thermal simulators [Wang and Chen 2003; Huang et al. 2006; Yang et al. 2007; Huang and Lee 2009] have been developed to predict on-chip temperature profile. However, these thermal simulators only provide thermal information with deterministic on-chip power consumption.

As predicted by the international technology roadmap for semiconductors (ITRS), leakage power consumption increases dramatically and has become a dominant portion of total power consumption [ITRS 2010]. Moreover, the scaling down of technology causes physical parameter variations to be non-ignorable, and this fact leads to substantial on-chip leakage power fluctuations. As pointed out by Pang and Nikolic, 8% of process variations can lead to about 25% of on-chip leakage power fluctuations [2009]. Therefore, physical parameter variations are essential to be considered for on-chip power estimation techniques [Chang and Sapatnekar 2007; Shen et al. 2010b, 2010a]. Since on-chip temperature is transfered from on-chip power consumption, its distribution is impacted by process variations inducing leakage power fluctuations. However, deterministic thermal analyzers [Wang and Chen 2003; Huang et al. 2006; Yang et al. 2007; Huang and Lee 2009], which did not take process variations into account in their power models, are not adequate to precisely provide thermal reliability estimation under process variations. Therefore, the on-chip temperature profile should be treated statistically under process variations, and statistical thermal simulation techniques are essential, especially for leakage dominated technologies [Huang et al. 2009; Jaffari and Anis 2008].

To provide the statistical characteristics of on-chip temperature distribution, Jaffari and Anis [2008] proposed a recursive log-normal approximation algorithm to obtain mean and standard deviation profiles of the on-chip temperature distribution. Compared with the Monte Carlo (MC) simulations, they have successfully demonstrated its efficiency and accuracy for estimating mean and standard deviation profiles of the temperature distribution in the macro-architectural level. Instead of constructing the different leakage power model for each different macro/gate type, they built the different leakage power model for each bin (grid) on a die. Since optimization engines, such as floorplanners or placers, might disturb positions of macros or gates, their related leakage power models need to be rebuilt after an optimization loop is executed. Therefore, the efficiency of their approach will be degraded while casting into thermal-aware design flows, because their approach needs to execute the time-consuming HSPICE simulation numerous times and fit curves for reestablishing leakage power models. In addition, their recursive log-normal approximation algorithm is restricted to the form of their proposed leakage power models. However, the leakage power model is getting more complicated for maintaining an acceptable accuracy level as the technology continuously scales down. To overcome the leakage power model restriction, a statistical thermal simulation framework that has high capability of adopting complex and accurate power models for any technology generations is required.

Besides the leakage power model issues, the figure of merit for on-chip temperature distribution is still ambiguous if only its mean and standard deviation profiles are

reported.[1] Therefore, instead of only reporting mean and standard deviation profiles, a more precise figure of merit for the statistical characteristics of on-chip temperature distribution should be addressed to ensure the thermal reliability or to provide the thermal related cost for thermal-aware design engines.

In this work, a statistical simulation framework is developed for characterizing the on-chip temperature distribution. With the aim of dealing with the restriction issues of leakage power models, providing a more precise figure of merit for ensuring thermal reliability and being more easily incorporated into statistical performance analysis and design engines, technical key points and advantages of this work are summarized as follows.

(1) Compared with the bin-based model [Jaffari and Anis 2008], a cell-based model is adopted for characterizing leakage powers. With the precharacterizing property, the reestablishing process of leakage power models can be avoided, while macros or gates are exchanged by optimization engines, such as floorplanners or placers.
(2) Adopting the concept of sparse collocation-based methods, a statistical electrothermal simulation framework is developed to generate the statistical polynomial expression of on-chip temperature distribution. Compared with that of Jaffari and Anis [2008], the developed framework is more flexible for complex and precise leakage powers models.
(3) This work not only provides the mean and standard deviation profiles of on-chip temperature distribution, but also introduces the concept of *thermal yield profile* to statistically characterize the on-chip temperature distribution more precisely, and builds an efficient estimating technique for this figure of merit.
(4) Without sacrificing the accuracy, a mixed-mesh strategy is presented and integrated into the baseline method of our statistical electrothermal simulation engine to further enhance its efficiency.

The rest of this article is organized as follows. Section 2 motivates the concept and essentialness of on-chip *thermal yield profile*, investigates the accuracy of existing cell-based leakage power models, and indicates that complex leakage current models are required for maintaining acceptable accuracy level. After that, the problem formulation and the modeling technique of device parameters are described in Section 3. Then, the developed statistical electrothermal simulator is detailed in Section 4, and experimental results are given in Section 5. Finally, the conclusion and potential applications of the developed simulation framework are presented in Section 6.

2. MOTIVATION ILLUSTRATIONS

2.1. Concept of On-Chip Thermal Yield Profile

Because of process variations, on-chip temperature at an arbitrary position $\mathbf{r}$ is a random variable. Therefore, the deterministic thermal analysis with nominal on-chip power profile can no longer be a good figure of merit for identifying the hotspot location of the chip. Moreover, even if the temperature at any arbitrary position $\mathbf{r}$ has been treated as a random variable, it will still be ambiguous if only the mean ($\mu_T(\mathbf{r})$) and standard deviation ($\sigma_T(\mathbf{r})$) profiles of on-chip temperature distribution are delivered. For example, only using the mean profile of on-chip temperature distribution as a figure of merit is very likely (about 50%) to incorrectly indicate hotspot locations. Furthermore, as both $\mu_T(\mathbf{r})$ and $\sigma_T(\mathbf{r})$ are delivered, by utilizing the Chebyshev inequality, a large temperature value will be estimated to ensure the lower bound

[1]Please see the interpretation stated in Section 2.1.

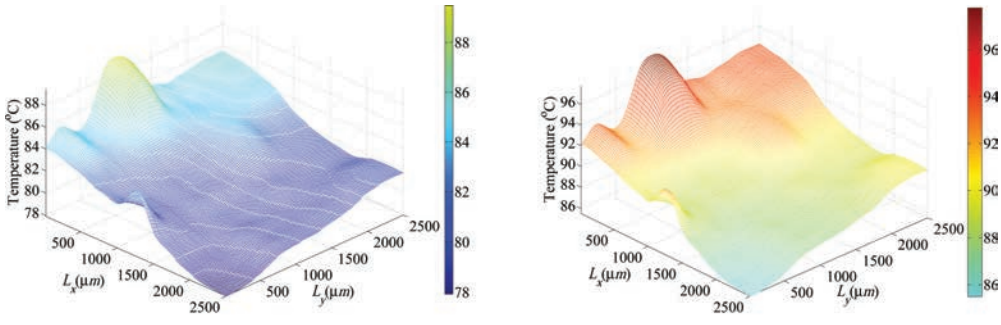

Fig. 1. An example to demonstrating the gap between the temperature profile that really achieves 90% thermal reliability and the temperature profile that satisfies the lower bound of 90% thermal reliability in the Chebyshev inequality: (a) the temperature profile that really achieves 90% thermal reliability, $T^{90}_{\text{Real}}(\mathbf{r})$; (b) the temperature profile that satisfies the lower bound of 90% thermal reliability in the Chebyshev inequality, $T^{90}_{\text{Chebyshev}}(\mathbf{r})$.

of 90% thermal reliability, that is, $T^{90}_{\text{Chebyshev}}(\mathbf{r})$ needs to be $\mu_T(\mathbf{r}) + 3\sigma_T(\mathbf{r})$ to ensure $\mathbf{Prob}(T(\mathbf{r}) \leq T^{90}_{\text{Chebyshev}}(\mathbf{r})) \geq 0.9$. Here, $T(\mathbf{r})$ is the on-chip temperature distribution.

Since the Chebyshev inequality does not always get a tight lower bound for any type of random variables, there will be a gap between the temperature profile, $T^{90}_{\text{Real}}(\mathbf{r})$, that really achieves 90% thermal reliability (i.e., $\mathbf{Prob}(T(\mathbf{r}) \leq T^{90}_{\text{Real}}(\mathbf{r})) = 0.9$), and $T^{90}_{\text{Chebyshev}}(\mathbf{r})$.[2] As shown in Figure 1, the gap between $T^{90}_{\text{Real}}(\mathbf{r})$ and $T^{90}_{\text{Chebyshev}}(\mathbf{r})$ can achieve about 10°C in our experimental results. Therefore, using the temperature profile of the Chebyshev bound might be an immoderately conservative constraint for thermal reliability. This undesirable phenomenon can result in immoderate guard-banding for circuit design.

On the other hand, from the aspect of circuit design, the specifications of circuit performance and the timing constrains of primary I/Os are usually specified in the system-level design stage. Moreover, in the timing and thermal cooptimization process, timing issues usually take higher priority than thermal issues. Designers try to minimize the circuit delay and meet the temperature requirement. Therefore, in the presence of process variations, to identify possible hotspot regions of a chip, the *thermal yield profile*, $T\,yield(\mathbf{r}, T_{\text{spec}}(\mathbf{r}))$, can be defined as the probability profile of the on-chip temperature at an arbitrary position $\mathbf{r}$ being at or less than a specified temperature $T_{\text{spec}}(\mathbf{r})$.

The thermal yield profile can identify the hotspot regions and can also quantify the probability of a region that could be a hotspot. Therefore, it is a suitable figure of merit for the thermal-related cost in timing-thermal cooptimization physical design stages.

2.2. Leakage Power Modeling

Leakage currents of a gate not only depend on physical device parameters and operating temperatures but also on its input patterns [Chang and Sapatnekar 2007]. To build leakage power models, different input patterns, physical parameters, and operating temperatures are set for each gate in the cell library, and HSPICE simulation is

[2]For example, given a standard normal random variable x with $\mathbf{Prob}(x \leq 1.28\sigma_x) = 0.9$, however, the Chebyshev inequality requires a larger reference value to obtain the same probability as the lower bound, i.e., $\mathbf{Prob}(x \leq 3\sigma_x) \geq 0.9$.

Table I. Accuracy Comparison of I_g and I_s with HSPICE Simulation Results for an NAND Gate

$f_g(L, t_{ox}, T)$		maximum error	average error	error > 3%
Without temperature	t_{ox}, L, t_{ox}^2, L^2 [Chang and Sapatnekar 2007; Shen et al. 2010b]	6.48%	2.70%	4.37%
With temperature	L, t_{ox}, T	3.20%	0.97%	0.35%
	†L, t_{ox}, T, t_{ox}^2	1.55%	0.29%	0.00%
$f_s(L, t_{ox}, T)$		maximum error	average error	error > 3%
Without temperature	$L, t_{ox}, t_{ox}^2, t_{ox}^{-1}$ [Chang and Sapatnekar 2007]	347.32%	70.65%	98.27%
	$L, t_{ox}, Lt_{ox}, L^2, t_{ox}^2, t_{ox}^{-1}, Lt_{ox}^{-1}, L^{-1}t_{ox}$ [Shen et al. 2010b, 2010a]	314.13%	70.52%	100.00%
With temperature	L, T, t_{ox} [Yu et al. 2009]	32.23%	8.73%	76.62%
	(L, t_{ox}, T) are fully expanded to 2nd order $\Longrightarrow$ $L, t_{ox}, T, Lt_{ox}, t_{ox}T, TL, L^2, t_{ox}^2, T^2$	10.31%	1.53%	8.47%
	†(L, t_{ox}, T) are fully expanded to 3rd order $\Longrightarrow$ $L, t_{ox}, T, Lt_{ox}, t_{ox}T, TL, L^2, t_{ox}^2, T^2, Lt_{ox}T, L^2t_{ox}, t_{ox}^2T, T^2L, L^3, t_{ox}^3, T^3$	1.31%	0.19%	0.00%

† The adoptive forms of f_g and f_s in this work.
Note: The second column shows the fitting components of f_g and f_s adopted by existing and proposed models.

performed with the industry design kit to generate data of the leakage currents. After that, average leakage currents of the input patterns are fitted by the least squares method. With leakage currents exponentially relating with physical parameters and operating temperatures and using the least squares fitting method, two major leakage currents—gate tunneling leakage current (I_g) and subthreshold leakage current (I_s)—for each gate type can be fitted [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b; Yu et al. 2009].

$$I_g(L, t_{ox}, T) = a_0 \exp(f_g(L, t_{ox}, T)), \quad (1)$$

$$I_s(L, t_{ox}, T) = b_0 \exp(f_s(L, t_{ox}, T)). \quad (2)$$

Here, a_0 and b_0 are fitting constants, L is the channel length, t_{ox} is the oxide thickness, T is the operating temperature, and $f_g(\cdot)$ and $f_s(\cdot)$ are specific fitting forms.[3]

Basically, I_g occurs in both on and off states, and I_s is the off-state leakage mechanism. Therefore, the leakage power of a gate can be represented as follows [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b; Yu et al. 2009].

$$P_{leak}(L, t_{ox}, T) = V_{dd}I_g + (1 - \mathrm{Sw})V_{dd}I_s, \quad (3)$$

where V_{dd} is the supply voltage and Sw is the switching activity.

To adopt suitable leakage power models, different cell-based leakage current models are investigated [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b; Yu et al. 2009]. To examine their accuracy, we have implemented their proposed models and compared their results with that of HSPICE simulation under TSMC 65nm design kit. Since leakage currents are temperature-dependent, simulation results show that the ignorance of temperature effect in the models [Chang and Sapatnekar 2007; Shen et al. 2010b, 2010a] leads to considerable errors. As shown in the second row of Table I, the

[3]Variations of the device channel length and oxide thickness are considered in this work, since leakage power is more sensitive to these parameters [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b]. It should be noted that although only these two parameters are considered, the developed framework can be easily extended to include any other process variation types, such as the channel dopant variation.

Table II. Error of Leakage Current Models Proposed by [Jaffari and Anis 2008] for an NAND Gate under 65nm Technology Node

Leakage Current	maximum error	average error	error > 3%
Subthreshold	35.53%	9.82%	79.34%
Gate Tunneling	4.51%	1.07%	6.32%

model adopted by Chang and Sapatnekar [2007] and Shen et al. [2010b] can provide acceptable accuracy for the gate tunneling leakage current because of its insensitivity to operating temperatures. However, since the subthreshold leakage current is sensitive to operating temperatures, as shown in the sixth and seventh rows of Table I, the models [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b] are not adequate for preserving the accuracy.

To simultaneously take temperatures and process variations into account, Yu et al. proposed a first-order exponential model, $b_0 \exp(b_1 L + b_2 t_{ox} + b_3 T)$, for the subthreshold leakage current [2009]. Their model can provide accurate results for the 90nm technology node. However, since the variability of subthreshold leakage current, because of operating temperatures and physical device parameters, will increase for more advanced technologies, considerable errors occur for simulation results under the 65nm technology node, as shown in the eighth row of Table I. To improve the accuracy for modeling leakage currents, the orders of fitting components for $f_g(L, t_{ox}, T)$ and $f_s(L, t_{ox}, T)$ shown in Eqs. (1) and (2) are increased. As shown in the fourth and tenth rows of Table I, compared with some models [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b; Yu et al. 2009], our models can present accurate results for both gate tunneling and subthreshold leakage currents. As demonstrating by the test results, exquisite approaches are still required for modern statistical power analyzers [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b] to refine the estimated result, while the temperature-dependent leakage power model is included.

Besides the cell-based leakage current models [Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b; Yu et al. 2009], Jaffari and Anis proposed a bin (grid)-based leakage power model that also simultaneously contains temperature and process variation effects [2008]. Adopting the bin (grid)-based leakage power model, Haghdad and Anis developed a power yield analysis engine that simultaneously considers temperature and process variation effects [2012]. However, the power dissipation of several bins might be changed because of disturbing macros/cells after each optimization iteration of thermal-aware design engines, such as floorplanners or placers. Therefore, the time-consuming HSPICE simulation and least-squares fitting process need to be reperformed for rebuilding leakage power models of the disturbed bins (grids). This will degrade its efficiency of providing thermal reliability information or thermal-related cost for thermal-aware design engines. We have implemented their leakage power models for examining the accuracy. With the fitting forms adopted in Jaffari and Anis [2008] and Haghdad and Anis [2012], both subthreshold and gate tunneling leakage currents of each gate type in the cell library are fitted as $a_0(1+a_1T+a_2T^2)\exp(a_3L+a_4t_{ox})$. The fitting results are also compared with those of HSPICE simulations under the TSMC 65nm model card, and listed in Table II. The results show that their adopted leakage power models present considerable errors under the 65nm technology node, although adequate accuracy of these models for the 90nm technology node has been reported [Jaffari and Anis 2008]. As shown in Table II, their adopted leakage power models result in the maximum error being 35.53% and the average error being 9.82% for the subthreshold leakage current of an NAND gate under the 65nm technology node.

Demonstrating results, more accurate leakage power models should be adopted in the electrothermal analysis frameworks [Jaffari and Anis 2008; Haghdad and Anis 2012] to refine the estimated results, since the temperature is transferred from power consumption. Although the electrothermal analysis frameworks proposed [Jaffari and

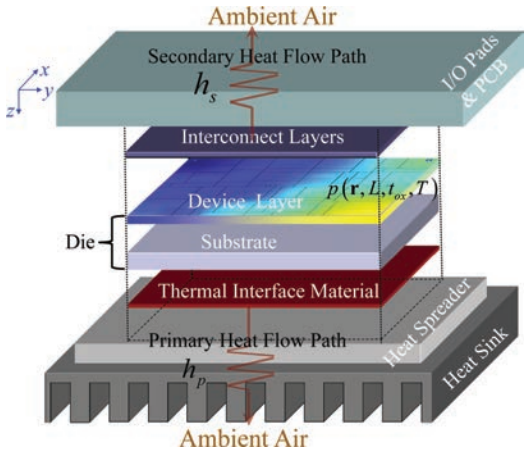

Fig. 2. Compact thermal model of physical design stages.

Anis 2008; Haghdad and Anis 2012] can be very efficient, their baseline temperature calculation frameworks require exquisite extending strategies, because their log-normal temperature approximation algorithm can not be applied to the leakage power models that are not expressed as log-normal random variables, that is, the first-order regression models for $f_g(L, t_{ox}, T)$ and $f_s(L, t_{ox}, T)$ in Eqs. (1) and (2) are not sufficient for the accuracy.

Compared with the framework in Jaffari and Anis [2008] and Haghdad and Anis [2012], our proposed thermal reliability estimator can handle accurate and more complicated leakage power models and present accurate estimated results.

3. PROBLEM FORMULATION AND PHYSICAL PARAMETER MODELING

3.1. Problem Formulation

The compact thermal model for physical design stages is shown in Figure 2 [Wang and Chen 2003; Huang et al. 2006; Yang et al. 2007; Huang and Lee 2009]. It consists of three portions. The primary heat flow path is composed of the thermal interface material, heat spreader, and heat sink. The secondary heat flow path contains interconnect layers, I/O pads, and the print circuit board. Functional blocks of the die are modeled as many power-generating sources attached to a thin layer close to the top surface of the die with the thickness being equal to the junction depth of device [Lallement et al. 2004]. Due to variations of physical parameters, the power consumption of functional blocks is treated statistically. Therefore, the profile of power generating sources, $p(\mathbf{r}, L, t_{ox}, T)$, shown in Figure 2 is modeled as a function of device channel length L, oxide thickness t_{ox}, and on-chip temperature distribution T.

Combining the compact thermal model and the statistical power consumption of functional blocks, the on-chip temperature distribution $T(\mathbf{r}, L, t_{ox})$ can be governed by the statistical steady state heat transfer equation.[4]

$$\nabla \cdot (\kappa(\mathbf{r}, T)\nabla T(\mathbf{r}, L, t_{ox})) = -p(\mathbf{r}, L, t_{ox}, T), \tag{4}$$

[4]Since the time constant of heat conduction is much larger than the clock period of the circuit [Wang and Chen 2003; Skadron et al. 2004], steady-state characteristics of the on-chip temperature distribution are more concerned in thermal-aware physical design engines [Han and Koren 2007; Chakraborty et al. 2008; Tsai et al. 2006]. The scope of this article is to provide a simulation framework for thermal-aware physical design engines, although temporary characteristics of the on-chip temperature distribution are also important for the post-floorplanning or placement real-time task scheduling or workload assignment [Reda et al. 2011].

1 Set T_{nom} and $T_{\text{nom}}^{\text{old}}$ to be the room temperature values;
2 Obtain thermal conductivity by using T_{nom};
3 Obtain P_{nom} by T_{nom}, and nominal values of L and t_{ox}, and set $P_{\text{nom}}^{\text{old}}$ to be P_{nom};
4 $\text{error}_P \leftarrow 1.0$;
5 **While** $\text{error}_P > \varepsilon_P$
6 $\quad \text{error}_T \leftarrow 1.0$;
7 $\quad$ **While** $\text{error}_T > \varepsilon_T$
8 $\quad\quad$ Obtain T_{nom} by the 1-D thermal model [Huang and Lee 2009] with $P_{\text{nom}}^{\text{old}}$;
9 $\quad\quad$ Update thermal conductivity κ by using T_{nom};
10 $\quad\quad \text{error}_T \leftarrow \dfrac{|T_{\text{nom}} - T_{\text{nom}}^{\text{old}}|}{T_{\text{nom}}}$;
11 $\quad\quad T_{\text{nom}}^{\text{old}} \leftarrow T_{\text{nom}}$;
12 $\quad$ **EndWhile**
13 $\quad$ Obtain P_{nom} by T_{nom}, and nominal values of L and t_{ox};
14 $\quad \text{error}_P \leftarrow \dfrac{|P_{\text{nom}} - P_{\text{nom}}^{\text{old}}|}{P_{\text{nom}}}$;
15 $\quad P_{\text{nom}}^{\text{old}} \leftarrow P_{\text{nom}}$;
16 **EndWhile**

Fig. 3. An iterative scheme for computing the appropriate thermal conductivity and approximated average of steady-state nominal temperature values. T_{nom} is the average of on-chip mean temperature values, and P_{nom} is the total nominal power consumption after executing an iteration. With nominal values of the physical device parameters and T_{nom}, P_{nom} can be obtained by summing up the power values of all gates in the design.

subject to the boundary condition

$$\kappa(\mathbf{r}_{b_s}, T)\frac{\partial T(\mathbf{r}_{b_s}, L, t_{ox})}{\partial \vec{n}_{b_s}} + h_{b_s} T(\mathbf{r}_{b_s}, L, t_{ox}) = f_{b_s}(\mathbf{r}_{b_s}). \tag{5}$$

Here, $\mathbf{r} = (x, y, z) \in D$, $D = (0, L_x) \times (0, L_y) \times (-L_z, 0)$ is the domain of die, L_x and L_y are the lateral sizes of die, L_z is the thickness of die, $\kappa(\mathbf{r}, T)$ is the thermal conductivity (W/m $\cdot$ °C) of die, and ∇ is a diverge operator. b_s is any specific boundary surfaces of the die, $\mathbf{r}_{b_s}$ is the position on b_s, h_{b_s} is the heat transfer coefficient on b_s, $f_{b_s}(\mathbf{r}_{b_s})$ is the heat flux function on b_s, and $\partial/\partial\vec{n}_{b_s}$ is the differentiation along the outward direction which is normalized to b_s. $p(\mathbf{r}, L, t_{ox}, T)$ is the power density profile that consists of the deterministic dynamic power density profile $p_d(\mathbf{r})$, the statistical gate tunneling leakage power density profile $p_g(\mathbf{r}, L, t_{ox}, T)$, and the statistical subthreshold leakage power density profile $p_s(\mathbf{r}, L, t_{ox}, T)$. Since the major part of device current flows through the channel, power density distribution has its value only when $\mathbf{r} \in (0, L_x) \times (0, L_y) \times (-j_d, 0)$. Here, j_d is the junction depth of device [Lallement et al. 2004].

Generally, the values of $\kappa(\mathbf{r}, T)$ are temperature dependent. In practice, they can be treated as appropriate constant values while performing temperature-aware physical design procedures [Tsai et al. 2006]. Given nominal values of the physical device parameters, the appropriate thermal conductivity can be computed by using the approximated average of steady-state nominal temperatures calculated by an iterative computation scheme shown in Figure 3.

With the appropriate thermal conductivity, the statistical steady-state heat transfer equation can be rewritten as

$$\kappa\nabla^2 T(\mathbf{r}, L, t_{ox}) = -p(\mathbf{r}, L, t_{ox}, T), \tag{6}$$

subject to the boundary condition

$$\kappa\frac{\partial T(\mathbf{r}_{b_s}, L, t_{ox})}{\partial \vec{n}_{b_s}} + h_{b_s} T(\mathbf{r}_{b_s}, L, t_{ox}) = f_{b_s}(\mathbf{r}_{b_s}), \tag{7}$$

where κ is the thermal conductivity of the die that is obtained by utilizing the procedure presented in Figure 3.

With Eqs. (6) and (7), the goals of this work are to estimate the mean profile, the standard deviation profile, and the thermal yield profile of on-chip temperature distribution.

3.2. Physical Parameter Modeling

Generally, variations of physical parameters can be classified into two categories, die-to-die (D2D) variations and within-die (WID) variations. Due to different stages of the fabrication process, D2D and WID variations can be treated as two independent variation sources. Since D2D variations are smooth within a die, it is reasonable to model all devices having the same D2D variation. On the other hand, WID variations present considerable gradients within a die, and they are spatially correlated because spatial imperfection of the chemical-mechanical polishing and lithography processes. As indicated by the measured results of Cline et al. [2006] and Cheng et al. [2011], distributions of the physical parameters are similar to those of the Gaussian random variables; generally, WID variations are assumed to be a correlated Gaussian random process, and D2D variations are treated as a Gaussian random variable for all devices [Bhardwaj et al. 2008; Chang and Sapatnekar 2007; Shen et al. 2010a, 2010b].

Combining the models of D2D and WID variations, the physical parameter $Par(\mathbf{r}_{xy})$ with its nominal value $\mu_{Par}(\mathbf{r}_{xy})$ at position $\mathbf{r}_{xy} = (x, y) \in (0, L_x) \times (0, L_y)$ can be represented as

$$Par(\mathbf{r}_{xy}) = \mu_{Par}(\mathbf{r}_{xy}) + \delta_{WID}(\mathbf{r}_{xy}) + \delta_{D2D}, \tag{8}$$

where $\delta_{WID}(\mathbf{r}_{xy})$ is a Gaussian random process of WID variations, and δ_{D2D} is a Gaussian random variable of D2D variations.

Since the spatial correlation of $\delta_{WID}(\mathbf{r}_{xy})$ has different decreasing rates in the x-direction and y-direction [Cline et al. 2006], the following spatial covariance function [Bhardwaj et al. 2008] is adopted for modeling the spatial correlation of $\delta_{WID}(\mathbf{r}_{xy})$.[5]

$$C(\mathbf{r}_{x_1y_1}, \mathbf{r}_{x_2y_2}) = \sigma^2 \exp\left(-\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y}\right), \tag{9}$$

where $\mathbf{r}_{x_1y_1} = (x_1, y_1)$ and $\mathbf{r}_{x_1y_2} = (x_2, y_2)$, λ_x and λ_y are correlation lengths of δ_{WID} in the x- and y-directions, respectively, and σ is the standard deviation of $\delta_{WID}(\mathbf{r}_{xy})$.

In this work, the Karhunen-Loève (KL) expansion is utilized to simplify $\delta_{WID}(\mathbf{r}_{xy})$, since its number of transformed random variables is much smaller than that of principal component analysis (PCA) [Bhardwaj et al. 2008]. By applying the KL expansion, $\delta_{WID}(\mathbf{r}_{xy})$ with the spatial covariance function shown in Eq. (9) can be approximated as

$$\delta_{WID}(\mathbf{r}_{xy}) \approx \sum_{l=1}^{N_{Par}} \sqrt{\chi_l}\vartheta_l(\mathbf{r}_{xy})\zeta_l. \tag{10}$$

Here, N_{Par} is the truncation number, each $(\chi_l, \vartheta_l(\mathbf{r}_{xy}))$ is an eigen-pair of $C(\mathbf{r}_{x_1y_1}, \mathbf{r}_{x_2y_2})$, and ζ_l's are independent standard normal random variables.

[5]Although this specific spatial covariance function is adopted, the Karhunen-Loève expansion of a Gaussian random process with an arbitrary spatial covariance function can be efficiently obtained by a finite-element method [Schwab and Todor 2006]. Hence, more advanced spatial covariance functions [Gao et al. 2011; Liu 2007; Cheng et al. 2011] can also be incorporated into our framework.

The closed-form expressions of an eigen-pair $(\chi_l, \vartheta_l(\mathbf{r}_{xy}))$ for $C(\mathbf{r}_{x_1y_1}, \mathbf{r}_{x_2y_2})$ shown in Eq. (9) can be derived as follows [Zhang and Lu 2004].

$$\chi_l = \frac{4\sigma^2\lambda_x\lambda_y}{\left(\lambda_x^2\nu_{x,i}^2+1\right)\left(\lambda_y^2\nu_{y,j}^2+1\right)}, \tag{11}$$

$$\vartheta_l(\mathbf{r}_{xy}) = \vartheta_{x,i}(x)\vartheta_{y,j}(y), \tag{12}$$

where l, i, and j are indices, and the mapping between (i, j) and l is one to one.

$\vartheta_{x,i}(x)$ and $\vartheta_{y,j}(y)$ are equal to

$$\vartheta_{x,i}(x) = \frac{\lambda_x\nu_{x,i}\cos(\nu_{x,i}x)+\sin(\nu_{x,i}x)}{\sqrt{\left(\lambda_x^2\nu_{x,i}^2+1\right)L_x/2+\lambda_x}}, \tag{13}$$

$$\vartheta_{y,j}(y) = \frac{\lambda_y\nu_{y,j}\cos(\nu_{y,j}y)+\sin(\nu_{y,j}y)}{\sqrt{\left(\lambda_y^2\nu_{y,j}^2+1\right)L_y/2+\lambda_y}}. \tag{14}$$

Here, $\nu_{x,i}$ and $\nu_{y,j}$ are positive values that satisfy

$$(\lambda^2\nu^2-1)\sin(\nu\gamma) = 2\lambda\nu\cos(\nu\gamma), \tag{15}$$

with $(\nu = \nu_{x,i}, \gamma = L_x, \lambda = \lambda_x)$ and $(\nu = \nu_{y,j}, \gamma = L_y, \lambda = \lambda_y)$, respectively.

To get reasonable truncation numbers of KL expansions for the physical parameters L and t_{ox}, in this work, N_{Par} for $Par = L$ or $Par = t_{ox}$ is decided by the following criterion.

$$\frac{\chi_{N_{Par}+1}}{\sum_{i=1}^{N_{Par}+1}\chi_i} \le \varepsilon, \tag{16}$$

with $\varepsilon = 1\%$.

Generally, devices located adjacently have similar physical characteristics [Chang and Sapatnekar 2007; Shen et al. 2010b, 2010a]. Therefore, the active layer is partitioned into several rectangular grids for modeling physical parameters. After that, with the KL expansion, the device channel length $L_{\underline{m}}$ and oxide thickness $t_{ox\underline{m}}$ in the mth modeling grid can be approximated as

$$L_{\underline{m}} = \mu_{L_{\underline{m}}} + \boldsymbol{g}_{L_{\underline{m}}}^T\boldsymbol{\eta}_L, \tag{17}$$

$$t_{ox\underline{m}} = \mu_{t_{ox\underline{m}}} + \boldsymbol{g}_{t_{ox\underline{m}}}^T\boldsymbol{\eta}_{t_{ox}}. \tag{18}$$

Here, $\mu_{L_{\underline{m}}}$ and $\mu_{t_{ox\underline{m}}}$ are nominal values of $L_{\underline{m}}$ and $t_{ox\underline{m}}$, respectively. $\boldsymbol{g}_{L_{\underline{m}}}$ and $\boldsymbol{g}_{t_{ox\underline{m}}}$ are coefficient vectors for $\boldsymbol{\eta}_L$ and $\boldsymbol{\eta}_{t_{ox}}$, respectively. $\boldsymbol{\eta}_L = [\eta_{L_1}, \ldots, \eta_{L_{N_L}}]^T$ and $\boldsymbol{\eta}_{t_{ox}} = [\eta_{t_{ox1}}, \ldots, \eta_{t_{ox}N_{t_{ox}}}]^T$ are standard normal random vectors constituted by the KL expanded WID and D2D random variables for representing the device channel length and the oxide thickness in all modeling grids, respectively.

In the rest of this article, $\boldsymbol{\xi}$ is employed to represent $[\eta_{L_1}, \ldots, \eta_{L_{N_L}}, \eta_{t_{ox1}}, \ldots, \eta_{t_{ox}N_{t_{ox}}}]^T$ for the sake of notation simplicity.

4. STATISTICAL ELECTRO-THERMAL SIMULATOR

The executing flow of the proposed statistical electrothermal simulator is summarized in Figure 4. Given the information of physical parameters, the KL expansion is performed to transform the spatial correlated physical parameters into a set of uncorrelated random variables. Then, the statistical polynomial expression of the on-chip temperature distribution is generated by the developed stochastic collocation-based

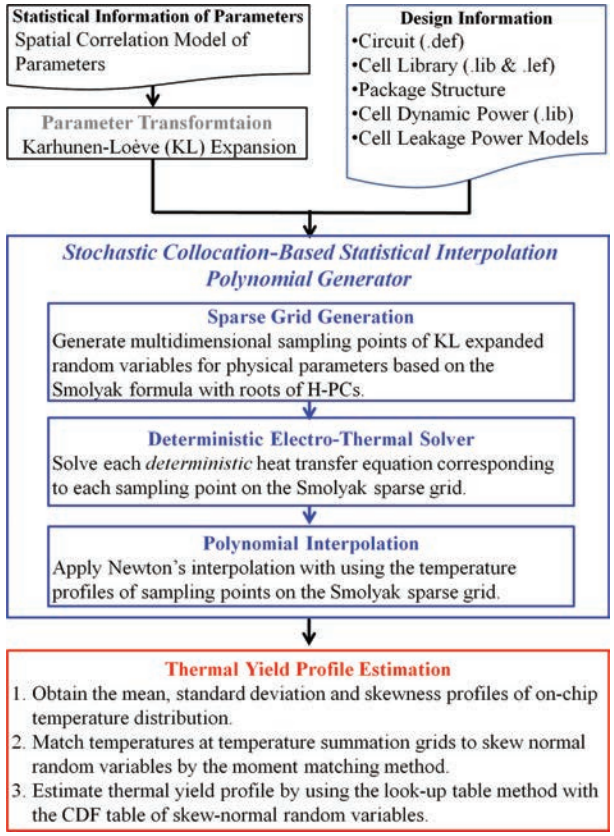

Fig. 4. Flow of the developed statistical electrothermal simulator.

statistical interpolation polynomial generator. After that, the on-chip thermal yield profile is estimated by the developed thermal yield profile estimation engine.

The stochastic collocation-based statistical interpolation polynomial generator, the thermal yield profile estimation engine, and a mixed-mesh strategy for enhancing the statistical electrothermal simulator will be described in the following three sections.

4.1. Stochastic Collocation-Based Statistical Interpolation Polynomial Generator

The generator takes three steps to construct the statistical interpolation polynomial of the on-chip temperature distribution. First, the multidimensional sampling points of KL expanded random variables are generated by using the Smolyak sparse grid formula with sampling points being the roots of Hermite polynomials (HPs). Then, for each sampling point of the physical parameters, its corresponding temperature profile can be obtained by solving the corresponding deterministic heat transfer equation. Finally, the approximated expression of on-chip temperature distribution is built by utilizing Newton's interpolation polynomial formula. The details are presented in the rest of this section.

4.1.1. Smolyak Sparse Grid Generation. The primary advantage of Smolyak sparse grid formulation is to construct an interpolating polynomial for approximating a multivariate function $u \in C^r$ by using much fewer samples of the desired function than those of the full tensor product interpolation formula and the MC method but still to

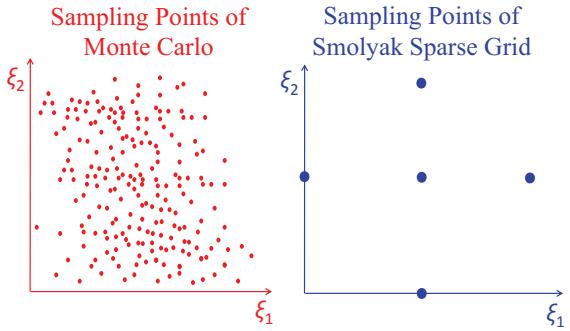

Fig. 5. The number of sampling random variables comparison between the Monte Carlo method and the Smolyak sparse grid formulation. Here, the samples of Smolyak sparse grid are adopted for achieving a level-two approximation.

maintain an acceptable error bound [Smolyak 1963; Barthelmann et al. 2000]. Here, C^r is the set of all functions that have continuous derivatives of all orders up to r. With this stochastic collocation technique, the statistical interpolation polynomial of on-chip temperature distribution can be efficiently constructed.

The MC method randomly generates samples of the random variables and hence requires a large number of samples for achieving an accurate estimate. In contrast, the Smolyak sparse grid technique uses the roots of HPs or the extrema of the Chebyshev polynomial [Barthelmann et al. 2000] to generate samples of the random variables and employs these fewer samples to effectively interpolate the desired solution. For example, Figure 5 illustrates that the number of possible sample points of the MC method is much larger than that of the Smolyak sparse grid formulation for a two-dimensional random variable.

According to the Smolyak sparse grid formulation [Smolyak 1963], on-chip temperature distribution can be explicitly approximated as follows [Barthelmann et al. 2000].

$$\widehat{T}_q^{N_{KL}}(\mathbf{r},\boldsymbol{\xi}) = \sum_{q-N_{KL}+1\le|\mathbf{i}|\le q} (-1)^{q-|\mathbf{i}|}\binom{N_{KL}-1}{q-|\mathbf{i}|}\left(\mathcal{Q}^{i_1}(T)\otimes\cdots\otimes\mathcal{Q}^{i_n}(T)\otimes\cdots\otimes\mathcal{Q}^{i_{N_{KL}}}(T)\right). \quad (19)$$

Here, $N_{KL} = N_{t_{ox}} + N_L$ is the number of random variables in $\boldsymbol{\xi}$, $q = N_{KL} + l$, $l \ge 1$ is the formulation level, and $|\mathbf{i}| = \sum_{n=1}^{N_{KL}} i_n$, with each $i_n \ge 1$. $\mathcal{Q}^{i_n}(T)$ is an interpolating polynomial of $T(\mathbf{r},\boldsymbol{\xi})$ by only utilizing a single random variable ξ_n, i_n is the index to decide the sample number (m_{i_n}) for $\mathcal{Q}^{i_n}(T)$, and $\otimes$ is the cross product operator for functions. As suggested by Barthelmann et al. [2000], $m_{i_n=1}$ is set to 1, and m_{i_n} is equal to $2^{i_n-1}+1$ for $i_n > 1$. From Eq. (19), only the corresponding temperature values of a small set of samples for $\boldsymbol{\xi}$ need to be known [Barthelmann et al. 2000]. This set is called the sparse grid and can be represented as

$$\mathcal{H}(q, N_{KL}) = \bigcup_{q-N_{KL}+1\le|\mathbf{i}|\le q} \left(\hbar^{i_1}\times\cdots\times\hbar^{i_n}\times\cdots\times\hbar^{i_{N_{KL}}}\right), \quad (20)$$

where $\hbar^{i_n} = \{\xi_n^1,\ldots,\xi_n^{m_{i_n}}\}$ is the set of sample points used by $\mathcal{Q}^{i_n}(T)$, and $\times$ is the cross product operator for sets.

The number of sample points from the Smolyak sparse grid formulation is in the order of $O\left(N_{KL}^l/l!\right)$, and the runtime complexity for obtaining $\widehat{T}_q^{N_{KL}}(\mathbf{r},\boldsymbol{\xi})$ is in the order of $t_{\text{det}}\cdot O(N_{KL}^l/l!)$. Here, t_{det} is the runtime complexity for performing the deterministic

electrothermal simulation once. The Smolyak sparse grid formulation ensures a error bound, $\frac{c_{N_{KL},r}\cdot(\log N_{\mathcal{H}})^{(r+1)(N_{KL}-1)}}{N_{\mathcal{H}}^{r}}$, for the function having bounded derivatives up to order r [Barthelmann et al. 2000]. Here, $N_{\mathcal{H}}$ is the number of sample points in $\mathcal{H}(q, N_{KL})$, and $c_{N_{KL},r}$ is a constant that depends on N_{KL} and r. According to our experience, the accurate estimation of the thermal yield profile can be obtained by setting level l to be 1. The number of sample points for the Smolyak sparse grid formulation can be much less than that of the MC method. A simple example is presented in Appendix A to illustrate the Smolyak sparse grid formulation.

The sampling values of $\hbar^{i_n}$ for each i_n must be properly decided. Adopting the roots of H-PCs with its order corresponding to each i_n can achieve the most accurate result, as $\boldsymbol{\xi}$ is a normal random vector [Phillips 2003]. Choosing the extrema of the Chebyshev polynomial with its order corresponding to i_n can achieve the nested sparse grid structure, that is, $\hbar^{i_n=j} \subset \hbar^{i_n=k}$ as $j < k$, and the acceptable accuracy [Barthelmann et al. 2000]. In this work, we select the roots of H-PCs as the sampling values, since the result is shown to be very accurate by using the low-level approximation, and the nested sparse grid structure is still preserved for $q = N_{KL} + 1$.[6]

4.1.2. Deterministic Electrothermal Solver: Temperature Profile Calculation for a Given Sample Point. After constructing the sparse grid $\mathcal{H}(q, N_{KL})$ of $\boldsymbol{\xi}$, the samples of channel length and oxide thickness in the mth modeling grid corresponding to the jth sampling vector $\boldsymbol{\xi}^j$ in $\mathcal{H}(q, N_{KL})$ can be calculated by Eqs. (17) and (18), respectively, and the deterministic power density profile corresponding to $\boldsymbol{\xi}^j$ can also be obtained. Hence, we have the deterministic steady-state heat transfer equation as

$$\kappa \nabla^2 T(\mathbf{r}, \boldsymbol{\xi}^j) = -p(\mathbf{r}, \boldsymbol{\xi}^j, T), \tag{21}$$

subject to the boundary condition

$$\kappa \frac{\partial T(\mathbf{r}_{b_s}, \boldsymbol{\xi}^j)}{\partial \vec{n}_{b_s}} + h_{b_s} T(\mathbf{r}_{b_s}, \boldsymbol{\xi}^j) = f_{b_s}(\mathbf{r}_{b_s}). \tag{22}$$

Here, $T(\mathbf{r}, \boldsymbol{\xi}^j)$ and $p(\mathbf{r}, \boldsymbol{\xi}^j, T)$ are the deterministic temperature and power density profiles with the sampling point $\boldsymbol{\xi}^j$, respectively. Since the power density profile in Eq. (21) is temperature dependent, a deterministic electrothermal solver is summarized in Figure 6 and built to obtain each $T(\mathbf{r}, \boldsymbol{\xi}^j)$.

The implementation of the developed deterministic electrothermal solver is illustrated in Figure 7. The accumulated area of each gate type in each simulating temperature grid can be precalculated and stored in the precalculation stage. With this precalculated data, the deterministic power density profile for each sampling point in $\mathcal{H}(q, N_{KL})$ can be obtained and updated in the order of $O(N_x N_y N_{type})$ during the electrothermal simulation loop. Here, N_x and N_y are the division numbers of the simulation grid along $x-$ and $y-$directions, respectively, and N_{type} is the number of gate types for the given design. Generally, N_{type} is determined by the specific cell library, and it is far less than the number of simulation grids, $N_x N_y$. After obtaining or updating the deterministic power density profile, an efficient deterministic thermal simulator [Huang and Lee 2009] is adopted to solve the deterministic heat transfer equation. These updating and solving procedures are repeated until the result is converged.

[6]If high-order approximation is needed for the accuracy, we suggest using the extrema of the Chebyshev polynomial because the nested sparse grid structure is always preserved. Hence, the number of sample points can be smaller.

Algorithm: Temperature Profile Calculation for a Sample Point
Input: A sampling point $\boldsymbol{\xi}^j$, initial temperature $T^{ini}_{\boldsymbol{\xi}^j}$ and $p_d(\mathbf{r})$
Output: Temperature profile $T(\mathbf{r}, \boldsymbol{\xi}^j)$

1 **Begin**
2 $T(\mathbf{r}, \boldsymbol{\xi}^j) \leftarrow T^{ini}_{\boldsymbol{\xi}^j}$;
3 $MaxError \leftarrow \infty$
4 Obtain $t_{oxm}(\boldsymbol{\xi}^j)$ and $L_m(\boldsymbol{\xi}^j)$ for each mth modeling grid according to $\boldsymbol{\xi}^j$;
5 **While** $(MaxError > \epsilon)$
6 $T_{pre}(\mathbf{r}, \boldsymbol{\xi}^j) \leftarrow T(\mathbf{r}, \boldsymbol{\xi}^j)$;
7 Update $p_{leak}(\mathbf{r}, \boldsymbol{\xi}^j, T_{pre})$ by $T_{pre}(\mathbf{r}, \boldsymbol{\xi}^j)$;
8 $p(\mathbf{r}, \boldsymbol{\xi}^j, T_{pre}) \leftarrow p_{leak}(\mathbf{r}, \boldsymbol{\xi}^j, T_{pre}) + p_d(\mathbf{r})$;
9 †Solve (21) and (22) with $p(\mathbf{r}, \boldsymbol{\xi}^j, T_{pre})$ to obtain a new $T(\mathbf{r}, \boldsymbol{\xi}^j)$;
10 **if** $(T(\mathbf{r}, \boldsymbol{\xi}^j) = \infty)$ **then** Thermal runaway;
11 $MaxError \leftarrow \max_{\mathbf{r}} \left|T(\mathbf{r}, \boldsymbol{\xi}^j) - T_{pre}(\mathbf{r}, \boldsymbol{\xi}^j)\right|$;
12 **EndWhile**
13 **End**

†Any deterministic thermal simulators can be used to execute *Line* 9.
Here, the simulator developed in [Huang and Lee 2009] is adopted.

Fig. 6. Deterministic electrothermal solver for each sampling point ($\boldsymbol{\xi}^j$) in the sparse grid. p_{leak}, p_d, and p are the leakage, dynamic, and total power density profiles for each sampling point, respectively.

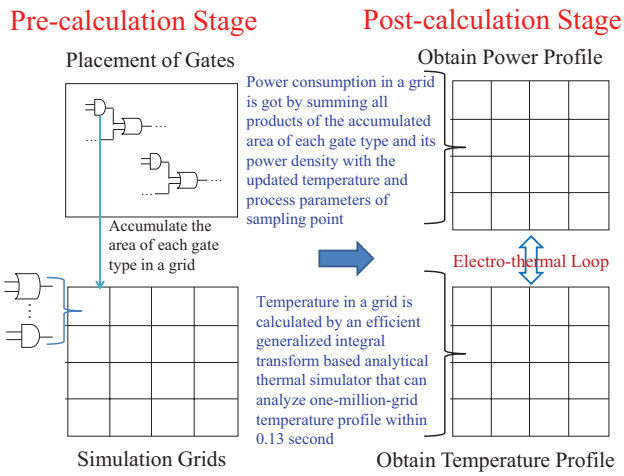

Fig. 7. Implementation of solving the deterministic heat transfer equation.

4.1.3. Temperature Profile Construction by Using Polynomial Interpolation. With each obtained $T(\mathbf{r}, \boldsymbol{\xi}^j)$, the polynomial interpolation technique can be applied to construct the interpolating polynomial for the statistical temperature. As suggested by Barthelmann et al. [2000], the Lagrange polynomial can be applied to construct the interpolating polynomial $\mathcal{Q}^{i_1}(T) \otimes \cdots \otimes \mathcal{Q}^{i_{N_{KL}}}(T)$ for each different $|\mathbf{i}|$. However, the suggested interpolation method requires performing the cross-product operation of functions; this can be slightly complicated for the implementation. Therefore, we adopt Newton's interpolating method to globally interpolate $T(\mathbf{r}, \boldsymbol{\xi})$, because it can be implemented more easily and can interpolate the same polynomial as Barthelmann's method [Phillips 2003]. Therefore, the Smolyak's error bound can still be preserved.

Algorithm: Stochastic Collocation-Based Statistical Interpolation Polynomial Generation
Input: Design information such as .def, .lef, and .lib files;
Geometry of the die and Package structure;
Leakage power models;
Spatial correlation models of the device channel length and oxide thickness.
Output: The statistical interpolation polynomial, $\widehat{T}(\mathbf{r}, \boldsymbol{\xi})$, of the on-chip temperature distribution.

Begin
Set thermal parameters and the initial average mean temperature, μ_T^{ini}, of the die by 1-D thermal model;
For $m \leftarrow 1$ to N_g
  Obtain $\boldsymbol{g}_{L_{\underline{m}}}$ of $L_{\underline{m}}$ and $\boldsymbol{g}_{t_{ox\underline{m}}}$ of $t_{ox\underline{m}}$ by the KL expansion;
EndFor
Generate $\mathcal{H}(q, N_{KL})$ for the KL expanded random variables.
For $n \leftarrow 0$ to $N_{\mathcal{H}} - 1$
  Obtain $T(\mathbf{r}, \boldsymbol{\xi}^n)$ by using the algorithm presented in Fig. 6.
EndFor
Solve (25) to obtain Newton's interpolation formula in (23).
End

Fig. 8. Stochastic collocation-based statistical interpolation polynomial generation algorithm.

With Newton's interpolation formula, the on-chip temperature at an arbitrary position $\mathbf{r}^*$ of the die can be approximated as

$$\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \sum_{j=0}^{N_{\mathcal{H}}-1} \hat{u}_j(\mathbf{r}^*)\phi_j(\boldsymbol{\xi}). \tag{23}$$

Here, each $\phi_j(\boldsymbol{\xi})$ is a fundamental polynomial with respect to the jth sampling vector $\boldsymbol{\xi}^j$, and the form of each $\phi_j(\boldsymbol{\xi})$ can be found in Phillips [2003]. $N_{\mathcal{H}}$ is the number of sampling vectors in the sparse grid $\mathcal{H}(q, N_{KL})$. Each $\hat{u}_j(\mathbf{r}^*)$ is an unknown coefficient that needs to be determined.

Based on the basic idea of interpolation that the approximated function must match each known data, the interpolated polynomial in Eq. (23) satisfies the following equation for each $\boldsymbol{\xi}^n$.

$$\sum_{j=0}^{N_{\mathcal{H}}-1} \hat{u}_j(\mathbf{r}^*)\phi_j(\boldsymbol{\xi}^n) = T(\mathbf{r}^*, \boldsymbol{\xi}^n). \tag{24}$$

With the property of fundamental polynomial described in [Phillips 2003], Eq. (24) can be rewritten as the matrix form for finding each $\hat{u}_j(\mathbf{r}^*)$ at position $\mathbf{r}^*$.

$$\begin{bmatrix} \phi_0(\boldsymbol{\xi}^0) & 0 & \cdots & 0 \\ \phi_0(\boldsymbol{\xi}^1) & \phi_1(\boldsymbol{\xi}^1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{\xi}^{N_{\mathcal{H}}-1}) & \phi_1(\boldsymbol{\xi}^{N_{\mathcal{H}}-1}) & \cdots & \phi_{N_{\mathcal{H}}-1}(\boldsymbol{\xi}^{N_{\mathcal{H}}-1}) \end{bmatrix} \begin{bmatrix} \hat{u}_0(\mathbf{r}^*) \\ \hat{u}_1(\mathbf{r}^*) \\ \vdots \\ \hat{u}_{N_{\mathcal{H}}-1}(\mathbf{r}^*) \end{bmatrix} = \begin{bmatrix} T(\mathbf{r}^*, \boldsymbol{\xi}^{\mathbf{0}}) \\ T(\mathbf{r}^*, \boldsymbol{\xi}^1) \\ \vdots \\ T(\mathbf{r}^*, \boldsymbol{\xi}^{N_{\mathcal{H}}-1}) \end{bmatrix} \tag{25}$$

Each $\hat{u}_j(\mathbf{r}^*)$ can be calculated by using the forward substitution.

The algorithm for generating the statistical interpolation polynomial of on-chip temperature distribution is shown in Figure 8.

4.2. Thermal Yield Profile Estimation Engine

With the generated statistical interpolation polynomial of on-chip temperature distribution, the mean, standard deviation, and skewness profiles of on-chip temperature

distribution are computed. After that, the temperature at each arbitrary position is approximated to be a corresponding skew normal random variable by the moment matching technique. Finally, the on-chip thermal yield profile is estimated by looking up the cumulative distribution function (CDF) table of those corresponding skew normal random variables. The detailed description of this thermal yield profile estimation engine is shown next.

As mentioned in Section 2.1, the on-chip thermal yield profile at an arbitrary position $\mathbf{r}^*$ of the die can be defined as

$$Tyield(\mathbf{r}^*, T_{\text{spec}}(\mathbf{r}^*)) \stackrel{\text{def}}{=} \mathbf{Prob}(T(\mathbf{r}^*, \boldsymbol{\xi}) \leq T_{\text{spec}}(\mathbf{r}^*)). \tag{26}$$

With the definition given in Eq. (26), our target is to approximate the CDF of $T(\mathbf{r}^*, \boldsymbol{\xi})$. To obtain the approximated expression of $T(\mathbf{r}^*, \boldsymbol{\xi})$, the formulation level l is set as 1 for generating the sparse grid $\mathcal{H}(q, N_{KL})$ with $q = N_{KL} + l$ in Eq. (20). Then, applying Newton's interpolating method, the approximated expression of $T(\mathbf{r}^*, \boldsymbol{\xi})$, shown in Eq. (23), can be rewritten as[7]

$$\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \sum_{k=1}^{N_{KL}} \left(\hat{a}_k(\mathbf{r}^*) \xi_k^2 + \hat{b}_k(\mathbf{r}^*) \xi_k \right) + \hat{c}(\mathbf{r}^*), \tag{27}$$

where $\hat{a}_k(\mathbf{r}^*)$, $\hat{b}_k(\mathbf{r}^*)$, and $\hat{c}(\mathbf{r}^*)$ are the coefficients and can be obtained by performing the algorithm shown in Figure 8.

After several manipulations, Eq. (27) can be rewritten as

$$\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \sum_{k=1}^{N_{KL}} \hat{a}_k(\mathbf{r}^*) \chi_k(\mathbf{r}^*, \xi_k) + \tilde{c}(\mathbf{r}^*). \tag{28}$$

Here, each $\chi_k(\mathbf{r}^*, \xi_k) = (\xi_k + \frac{\hat{b}_k(\mathbf{r}^*)}{2\hat{a}_k(\mathbf{r}^*)})^2$ is a non-central chi-square random variable, because ξ_k is a normal random variable, $\tilde{c}(\mathbf{r}^*) = \hat{c}(\mathbf{r}^*) - \sum_{k=1}^{N_{KL}} \frac{\hat{b}_k^2(\mathbf{r}^*)}{4\hat{a}_k(\mathbf{r}^*)}$ is a constant, and $\chi_k(\mathbf{r}^*, \xi_k)$'s are independent because ξ_k's are independent. Therefore, $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ is a weighted sum of independent non-central chi-square random variables.

The estimation of Eq. (26) can be done by calculating the CDF of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ represented in Eq. (28). Since $\boldsymbol{\xi}$ is an independent normal random vector, theoretically, the PDF of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ could be obtained by convolving the PDFs of $\chi_k(\mathbf{r}^*, \xi_k)$'s. However, it is not practical because of numerous numerical convolutions. The moment matching-based CDF estimation techniques are another choice for efficiently approximating the CDF of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$. APEX [Li et al. 2004], a state-of-the-art method, approximates the CDF of a random variable with the similar form of Eq. (27) by linearly combining exponential waveforms and can achieve an arbitrarily required matching order of statistical moments. Padé approximation is essential during performing APEX, although it cannot guarantee being stable for obtaining poles/zeros, even in the low-order approximation. To remedy this unstable issue, the technique proposed by Tutuianu et al. [1996] can be adopted to obtain the first two dominated pole/zero pairs for APEX. However, the first two dominated pole/zero pairs only can construct an approximated CDF of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ that

[7]To get a more accurate approximated expression of $T(\mathbf{r}^*, \boldsymbol{\xi})$, one can set the formulation level l as 2 to capture the cross-product terms of ξ_k's in the second-order polynomial approximation. As shown in Appendix C, with cross-product terms of ξ_k's, the second-order polynomial approximated expression of $T(\mathbf{r}^*, \boldsymbol{\xi})$ can be transformed to the form that is consistent with Eq. (27). Therefore, the proposed thermal yield profile estimation engine can be extended to the second-order polynomial approximation that has cross-product terms of ξ_k's.

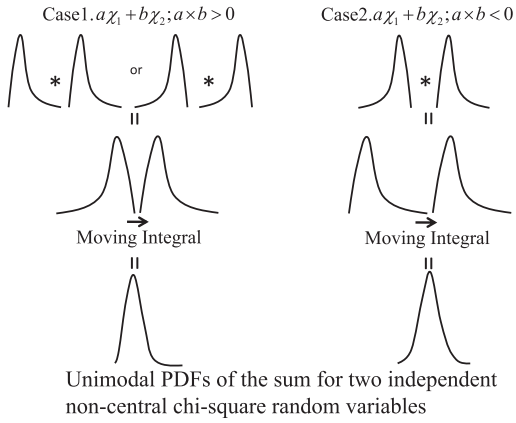

Fig. 9. Sketch of PDF for the weighted sum of two independent non-central chi-square random variables. Case 1: the convolution result of two right-skewed distributions or two left-skewed distributions. Case 2: the convolution result of one left-skewed distribution and one right-skewed distribution.

matches up to the first two statistical moments. Refer to Li et al. [2004] and Tutuianu et al. [1996] for the details of APEX and the stable two-pole technique, respectively.

Here, we are going to develop a moment matching-based technique to match the statistical moments of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ up to the third order for approximating the CDF of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$. The basic idea of this approach is to approximate a random variable with a unimodal and skewed PDF by matching its mean, variance, and skewness to be a skew-normal random variable. To explain the unimodal PDF property of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$, the sketches of PDFs corresponding to two different cases for the weighted sum of two independent non-central chi-square random variables are shown in Figure 9. Case 1 shows the convolution result of two right-skewed distributions or two left-skewed distributions, and Case 2 presents the convolution result of one left-skewed distribution and one right-skewed distribution. Although, depending on the leading coefficients, the skewness of resulting random variables might increase or decrease, both resulting random variables have unimodal PDFs. Since $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ is the weighted sum of independent non-central chi-square random variables, its PDF can be obtained by performing the convolution of two random variables successively. Therefore, it still has a unimodal PDF.

As indicated in Azzalini [2005], the skew-normal random variable is suitable for approximating the random variable with unimodal and skewed PDF. Hence, the skew-normal random variable can be a suitable model for approximating $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$.

From the representation of Eq. (28), the thermal yield at an arbitrary location $\mathbf{r}^*$ can be approximated as

$$\begin{aligned} Tyield\big(\mathbf{r}^*, T_{\text{spec}}(\mathbf{r}^*)\big) &\approx \mathbf{Prob}\big(\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) \le T_{\text{spec}}(\mathbf{r}^*)\big) \\ &= \mathbf{Prob}\big(\Delta\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) \le \rho_{\widehat{T}}(\mathbf{r}^*)\big), \end{aligned} \tag{29}$$

where $\rho_{\widehat{T}}(\mathbf{r}^*) = \frac{T_{\text{spec}}(\mathbf{r}^*) - \mu_{\widehat{T}}(\mathbf{r}^*)}{\sigma_{\widehat{T}}(\mathbf{r}^*)}$, and $\Delta\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \frac{\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) - \mu_{\widehat{T}}(\mathbf{r}^*)}{\sigma_{\widehat{T}}(\mathbf{r}^*)}$. $\mu_{\widehat{T}}(\mathbf{r}^*)$ and $\sigma_{\widehat{T}}(\mathbf{r}^*)$ are the mean and the standard deviation of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$, respectively; they can be computed in the order of $O(N_{KL})$, since $\chi_k(\mathbf{r}^*, \xi_k)$'s are independent.

To approximate $\Delta\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ to be a skew-normal random variable, $Z \sim SN(\upsilon_{\mathbf{r}^*}, \omega_{\mathbf{r}^*}, \alpha_{\mathbf{r}^*})$, the parameters $\upsilon_{\mathbf{r}^*}$, $\omega_{\mathbf{r}^*}$, and $\alpha_{\mathbf{r}^*}$ [Azzalini 2005] need to be calculated. The first three

moments of Z can be formulated as follows with these parameters.

$$\mathrm{E}\{Z\} = \upsilon_{\mathbf{r}^*} + \omega_{\mathbf{r}^*}\delta_{\mathbf{r}^*}, \tag{30}$$

$$\mathrm{Var}\{Z\} = \omega_{\mathbf{r}^*}^2\left(1-\delta_{\mathbf{r}^*}^2\right), \tag{31}$$

$$\mathrm{Skew}\{Z\} = \frac{4-\pi}{2}\frac{\delta_{\mathbf{r}^*}^3}{\sqrt{\left(1-\delta_{\mathbf{r}^*}^2\right)^3}}, \tag{32}$$

where

$$\delta_{\mathbf{r}^*} = \sqrt{\frac{2}{\pi}}\frac{\alpha_{\mathbf{r}^*}}{\sqrt{1+\alpha_{\mathbf{r}^*}^2}}. \tag{33}$$

After matching the first three moments of $\Delta\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ with Eqs. (30)–(32), we have

$$\upsilon_{\mathbf{r}^*} = -\omega_{\mathbf{r}^*}\delta_{\mathbf{r}^*}, \tag{34}$$

$$\omega_{\mathbf{r}^*} = \sqrt{\frac{1}{1-\delta_{\mathbf{r}^*}^2}}, \tag{35}$$

$$\alpha_{\mathbf{r}^*} = \sqrt{\frac{\pi}{2}}\frac{\delta_{\mathbf{r}^*}}{\sqrt{1-\frac{\pi}{2}\delta_{\mathbf{r}^*}^2}}, \tag{36}$$

where

$$\delta_{\mathbf{r}^*} = \sqrt{\frac{\gamma_{\Delta\widehat{T}}^{\frac{2}{3}}(\mathbf{r}^*)}{\left(\frac{4-\pi}{2}\right)^{\frac{2}{3}} + \gamma_{\Delta\widehat{T}}^{\frac{2}{3}}(\mathbf{r}^*)}}. \tag{37}$$

Here, $\gamma_{\Delta\widehat{T}}(\mathbf{r}^*)$ is the skewness of $\Delta\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$, and the sign of $\delta_{\mathbf{r}^*}$ is the same as the sign of $\gamma_{\Delta\widehat{T}}(\mathbf{r}^*)$.[8]

To obtain $\gamma_{\Delta\widehat{T}}(\mathbf{r}^*)$, $\mathrm{E}\{\Delta\widehat{T}^3(\mathbf{r}^*, \boldsymbol{\xi})\}$ is needed and can be calculated as

$$\mathrm{E}\{\Delta\widehat{T}^3(\mathbf{r}^*, \boldsymbol{\xi})\} = \frac{\mathrm{E}\{\widehat{T}^3(\mathbf{r}^*, \boldsymbol{\xi})\} - 3\sigma_{\widehat{T}}^2(\mathbf{r}^*)\mu_{\widehat{T}}(\mathbf{r}^*) - \mu_{\widehat{T}}^3(\mathbf{r}^*)}{\sigma_{\widehat{T}}^3(\mathbf{r}^*)}, \tag{38}$$

where $\mu_{\widehat{T}}(\mathbf{r}^*)$ and $\sigma_{\widehat{T}}(\mathbf{r}^*)$ are the mean and standard deviation of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$, respectively.

As shown in Eq. (38), $\mathrm{E}\{\widehat{T}^3(\mathbf{r}^*, \boldsymbol{\xi})\}$ is needed. However, the computational complexity of obtaining its value is $O(N_{KL}^3)$ if the expression of $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ shown in Eq. (28) is directly used. To reduce the complexity, $\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi})$ can be rewritten as

$$\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \sum_{k=1}^{N_{KL}} \hat{a}_k(\mathbf{r}^*)\hat{\chi}_k(\mathbf{r}^*, \xi_k) + \hat{d}(\mathbf{r}^*), \tag{39}$$

[8]Theoretically, the skew-normal random variable has an upper-bound skewness that can be achieved. In Appendix B, we examine the preceding stability issue, and address a method that can stably achieve the more higher-order approximation if more accurate approximation is required.

where $\hat{\chi}_k(\mathbf{r}^*, \xi_k) = \chi_k(\mathbf{r}^*, \xi_k) - \mu_{\chi_k}(\mathbf{r}^*)$, $\hat{d}(\mathbf{r}^*) = \tilde{c}(\mathbf{r}^*) - \sum_{k=1}^{N_{KL}} \hat{a}_k(\mathbf{r}^*)\mu_{\chi_k}(\mathbf{r}^*)$, and $\mu_{\chi_k}(\mathbf{r}^*) = \mathrm{E}\{\chi_k(\mathbf{r}^*, \xi_k)\}$. Since $\hat{\chi}_k(\mathbf{r}^*, \xi_k)$'s have zero mean and are independent, we have

$$\mathrm{E}\left\{\left(\sum_{k=1}^{N_{KL}} \hat{a}_k(\mathbf{r}^*)\hat{\chi}_k(\mathbf{r}^*, \xi_k)\right)^i\right\} = \sum_{k=1}^{N_{KL}} \hat{a}_k^i(\mathbf{r}^*)\mathrm{E}\left\{\hat{\chi}_k^i(\mathbf{r}^*, \xi_k)\right\}, \tag{40}$$

for $i = 1$, 2, and 3.

Therefore, $\mathrm{E}\{\widehat{T}^3(\mathbf{r}^*, \boldsymbol{\xi})\}$ can be obtained in the order of $O(N_{KL})$, and the computational complexity of evaluating $\mathrm{E}\{\Delta\widehat{T}^3(\mathbf{r}^*, \boldsymbol{\xi})\}$ is also $O(N_{KL})$.

With $\upsilon_{\mathbf{r}^*}$, $\omega_{\mathbf{r}^*}$, and $\alpha_{\mathbf{r}^*}$, $Tyield(\mathbf{r}^*, T_{\mathrm{spec}}(\mathbf{r}^*))$ can be estimated by the CDF of the matched skew-normal random variable. Finally, we have

$$Tyield(\mathbf{r}^*, T_{\mathrm{spec}}(\mathbf{r}^*)) \approx \Phi(\beta_{\mathbf{r}^*}) - 2T_{Owen}(\beta_{\mathbf{r}^*}, \alpha_{\mathbf{r}^*}). \tag{41}$$

Here, $\Phi(\cdot)$ is the CDF of the standard normal random variable, $T_{Owen}(\cdot)$ is Owen's T function [Azzalini 2005], and $\beta_{\mathbf{r}^*} = \frac{\rho_{\widehat{T}}(\mathbf{r}^*) - \upsilon_{\mathbf{r}^*}}{\omega_{\mathbf{r}^*}}$.

With Eqs. (34)–(36) and Eq. (41), $Tyield(\mathbf{r}^*, T_{\mathrm{spec}}(\mathbf{r}^*))$ can be efficiently evaluated by the lookup table method.

4.3. Mixed-Mesh Strategy for Enhancing the Statistical Electrothermal Simulator

As stated in Section 4.1.3, the deterministic electrothermal solver presented in Figure 6 needs to be executed $N_{\mathcal{H}}$ times to generate the statistical interpolation polynomial of on-chip temperature distribution shown in Eq. (23) with the level-1 Smolyak sparse grid formula. Hence, although the developed thermal yield profile estimation engine can be done efficiently, the total runtime for obtaining the thermal yield profile is still dominated by the statistical interpolation polynomial generation. Here, we will present a mixed-mesh strategy to speed up the statistical interpolation polynomial generator without sacrificing the accuracy of the estimated thermal yield profile.

The developed mixed-mesh strategy is inspired by the following observations. The statistical interpolation polynomial generator needs to perform the deterministic electrothermal solver once for calculating the temperature profile with nominal device parameters and execute the deterministic electrothermal solver $N_{\mathcal{H}}$-1 times for obtaining temperature variations corresponding to the nominal temperature profile. The temperature profile from the first part contributes the major portion of the mean profile of temperature distribution, and the temperature variations from the second part contribute a large portion of the variance and skewness profiles of temperature distribution. In practice, the magnitude of the mean temperature profile is larger than those of the variance and skewness profiles of temperature distribution, since process variations of parameters are usually within a controllable range.

Based on the preceding observations, the mixed-mesh strategy for generating the statistical interpolation polynomial of on-chip temperature distribution is illustrated in Figure 10. Since the mean profile contributes a major portion to the thermal yield profile, the nominal temperature profile is built by the deterministic electrothermal solver with a fine mesh having $N_{\mathrm{F}}N_{\mathrm{F}}$ grids to preserve the estimation accuracy of the mean temperature profile. Then, the difference between the maximum and minimum temperature values, $\Delta\overline{T}_{\max}$, of the nominal temperature profile is calculated, and an allowable temperature resolution, T_{res}, is chosen. After that, the remaining $N_{\mathcal{H}} - 1$ deterministic electrothermal simulations are executed with a coarse mesh having $N_{\mathrm{C}}N_{\mathrm{C}}$ grids. Here, N_{C} is equal to $\lceil \Delta\overline{T}_{\max}/T_{\mathrm{res}} \rceil$. Finally, the mean, variance, and skewness profiles of on-chip temperature distribution can be approximated by the generated

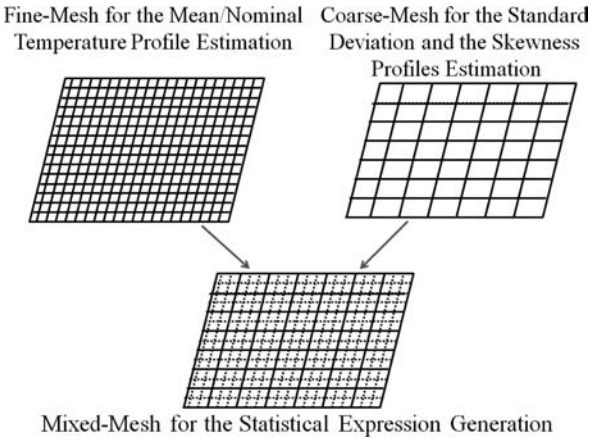

Fig. 10. The sketch of mixed-mesh strategy for generating the statistical interpolation polynomial of on-chip temperature distribution.

statistical interpolation polynomial, and these temperature profiles are utilized to calculate the thermal yield profile.

With the proposed mixed-mesh strategy, the runtime for generating the statistical interpolation polynomial of on-chip temperature distribution can be significantly reduced. In this work, an effective deterministic thermal simulator [Huang and Lee 2009] is adopted as the kernel of our developed deterministic electrothermal solver.[9] The computational complexity of the deterministic thermal simulator presented in Huang and Lee [2009] is $O(N_{\mathrm{M}}N_{\mathrm{M}} \log N_{\mathrm{Base}})$. Here, $N_{\mathrm{M}}N_{\mathrm{M}}$ is the mesh size, and N_{Base} is the number of bases for expressing the deterministic temperature profile. Generally, by using the average chip temperature calculated by the iterative computation scheme of the 1D thermal model shown in Figure 3 to be the initial operating temperature, the number of electrothermal loops for achieving convergence can be less than a small value. Hence, the computational complexity of the developed deterministic electrothermal solver is also $O(N_{\mathrm{M}}N_{\mathrm{M}} \log N_{\mathrm{Base}})$.

Therefore, the computational complexity of our baseline algorithm (the fine mesh is used for each deterministic electrothermal simulation) stated in Figure 4 is $O(N_{\mathcal{H}}N_{\mathrm{F}}N_{\mathrm{F}} \log N_{\mathrm{Base}})$, and the computational complexity by utilizing the developed mixed-mesh strategy can be reduced to $O((N_{\mathrm{F}}N_{\mathrm{F}} + (N_{\mathcal{H}} - 1)N_{\mathrm{C}}N_{\mathrm{C}}) \log N_{\mathrm{Base}})$.

The computational complexity ratio of the developed mixed-mesh strategy for generating the statistical interpolation polynomial to the deterministic electrothermal solver with nominal device parameters is equal to $1+(N_{\mathcal{H}}-1)\times(N_{\mathrm{C}}/N_{\mathrm{F}})^2$. In our experimental results, an accurate thermal yield profile can be estimated with $N_{\mathcal{H}} = 53$, $N_{\mathrm{F}} = 128$, $N_{\mathrm{C}} = 16$, and $T_{\mathrm{res}} = 0.65^{\circ}$C. The computational complexity ratio is 1.8125. Therefore, the mixed-mesh strategy does enhance the efficiency of the developed statistical electrothermal simulator for catching up with that of a deterministic electro-thermal simulator.

5. EXPERIMENTAL RESULTS

The developed statistical electrothermal simulator is implemented in C++ language and tested on a Linux system with Intel Xeon 3.0-GHz CPU and 32GB memory. The die size is 2.5 mm $\times$ 2.5 mm $\times$ 0.5 mm. The junction depth is set to 20nm for the 65nm

[9]As reported in Huang and Lee [2009], it took only 0.13 seconds to obtain the temperature profile of a chip with one million functional blocks and 1024×1024 simulation mesh.

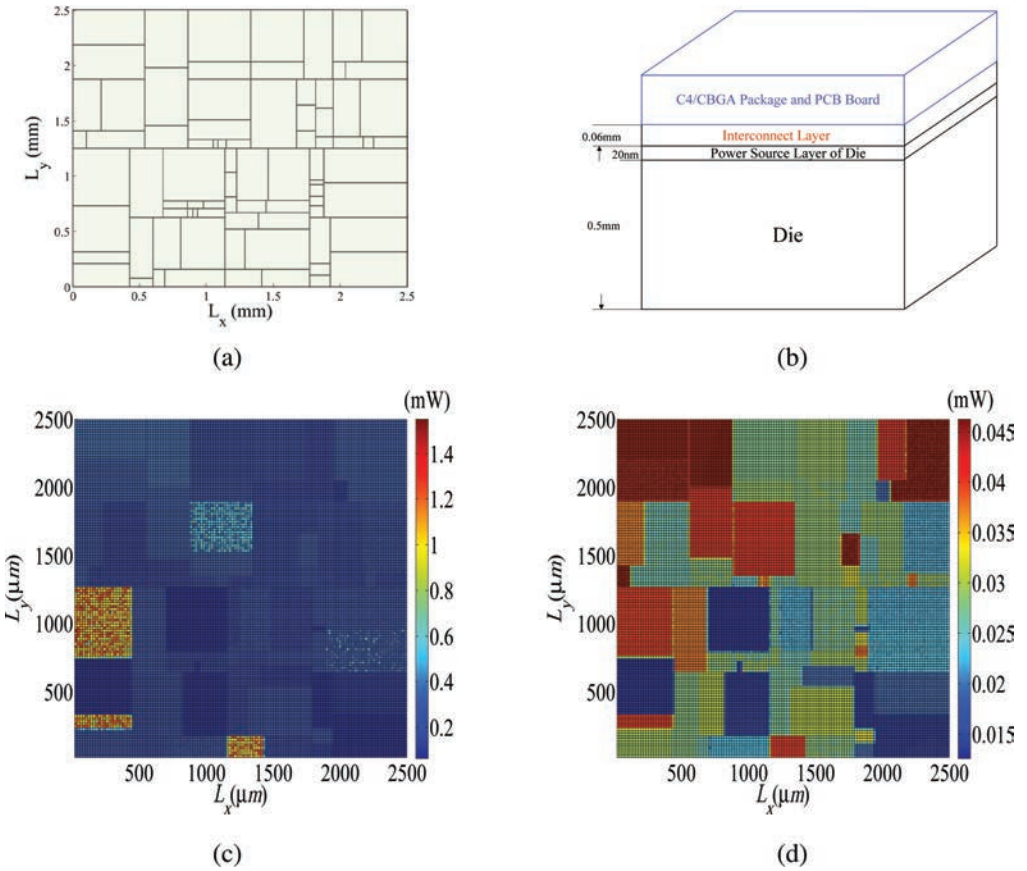

Fig. 11. Power map of the test chip: (a) floorplan, (b) geometries of the die and package, (c) the mean profile of the power map, (d) the standard deviation profile of the power map. Here, L_x and L_y are the width and length of the test die, respectively.

technology [Lallement et al. 2004], and the Debye length is 2nm [Bienacel et al. 2004]. The floorplan of a test chip having 1.2 million functional gates is shown in Figure 11(a), and the geometries of chip and package are shown in Figure 11(b).

By applying the modeling skill of thermal parameters mentioned in Figure 3 of Section 3.1 and the modeling skill for both heat transfer paths mentioned in Huang et al. [2006], the thermal conductivity and the equivalent heat transfer coefficients of the primary and secondary heat flow paths for executing the deterministic thermal simulator [Huang and Lee 2009] are summarized in Table III. The boundary condition of each vertical surface is set to be isothermal [Huang and Lee 2009].

The device parameters, the truncation points of KL expansions for the channel length (N_L) and the oxide thickness ($N_{t_{ox}}$), and the number of device modeling grid (N_{KL_g}) are summarized in Table IV. Both N_L and $N_{t_{ox}}$ are decided by using the criterion stated in Eq. (16) with $\varepsilon = 1\%$. To model the spatial correlation, both η_x/L_x and η_y/L_y are set to 0.98 for the correlation function shown in Eq. (9) [Cline et al. 2006]. The active layer of the test chip is divided into 128×128 simulated grids.

The estimated mean and standard deviation profiles of the power map under the settings of 60% WID and 40% D2D variations to the total variations are shown in Figures 11(c)–(d), respectively.

Table III. Equivalent Thermal Parameters

Parameter	Value
κ	104.6 W/(m·°C)
h_p	12,000 W/(m^2·°C)
h_s	2,017 W/(m^2·°C)

κ: the thermal conductivity of the die.
h_p: the equivalent primary heat transfer coefficient.
h_s: the equivalent secondary heat transfer coefficient.

Table IV. Parameters and Truncation Points for the Channel Length and Oxide Thickness

Nominal L	Nominal t_{ox}	$3\sigma_L$	$3\sigma_{t_{ox}}$	N_L	$N_{t_{ox}}$	N_{KL_g}
65nm	1.5nm	12%	5%	13	13	49

5.1. Electrothermal *vs.* Nonelectrothermal Simulations

With the number of grids being 100 for modeling device parameters and the ratios of WID and D2D variations to the total variations being 60% and 40%, respectively, the MC method with 2×10^4 samples is employed to demonstrate essentialness of the electrothermal simulation loop. The estimated mean and standard deviation profiles of the on-chip temperature distribution with and without considering the temperature-dependent issue of leakage power are shown in Figures 12(a) and 12(c), and Figures 12(b) and 12(d), respectively. For the mean profile estimation, the difference between Figure 12(a) and Figure 12(b) is over 16%. For the standard deviation profile estimation, the difference between Figure 12(c) and Figure 12(d) is over 31%. These results indicate that statistical electrothermal analysis is essential.

5.2. Accuracy and Efficiency

This section is going to demonstrate the correctness and efficiency of the developed statistical electrothermal simulator and show its efficiency improvement by using the mixed-mesh strategy.

Given three different ratio pairs of the WID and D2D variations to the total variations, (WID, D2D) = (40%, 60%), (50%, 50%), or (60%, 40%), the results from 2×10^4 MC simulations, which satisfy the maximum absolute relative error of variance is less than 1%, are utilized as the reference solution.

5.2.1. Mean and Standard Deviation Estimation. To demonstrate the accuracy and efficiency of the developed statistical interpolation polynomial generator, the level-1 Smolyak sparse grid formula with the sampling points being the roots of HPs is built. The deterministic electrothermal solver needs to be executed 53 times, since the number of sampling points ($N_{\mathcal{H}}$) for physical parameters to obtain the level-1 Smolyak sparse grid formula is equal to $2 \times (N_L + N_{t_{ox}}) + 1$, and both N_L and $N_{t_{ox}}$ are calculated to be 13, as shown in Table IV. The size of each simulation mesh is 128×128.

The maximum absolute errors of mean and standard deviation profiles are presented in Table V. The first two columns indicate the ratio pairs of WID and D2D variations to the total variations. Compared with 2×10^4 MC simulations, the maximum absolute errors of the estimated mean and standard deviation profiles from the developed statistical interpolation polynomial generator are shown in the fifth and sixth columns, respectively. As shown in Table V, the maximum absolute errors are less than 3.0% for all three different ratio pairs.

To fairly compare the runtime, the MC simulation is performed till achieving the same accuracy of standard deviation as the developed statistical interpolation polynomial generator. The number of MC simulations is shown in the #Samples column, and

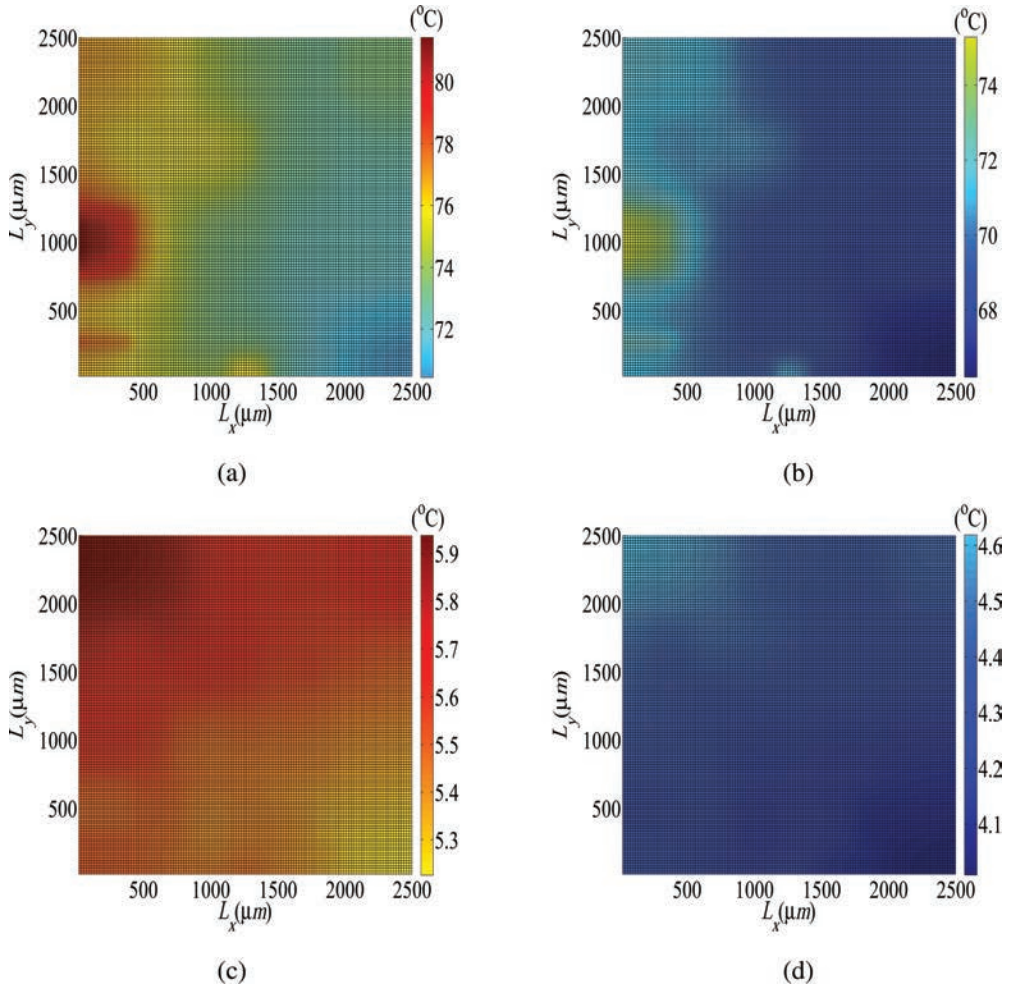

Fig. 12. Results of the MC method with or without considering the electrothermal effect: (a) and (b) are the mean temperature profiles with and without considering the electrothermal effect, respectively; (c) and (d) are the standard deviation profiles of temperature distribution with and without considering the electrothermal effect, respectively.

the Runtime column denotes the runtime for both methods. According to the Speedu column, the developed statistical interpolation polynomial generator can be orders of magnitude faster than the MC method. Under the ratio pair (WID, D2D) = (60%, 40%), the developed statistical interpolation polynomial generator takes 2.74 seconds to generate the interpolation polynomial of temperature profile for the 128×128 simulation mesh. It contains 0.47 seconds for executing 53 deterministic electrothermal simulations, and 2.27 seconds to generate the interpolation polynomials after the 53 sampling temperature profiles are obtained.

With the ratio pair (WID, D2D) = (60%, 40%), the estimated mean and standard deviation profiles of on-chip temperature distribution are shown in Figures 13(a) and 13(b), respectively. The error distributions of the mean and standard deviation of on-chip temperature distribution are presented in Figures 13(c) and 13(d), respectively.

Table V. Accuracy and Efficiency Comparison of the Developed Statistical Interpolation Polynomial Generator and the MC Method

WID Ratio	W2D Ratio	MC Method		Developed Statistical Interpolation Polynomial Generator			Speedup (×)
		#Samples	Runtime (s)	Maximum Mean Error	Maximum Standard Deviation Error	Runtime (s)	
40%	60%	6,921	442.94	0.91%	2.70%	2.68	165.2
50%	50%	7,011	448.70	0.91%	2.68%	2.72	164.9
60%	40%	7,031	449.98	0.90%	2.72%	2.74	164.2

†The error is calculated by comparing with 2×10^4 MC simulations.
"Runtime" does not include the time for parsing input files.

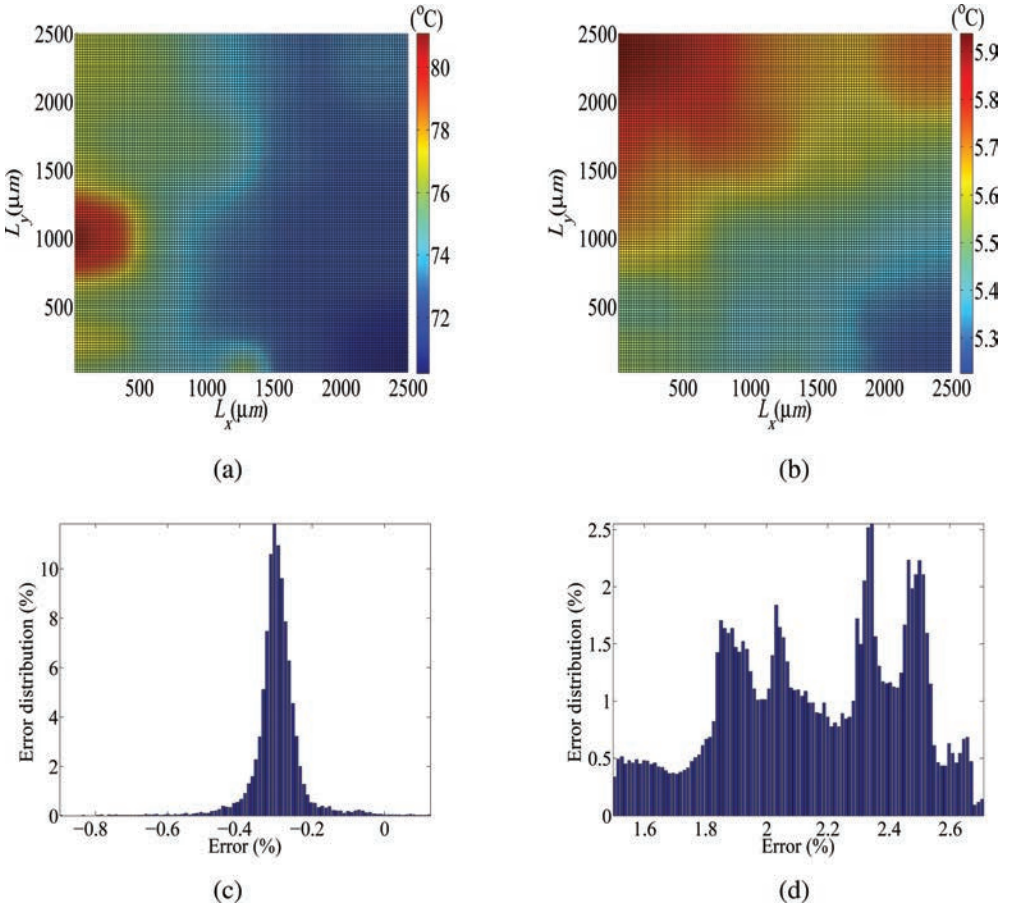

Fig. 13. Estimated mean and standard deviation profiles of on-chip temperature distribution with the ratio pair (WID, D2D) = (60%, 40%): (a) and (b) are the estimated mean and standard deviation profiles, respectively; (c) and (d) are the error distributions of the estimated mean and standard deviation profiles compared with the MC method, respectively.

Table VI. Accuracy and Efficiency Comparison of APEX [Li et al. 2004] and the Skew Normal Model Method

WID Ratio	D2D Ratio	T_{ref}	APEX [Li et al. 2004]		Skew Normal Model		Speedup (×)
			Runtime (s)	Maximum Error	Runtime (s)	Maximum Error	
40%	60%	88.40°C	2.80	1.97%	0.013	1.63%	215.38
50%	50%	88.48°C	2.80	1.90%	0.013	1.52%	215.38
50%	50%	88.54°C	2.80	2.32%	0.013	1.41%	215.38

†The error is calculated by comparing with 2×10^4 MC simulations.

5.2.2. Thermal Yield Estimation. To demonstrate the accuracy and efficiency of the developed thermal yield estimation engine by using the skew normal model, APEX [Li et al. 2004], a well-known cumulative distribution function estimation method, is also implemented. To avoid the instability of Padè approximation for APEX, a stable two-pole model [Tutuianu et al. 1996] is employed for finding poles and zeros. The developed statistical interpolation polynomial generator is utilized to generate the statistical expression of temperature distribution for both the skew normal model method and APEX. With the average mean ($\overline{\mu}_T$) and standard deviation ($\overline{\sigma}_T$) of temperature distribution obtained by 2×10^4 MC simulations, the specified reference temperature, T_{ref}, is assumed to be $\overline{\mu}_T + 2.5\overline{\sigma}_T$.

Table VI summarizes the results, and it shows that both the skew normal model method and APEX can accurately provide the thermal yield profile for each test case; furthermore, the developed skew normal model method is more accurate than that of APEX. The maximum error of the skew normal model method is 1.63%, and the maximum error of APEX is 2.32%. The results also reveal that the developed skew normal model method achieves 215× speedup over APEX. This is for two reasons. One is that APEX needs higher-order statistical moments to get a tight bound of its generalized Chebyshev inequality for the PDF/CDF shifting process. In our implementation, APEX requires moments up to the ninth order to achieve an accurate thermal yield profile estimate, even though it only needs moments up to the fourth order to get the first two dominated poles [Tutuianu et al. 1996]. The other is that APEX needs to solve zeros for constructing its exponential model after two dominated poles are computed. Compared with APEX, after the moments of temperature distribution up to the third order are computed, the skew normal model method simply looks up the CDF table of the skew-normal random variable to estimate the thermal yield profile.

The thermal yield profiles obtained by 2×10^4 MC simulations and estimated by the developed skew normal method and APEX under 60% of WID variation and 40% of D2D variation to the the total variations are presented in Figures 14(a)–14(c), respectively. The errors of estimated thermal yield profiles by the skew normal model method and APEX are plotted in Figures 14(d)–14(e), respectively. As shown in Figures 14(b)–14(c), the developed statistical interpolation polynomial generator can provide accurate statistical on-chip temperature expression for the thermal yield estimation, and the thermal yield profile can indicate thermal reliabilities at different locations of the die, that is, the smaller the thermal yield at a location, the less reliable the location. Figure 14(c) also shows that the estimated thermal yield profile of APEX might impractically exceed 100% in some region, since APEX does not guarantee generating a statistical model with preserving the property of CDF.

To further demonstrate the accuracy of the skew normal model method, the estimated CDF of temperature distribution at position **A** in Figures 14(a)–14(c) obtained by 2×10^4 MC simulations, the skew normal model method, and APEX with the ninth and fourth order moments for the PDF/CDF shifting process are drawn in Figure 15. As shown in Figure 15, the result of the skew normal model method fits that of MC simulations very well. However, APEX with the fourth order for the PDF/CDF shifting process cannot

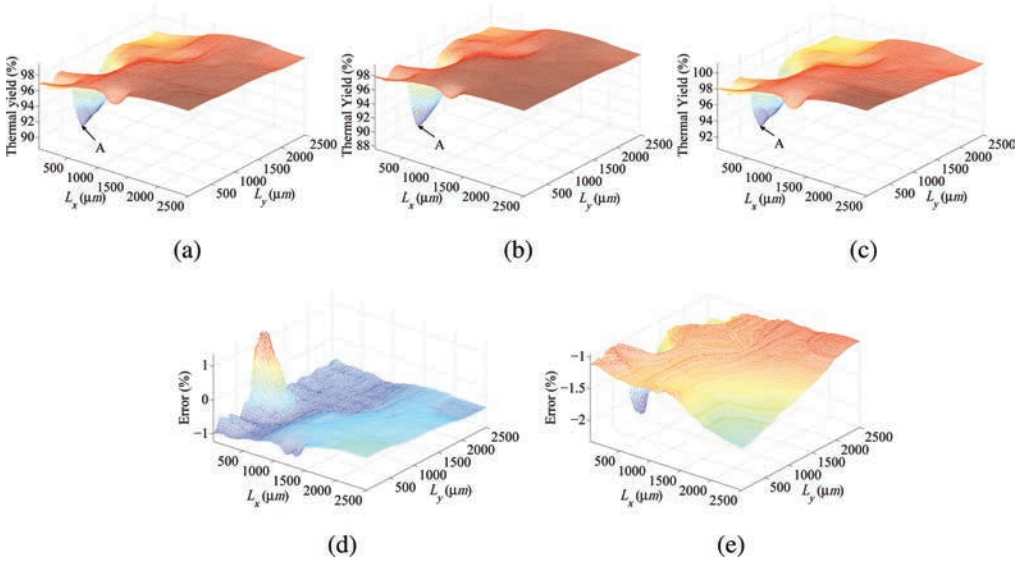

Fig. 14. Estimated thermal yield profiles and error profiles of the skew normal model and APEX: (a) the thermal yield profile by 2×10^4 MC simulations; (b) and (c) the estimated thermal yield profiles by the skew normal model method and APEX, respectively; (d) and (e) the error profiles of estimated thermal yields by the skew normal model and APEX, respectively.

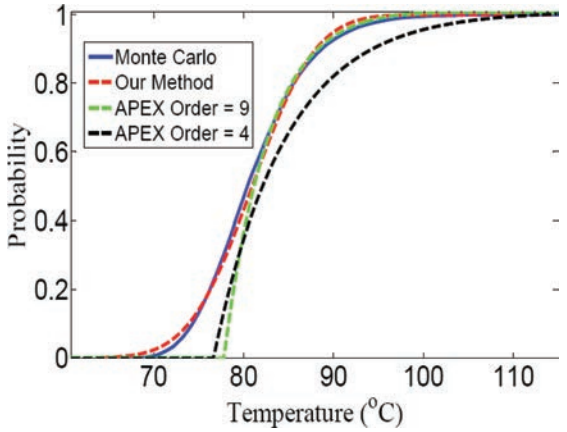

Fig. 15. CDFs of temperature at position **A** in Figures 14(a)–14(c) obtained by 2×10^4 MC simulations, the skew normal model method, APEX with the ninth order for the PDF/CDF shifting process, and APEX with the fourth order for the PDF/CDF shifting process.

meet the CDF obtained by the MC simulations, and APEX with the ninth order for the PDF/CDF shifting process does not provide the accurate estimate of thermal yield with smaller reference temperature values.

5.2.3. Mixed-Mesh Strategy for Thermal Yield Estimation. As demonstrated in Section 5.2.2, the developed skew normal model-based thermal yield profile estimation engine can be two orders of magnitude faster than APEX and can obtain the thermal yield profile in 0.013 seconds. However, as shown in Table V, it still requires couples of seconds to generate the statistical interpolation polynomial of temperature distribution. Here, we are going to show the superiority of the developed mixed-mesh strategy for thermal

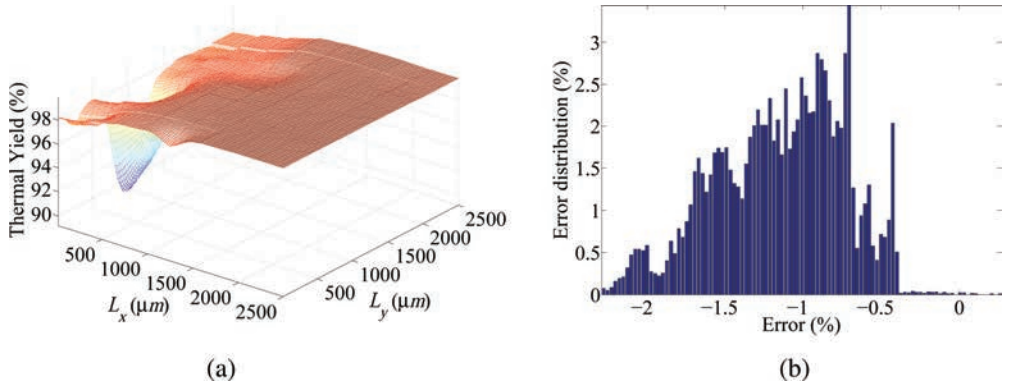

Fig. 16. Estimated thermal yield profile and error distribution of the developed mixed-mesh strategy.

Table VII. Accuracy and Efficiency Comparison between the Baseline and Mixed-Mesh Strategy Thermal Yield Profile Estimation Methods

Method	Mesh Size		Runtime (s)			Maximum Error	Speedup Ratio
	Nominal	52 (= $2N_{KL}$)	Expression Generation	Thermal Yield Profile	Total		
Baseline	128 × 128	128 × 128	2.740	0.013	2.753	1.41%	1.0
Mixed Mesh	128 × 128	16 × 16	0.049	0.013	0.062	2.28%	44.4

†The error is obtained by comparing with the MC method with 2×10^4 samples.

yield profile estimation with the test case having 60% WID variation ratio and 40% D2D variation ratio. For the accuracy verification, its results are compared with 2×10^4 MC simulations, and for the efficiency demonstration, its results are compared with the baseline method of the developed skew normal model thermal yield profile estimation engine with the size of the simulation mesh being equal to 128 × 128 for each deterministic electrothermal simulation corresponding to each sampling point in the Smolyak sparse grid.

By using the mixed-mesh strategy, first, the simulation mesh for calculating the nominal temperature profile of the test case is set to 128 × 128. After obtaining the nominal temperature profile, the difference between the maximum and minimum temperature values $\Delta\overline{T}_{\max}$ is calculated to be 10.1°C, and the temperature resolution T_{res} is set to 0.65°C for the coarse-mesh construction. Under this setting, the size of the coarse mesh is calculated as 16 × 16 for executing the remaining 52 (= 2 × N_{KL}) deterministic electrothermal simulations. The estimated thermal yield profile by using the mixed-mesh strategy is shown in Figure 16(a). Compared with Figure 14(a), the estimated thermal yield profile matches with that of the MC simulations. Its error distribution is drawn in Figure 16(b), and the plot shows that the errors are within the range [−2.28%, 0.13%].

Comparison between the developed baseline and mixed-mesh strategy thermal yield profile estimation methods is summarized in Table VII. The Expression Generation column denotes the runtime of generating the statistical interpolation polynomial, and it is 2.74 seconds and 0.049 seconds for the baseline method and the mixed-mesh strategy, respectively. The Thermal Yield Profile column shows the runtime to obtain the thermal yield profile by using the skew normal random variable model, and it is 0.013 seconds for both methods. From the last three columns of Table VII, we can see that the developed mixed-mesh strategy, with slightly trading off the accuracy, can achieve 44.4× speedup over the developed baseline method.

6. CONCLUSION AND POTENTIAL APPLICATIONS

An efficient statistical electrothermal simulator is proposed. The developed statistical electro-thermal simulator not only provides the mean and standard deviation profiles of on-chip temperature distribution, but also delivers the thermal yield profile of on-chip temperature distribution to designers and provides a proper figure of merit to indicate the thermal reliability of designed circuit.

Compared with the MC method and APEX [Li et al. 2004], the experimental results demonstrate that the proposed statistical electrothermal simulator can accurately and efficiently provide the mean, standard deviation, and thermal yield profiles of on-chip temperature distribution.

Potential extensions and applications of the developed framework are summarized as follows.

6.1. Statistical Electrothermal Analysis of 3D ICs

One possible strategy of extending our framework to 3D ICs is summarized as follows.

(1) Construct the equivalent modified nodal analysis (MNA) system for a 3D IC with the process variation and temperature-dependent power consumption vector at the right-hand side of MNA equation.
(2) Build the sparse grid sample points for KL expanded or PCA decomposed random variables of the physical parameters. Use these sample points to get the deterministic physical parameters, and hence, we have several deterministic temperature-dependent power consumption vectors corresponding to these deterministic physical parameters.
(3) Perform the deterministic electrothermal analysis with each deterministic temperature-dependent power consumption vector by using any existing deterministic electrothermal simulators to get the corresponding temperature profile.
(4) Interpolate the statistical polynomials of the on-chip temperature distribution, and apply the developed thermal yield profile estimation technique to it.

6.2. Thermal-Aware Statistical Timing Analysis

One possible extension strategy for the thermal-aware statistical timing analysis is briefed as follows. Given different channel-length variations, different oxide-thickness variations, different operating temperatures, and different load capacitances, the delay data of each gate type in the cell library can be constructed by utilizing the HSPICE. With the simulated data and the least square fitting methodology, the delay of a specific gate type can be fitted as the first-order canonical form

$$D = a_0 + a_1 L + a_2 t_{ox} + a_3 T, \tag{42}$$

where the fitting coefficients a_0, a_1, a_2, and a_3 are load-capacitance dependent.

After that, by using the formulation of KL expansion or PCA decomposition, $L(\mathbf{r}^*)$ and $t_{ox}(\mathbf{r}^*)$ at the arbitrary position $\mathbf{r}^*$ can be represented as a linear combination of their transformed Gaussian random variables. Then, using the level-1 Smolyak sparse grid formulation, the delay of an arbitrary gate type at position $\mathbf{r}^*$ can be expressed as

$$D = \sum_{i=1}^{N_{par}} \left(b_i \xi_i + c_i \xi_i^2 \right) + d. \tag{43}$$

Here, b_i's, c_i's, and d are constants, and ξ_i's are the Gaussian random variables transformed by KL expansion or PCA decomposition. With the expression stated in Eq. (43), any existing second-order statistical static timing analysis engine can be utilized to perform the thermal-aware statistical timing analysis.

6.3. Possible Utilizing Scenarios of the Thermal Yield Profile

As indicated in Tsai et al. [2006], several existing thermal-aware placers have successfully taken the temperature effect into account by using deterministic thermal analyzers to calculate the thermal-related cost. By using similar strategies presented in Tsai et al. [2006], the developed thermal yield profile estimation engine can also be utilized to calculate the thermal-related cost for thermal-aware placing flows, since the quantity of thermal yield is intrinsically deterministic.

A multiple supply voltage design method for 3D ICs was proposed in Yu et al. [2009], and it can also be used for 2D ICs. Yu et al. [2009] calculated the thermal-related cost incrementally by using the mean and standard deviation of on-chip temperature distribution. Since the thermal yield is a better figure of merit for quantifying the thermal effect, instead of using the mean and standard deviation of on-chip temperature distribution, the developed thermal yield profile can be and should be adopted for that work.

APPENDIXES

A. AN EXAMPLE FOR ILLUSTRATING THE SMOLYAK SPARSE GRID FORMULATION

A simple example with $N_{KL} = 2$ and $q = N_{KL} + 1 = 3$ is given to illustrate the Smolyak sparse grid formulation. We have $2 \leq |\mathbf{i}| \leq 3$, since $q - N_{KL} + 1 \leq |\mathbf{i}| \leq q$. Hence, $i_1 = 1, i_2 = 1$ for $|\mathbf{i}| = 2$, and $i_1 = 1$, $i_2 = 2$ or $i_1 = 2$, $i_2 = 1$ for $|\mathbf{i}| = 3$. The numbers of sample values for random variables ξ_1 and ξ_2 are $m_{i_1=1} = 1$ and $m_{i_2=1} = 1$ for $|\mathbf{i}| = 2$, and $m_{i_1=1} = 1$ and $m_{i_2=2} = 3$ or $m_{i_1=2} = 3$ and $m_{i_2=1} = 1$ for $|\mathbf{i}| = 3$, respectively. According to various values of i_1 or i_2, the interpolating polynomial forms can be determined by utilizing a single random variable ξ_1 or ξ_2.

After that, the interpolating polynomial forms corresponding to $\boldsymbol{\xi}^T = [\xi_1, \xi_2]$ at different combined values of i_1 and i_2 can be constructed by using the functional cross-product operation. For example, $\mathcal{Q}^{i_n=1}(T) = 1$ and $\mathcal{Q}^{i_n=2}(T) = p_0 + p_1\xi_n + p_2\xi_n^2$ are the first-order and second-order interpolating polynomial forms for $n \in \{1, 2\}$, respectively. $\mathcal{Q}^{i_1=1}(T) \otimes \mathcal{Q}^{i_2=2}(T) = 1 \otimes (b_0 + b_1\xi_2 + b_2\xi_2^2) = b_0 + b_1\xi_2 + b_2\xi_2^2$, and $\mathcal{Q}^{i_1=2}(T) \otimes \mathcal{Q}^{i_2=1}(T) = (a_0 + a_1\xi_1 + a_2\xi_1^2) \otimes 1 = a_0 + a_1\xi_1 + a_2\xi_1^2$. Here, a_j's and b_j's are constant coefficients that can be determined by using the sample values of $T(\mathbf{r}, \boldsymbol{\xi})$.

To obtain $\mathcal{Q}^{i_1}(T) \otimes \mathcal{Q}^{i_2}(T)$ at different combined values of i_1 and i_2, only the on-chip temperature distribution excited by the point that belongs to the following sample set of $\boldsymbol{\xi}$ needs to be known. Given $\hbar^1 = \{p_0^1\}$ and $\hbar^2 = \{p_0^2, p_1^2, p_2^2\}$, we have

$$\begin{aligned}
\mathcal{H}(3,2) &= (\hbar^{i_1=1} \times \hbar^{i_2=1}) \cup (\hbar^{i_1=1} \times \hbar^{i_2=2}) \cup (\hbar^{i_1=2} \times \hbar^{i_2=1}) \\
&= \left\{ \left[p_0^1, p_0^1\right]^T \right\} \cup \left\{ \left[p_0^1, p_0^2\right]^T, \left[p_0^1, p_1^2\right]^T, \left[p_0^1, p_2^2\right]^T \right\} \\
&\quad \cup \left\{ \left[p_0^2, p_0^1\right]^T, \left[p_1^2, p_0^1\right]^T, \left[p_2^2, p_0^1\right]^T \right\} \qquad (44) \\
&= \left\{ \left[p_0^1, p_0^1\right]^T, \left[p_0^1, p_0^2\right]^T, \left[p_0^1, p_1^2\right]^T, \left[p_0^1, p_2^2\right]^T, \left[p_0^2, p_0^1\right]^T, \left[p_1^2, p_0^1\right]^T, \left[p_2^2, p_0^1\right]^T \right\}
\end{aligned}$$

B. ENHANCEMENT STRATEGY OF THE STABILITY OF THE THERMAL YIELD PROFILE ESTIMATION ENGINE

As shown in Eqs. (34)–(37), $|\delta_{\mathbf{r}^*}|$ needs to be less than $\sqrt{2/\pi}$ to ensure the stability of the moment matching process of the skew-normal model. Under this situation, that is, setting $|\delta_{\mathbf{r}^*}| < \sqrt{2/\pi}$ in Eq. (37), the skewness of $|\gamma_{\Delta\widehat{T}}(\mathbf{r}^*)|$ that the model can match

requires

$$|\gamma_{\Delta\widehat{T}}(\mathbf{r}^*)| < \left(\frac{\left(\frac{4-\pi}{2}\right)^{2/3} \frac{2}{\pi}}{1-\frac{2}{\pi}} \right)^{3/2} \approx 0.99527. \tag{45}$$

Comparing with the variance of $\widehat{T}(\mathbf{r}^*)$, $\sigma_{\widehat{T}}(\mathbf{r}^*) = 1$, this would be an extreme situation in the practical application, because the PDF of $\widehat{T}(\mathbf{r}^*)$ will be fairly close to the right/left direction exponential shape, and the probability that the temperature is less/larger than its nominal value will be near to zero for the positive/negative skewness case.

However, this is indeed the limitation of the skew-normal model. Fortunately, with the technique proposed in Raphaeli [1996], the PDF of the weighted sum of independent non-central chi-square random variables, which is exactly consistent with the form of $\widehat{T}(\mathbf{r}^*)$ stated in Eq. (27), can be stably approximated by the series expansion up to any specific order. Due to the limitation of space, please refer to Raphaeli [1996] for the details. Although the technique stated in Raphaeli [1996] is stable for any specific order, its complexity is higher than that of the skew-normal moment matching technique if the three-order approximation is proceed. Therefore, we can still apply the skew-normal model for the temperature as its skewness satisfies Eq. (45), and apply the formula stated in Raphaeli [1996] for the temperature as the skewness violates Eq. (45). Since the situation of violating Eq. (45) is extreme, only a small amount of grids need to perform the formula stated in Raphaeli [1996]. Therefore, the efficiency of the thermal yield estimation can still be preserved.

C. EXTENSION OF THERMAL YIELD PROFILE ESTIMATION ENGINE FOR THE SECOND-ORDER POLYNOMIAL WITH CROSS PRODUCT TERMS

With level l in Smolyak sparse grid formulation being 2, the interpolation polynomial of statistical temperature at an arbitrary location $\mathbf{r}^*$ can be approximated as the following quadratic form.

$$\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \boldsymbol{\xi}^T \boldsymbol{A}(\mathbf{r}^*)\boldsymbol{\xi} + \boldsymbol{b}^T(\mathbf{r}^*)\boldsymbol{\xi} + c(\mathbf{r}^*). \tag{46}$$

Here, $\boldsymbol{A}(\mathbf{r}^*)$ is an $N_{KL} \times N_{KL}$ matrix with its ijth entry a_{ij} being the leading coefficient corresponding to $\xi_i\xi_j$ for $1 \le i, j \le N_{KL}$, $\boldsymbol{b}(\mathbf{r}^*)$ is the $1 \times N_{KL}$ vector with its kth entry b_k being the leading coefficient corresponding to ξ_k for $1 \le k \le N_{KL}$, and $c(\mathbf{r}^*)$ is a constant.

Utilizing the eigenvalue decomposition, $\boldsymbol{A}(\mathbf{r}^*)$ can be written as

$$\boldsymbol{A}(\mathbf{r}^*) = \boldsymbol{V}^T(\mathbf{r}^*)\boldsymbol{D}(\mathbf{r}^*)\boldsymbol{V}(\mathbf{r}^*), \tag{47}$$

where $\boldsymbol{V}(\mathbf{r}^*)$ and $\boldsymbol{D}(\mathbf{r}^*)$ are the eigenvector matrix and the diagonal eigenvalue matrix of $\boldsymbol{A}(\mathbf{r}^*)$, respectively. Plugging Eq. (47) into Eq. (46) and representing $\boldsymbol{z}(\mathbf{r}^*) = \boldsymbol{V}(\mathbf{r}^*)\boldsymbol{\xi}$, we have

$$\widehat{T}(\mathbf{r}^*, \boldsymbol{\xi}) = \widehat{T}(\mathbf{r}^*, \boldsymbol{z}(\mathbf{r}^*)) = \boldsymbol{z}^T(\mathbf{r}^*)\boldsymbol{D}(\mathbf{r}^*)\boldsymbol{z}(\mathbf{r}^*) + \hat{\boldsymbol{b}}^T(\mathbf{r}^*)\boldsymbol{z}(\mathbf{r}^*) + c(\mathbf{r}^*). \tag{48}$$

Here, $\hat{\boldsymbol{b}}(\mathbf{r}^*) = \boldsymbol{V}(\mathbf{r}^*)\boldsymbol{b}$. Since each ith entry of $\boldsymbol{z}(\mathbf{r}^*)$, $z_i(\mathbf{r}^*)$ is the linear combination of Gaussian random variables ξ_k's in $\boldsymbol{\xi}$, it is still a Gaussian random variable. Moreover, $\boldsymbol{D}(\mathbf{r}^*)$ is a diagonal matrix, Eq. (48) is consistent with the form in Eq. (27). Therefore, after performing the proceding manipulation, the proposed thermal yield profile estimation engine can also handle the second-order polynomial with cross-product terms.

Comparing with the level 1 Smolyak sparse grid formulation, we need to perform $O(N_{KL}^2)$ deterministic electrothermal simulations for generating the cross-product terms in Eq. (46). In this situation, our mixed-mesh thermal yield estimation can gain more benefits on efficiency, since the crossproduct terms contribute the values on variance and skewness, and these values can be computed with a coarser mesh.

REFERENCES

AZZALINI, A. 2005. The skew-normal distribution and related multivariate families. *Board Found. Scandinavian J. Stat. 32*, 159–188.

BARTHELMANN, V., NOVAK, E., AND RITTER, K. 2000. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math. 12,* 4, 273–288.

BHARDWAJ, S., VRUDHULA, S., AND GOEL, A. 2008. A unified approach for full chip statistical timing and leakage analysis of nanoscale circuits considering intradie process variations. *IEEE Tran. Comput. Aid. Des. Integ. Circ. Syst. 27,* 10, 1812–1825.

BIENACEL, J., BARGE, D., BIDAUD, M., EMONET, N., ROY, D., VISHNUBHOTLA, L., POUILLOUX, I., AND BARLA, K. 2004. Anticipation of nitrided oxides electrical thickness based on XPS measurement. *Mater. Sci. Semiconduct. Process. 7,* 4–6, 181–183.

BOTA, S., ROSALES, M., ROSELLO, J., KESHAVARZI, A., AND SEGURA, J. 2004. Within die thermal gradient impact on clock-skew: A new type of delay-fault mechanism. In *Proceedings of the IEEE International Test Conference*. 1276–1283.

CHAKRABORTY, A., DURAISAMI, K., SATHANUR, A., SITHAMBARAM, P., BENINI, L., MACII, A., MACII, E., AND PONCINO, M. 2008. Dynamic thermal clock skew compensation using tunable delay buffers. *IEEE Trans. Very Large Scale Integr. Syst. 16,* 6, 639–649.

CHANG, H. AND SAPATNEKAR, S. S. 2007. Prediction of leakage power under process uncertainties. *ACM Trans. Des. Autom. Electron. Syst. 12,* 2.

CHANG, H. C., HUANG, P. Y., LI, T. J., AND LEE, Y. M. 2010. Statistical electro-thermal analysis with high compatibility of leakage power models. In *Proceedings of the IEEE International SOC Conference*. 139–144.

CHENG, L., GUPTA, P., SPANOS, C. J., QIAN, K., AND HE, L. 2011. Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability. *IEEE Trans. Comput. Aid. Des. Integr. Circ. Syst. 30,* 3, 388–401.

CLINE, B., CHOPRA, K., BLAAUW, D., AND CAO, Y. 2006. Analysis and modeling of CD variation for statistical static timing. In *Proceedings of the International Conference on Computer Aided Design*. ACM, 60–66.

GAO, M., YE, Z., ZENG, D., WANG, Y., AND YU, Z. 2011. Robust spatial correlation extraction with limited sample via L1-norm penalty. In *Proceedings of the Asia and South Pacific Design Automation Conference*. IEEE Press, 677–682.

HAGHDAD, K. AND ANIS, M. 2012. Power yield analysis under process and temperature variations. *IEEE Trans. Very Large Scale Integr. Syst. 20,* 10, 1794–1803.

HAN, Y. AND KOREN, I. 2007. Simulated annealing based temperature aware floorplanning. *J. Low Power Electron. 3,* 2, 141–155.

HUANG, P. Y. AND LEE, Y. M. 2009. Full-chip thermal analysis for the early design stage via generalized integral transforms. *IEEE Trans. Very Large Scale Integr. Syst. 17,* 5, 613–26.

HUANG, P. Y., LEE, Y. M., AND PAN, C. W. 2012. On-chip statistical hot-spot estimation using mixed-mesh statistical polynomial expression generating and skew-normal based moment matching techniques. In *Proceedings of the Asia and South Pacific Design Automation Conference*. 603–608.

HUANG, P. Y., WU, J. H., AND LEE, Y. M. 2009. Stochastic thermal simulation considering spatial correlated within-die process variations. In *Proceedings of the Asia and South Pacific Design Automation Conference*. 55–60.

HUANG, W., GHOSH, S., VELUSAMY, S., SANKARANARAYANAN, K., SKADRON, K., AND STAN, M. R. 2006. HotSpot: A compact thermal modeling methodology for early-stage VLSI design. *IEEE Trans. Very Large Scale Integr. Syst. 14,* 5, 501–513.

ITRS. 2010. International technology roadmap for semiconductors. http://www.itrs.net/.

JAFFARI, J. AND ANIS, M. 2008. Statistical thermal profile considering process variation: Analysis and applications. *IEEE Trans. Comput. Aid. Des. Integr. Circ. Syst. 27,* 6, 1027–1040.

KUMAR, R. AND KURSUN, V. 2006. Reversed temperature-dependent propagationdelay characteristics in nanometer CMOS circuits. *IEEE Trans. Circ. Syst. Express Briefs 53,* 10, 1078–1082.

LALLEMENT, F., DURIEE, B., GROUILLET, A., AMAUD, F., TAVEL, B., WACQUANT, F., STALK, P., WOO, M., EROKHIN, Y., SCHEUER, J., GADET, L., WEEMAN, J., DISTASO, D., AND LENOBLEE, D. 2004. Ultra-low cost and high performance 65nm CMOS device fabricated with plasma doping. In *Proceedings of the Symposium on VLSl Technology Digest of Technical Papers*. 178–179.

LI, X., LE, J., GOPALAKRISHNAN, P., AND PILEGGI, L. T. 2004. Asymptotic probability extraction for non-normal distributions of circuit performance. In *Proceedings of the International Conference on Computer Aided Design*. 2–9.

Liu, C., Chen, R. X., Tan, J., Fan, S., Fan, J., and Makki, K. 2008. Thermal aware clock synthesis considering stochastic variation and correlations. In *Proceedings of the International Symposium on Circuits and Systems*. 1204–1207.

Liu, F. 2007. A general framework for spatial correlation modeling in VLSI design. In *Proceedings of the Design Automation Conference*. 817–822.

Pang, L. T. and Nikolic, B. 2009. Measurements and analysis of process variability in 90 nm CMOS. *IEEE J. Solid-State Circ. 44,* 5, 1655–1663.

Phillips, G. M. 2003. *Interpolation and Approximation by Polynomial*. Springer-Verlag, Berlin Heidelberg.

Raphaeli, D. 1996. Distribution of noncentral indefinite quadratic forms in complex normal variables. *IEEE Trans. Inf. Theory 42,* 3, 1002–1007.

Reda, S., Cochran, R., and Nowroz, A. 2011. Improved thermal tracking for processors using hard and soft sensor allocation techniques. *IEEE Trans. Comput. 60*, 6, 841–851.

Schwab, C. and Todor, R. A. 2006. Karhunen-Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Physics 217,* 1, 100–122.

Shen, R., Tan, S. X. D., Mi, N., and Cai, Y. 2010a. Statistical modeling and analysis of chip-level leakage power by spectral stochastic method. *Integr. VLSI J. 43,* 1, 156–165.

Shen, R., Tan, S. X. D., and Xiong, J. 2010b. A linear algorithm for full-chip statistical leakage power analysis considering weak spatial correlation. In *Proceedings of the Design Automation Conference*. ACM, 481–486.

Skadron, K., Stan, M. R., Sankaranarayanan, K., Huang, W., Velusamy, S., and Tarjan, D. 2004. Temperature-aware microarchitecture: Modeling and implementation. *ACM Trans. Architect. Code Optimi. 1,* 1, 94–125.

Smolyak, S. A. 1963. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Math. Doklady 4*, 240–243.

Tsai, J. L., Chen, C. C. P., Chen, G., Goplen, B., Qian, H., Zhan, Y., Kang, S. M., Wong, M. D. F., and Sapatnekar, S. S. 2006. Temperature-aware placement for SOCs. *Proc. IEEE 94,* 8, 1502–1518.

Tutuianu, B., Dartu, F., and Pileggi, L. 1996. An explicit RC-circuit delay approximation based on the first three moments of the impulse response. In *Proceedings of the Design Automation Conference*. 611–616.

Vassighi, A. and Sachdev, M. 2006. Thermal runaway in integrated circuits. *IEEE Trans. Device Mater. Reliab. 6,* 2, 300–305.

Wang, T. Y. and Chen, C. C. P. 2003. Thermal-ADI: A linear-time chip-level thermal simulation algorithm based on alternating-direction implicit (ADI) method. *IEEE Trans. Very Large Scale Integr. Syst. 11,* 4, 691–670.

Yang, Y., Gu, Z., Zhu, C., Dick, R. P., and Shang, L. 2007. ISAC: Integrated space-and-time-adaptive chip-package thermal analysis. *IEEE Trans. Comput. Aid. Desi. Integr. Circ. Syst. 26,* 1, 86–99.

Yu, S. A., Huang, P. Y., and Lee, Y. M. 2009. A multiple supply voltage based power reduction method in 3-D ICs considering process variations and thermal effects. In *Proceedings of the Asia and South Pacific Design Automation Conference*. 55–60.

Zhang, D. and Lu, Z. 2004. An efficient, high-order perturbation approach for flow in random porous media via Karhunen-Loève and polynomial expansions. *J. Comput. Physics 149,* 2, 773–794.

Received March 2012; revised September 2012; accepted January 2013