Accurate Temperature-Dependent Integrated Circuit Leakage Power Estimation is Easy

Yongpan Liu* ypliu99@mails.tsinghua.edu.cn
Robert P. Dick† dickrp@eecs.northwestern.edu
Li Shang‡ lshang@ee.queensu.ca
Huazhong Yang* yanghz@mail.tsinghua.edu.cn

* Electronic Engineering Dept. Tsinghua University Beijing, 100084, China
† EECS Dept. Northwestern University Evanston, IL 60208, U.S.A.
‡ ECE Dept. Queen's University Kingston, ON K7L 3N6, Canada

***Abstract*— It has been the conventional assumption that, due to the superlinear dependence of leakage power consumption on temperature, and widely varying on-chip temperature profiles, accurate leakage estimation requires detailed knowledge of thermal profile. Leakage power depends on integrated circuit (IC) thermal profile and circuit design style. We show that linear models can be used to permit highly-accurate leakage estimation over the operating temperature ranges in real ICs. We then show that for typical IC packages and cooling structures, a given amount of heat introduced at any position in the active layer will have similar impact on the average temperature of the layer. These two observations allow us to prove that, for wide ranges of design styles and operating temperatures, extremely fast, coarse-grained thermal models, combined with linear leakage power consumption models, permit highly-accurate system-wide leakage power consumption estimation. The results of our proofs are further confirmed via comparisons with leakage estimation based on detailed, time-consuming thermal analysis techniques. Experimental results indicate that the proposed technique yields a 59,259×–1,790,000× speedup in leakage power estimation while maintaining accuracy.**

I. Introduction

As a result of continued IC process scaling, the importance of leakage power consumption is increasing [1]. Leakage accounts for 40% of the power consumption of today's high-performance microprocessors [2]. Power consumption, temperature, and performance must now be optimized during the entire design flow. Leakage power consumption and temperature influence each other: increasing temperature increases leakage and vice versa . Leakage power estimation is frequently used in IC synthesis, within which it may be invoked tens of thousands of times: it must be both accurate and fast.

Researchers have developed a variety of techniques to characterize IC leakage power consumption, ranging from architectural level to device level [3]–[8]. However, most of these techniques neglect the dependence of leakage on temperature.

Leakage is a strong function of temperature. Therefore, thermal analysis must be embedded within the IC power analysis flow. Figure 1 shows a typical temperature-dependent IC leakage power estimation flow. Power consumption, including dynamic power and leakage power, is initially estimated at a reference temperature. The estimated power profile is then provided to a chip-package thermal analysis tool to estimate circuit thermal profile. This thermal profile is, in turn, used to update circuit leakage power estimation. This iterative process continues until power and temperature converge.

Recent work has considered the impact of temperature on leakage. Zhang et al. developed HotLeakage, a temperature-dependent cache leakage power model [9]. Su et al. proposed a full-chip leakage modeling technique that characterizes the impact of temperature and supply voltage fluctuations [10]. Liao et al. presented a temperature-dependent microarchitectural power model [11]. In leakage analysis, one can be confident of an accurate result by using a fine-grained thermal model. However, this is computationally intensive. One can also use a coarse-grained thermal model. Although fast, previous work has not demonstrated that this will permit accurate leakage estimation. Designers may select modeling granularity. However, without an understanding of the requirements necessary for accurate leakage prediction conservative designers are forced to use slow, fine-grained thermal models. This hinders the use of accurate IC leakage power estimation during IC synthesis.

This work was supported in part by the NSFC under awards 90207001 and 60506010; in part by the 863 Program under award 2006AA01Z224; in part by the NSF under award CNS-0347941; and in part by NSERC Discovery Grant 388694-01.

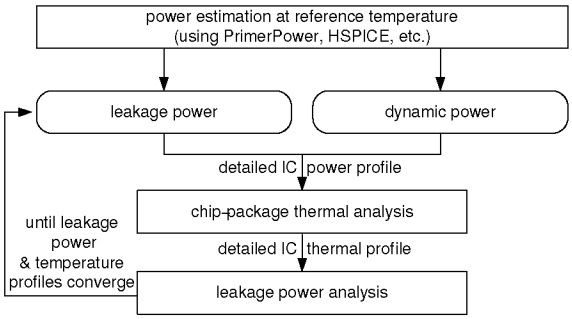

Fig. 1. Thermal-aware power estimation flow.

In this paper, we propose a very fast, accurate method of estimating IC leakage power consumption.

1) We demonstrate that, within the operating temperature ranges of ICs, using a linear leakage model for each functional unit results in less than 1% error in leakage estimation (Section II).

2) We demonstrate that IC packages and cooling structures have the useful property that a given amount of heat produced within the active layer of an IC will have similar impact on the average temperature of the active layer, regardless of its distribution (Section III).

3) We use the two properties described above to prove that within regions of uniform design style, knowledge of the average temperature is sufficient to accurately determine leakage power consumption. Based on this result, we show that leakage can be predicted using a simple, coarse-grained model without sacrificing accuracy (Section IV).

4) We validate the proposed technique via analytical proofs and simulation results. We demonstrate that for a wide range of ICs, a simplified thermal model in which only one thermal element is used for each functional unit permits a speedup in leakage estimation of 59,259×–1,790,000× while maintaining accuracy to within 1% (Section V), when compared with a conventional approach that uses a detailed thermal model.

II. Proposed Leakage Model

This section introduces IC leakage power consumption and characterizes leakage modeling linearization.

II.A. IC Leakage Sources

IC leakage current consists of various components, such as subthreshold leakage, gate leakage, reverse-biased junction leakage, punch-through leakage, and gate-induced drain leakage. Among these, subthreshold leakage and gate leakage are dominant [1]. They will be the focus of our analysis.

978-3-9810801-2-4/DATE07 © 2007 EDAA

Considering weak inversion Drain-Induced Barrier Lowering and body effect, the subthreshold leakage current of a MOS device can be modeled as follows [12]:

$$I_{subthreshold} = A_s \frac{W}{L} {v_T}^2 \left(1 - e^{\frac{-V_{DS}}{v_T}}\right) e^{\frac{(V_{GS}-V_{th})}{n v_T}} \quad (1)$$

- where A_s is a technology-dependent constant,
- V_{th} is the threshold voltage,
- L and W are the device effective channel length and width,
- V_{GS} is the gate-to-source voltage,
- n is the subthreshold swing coefficient for the transistor,
- V_{DS} is the drain-to-source voltage, and
- v_T is the thermal voltage.

$V_{DS} \gg v_T$ and $v_T = \frac{kT}{q}$. Therefore, Equation 1 can be reduced to

$$I_{subthreshold} = A_s \frac{W}{L} \left(\frac{kT}{q}\right)^2 e^{\frac{q(V_{GS}-V_{th})}{nkT}} \quad (2)$$

The gate leakage of a MOS device results from tunneling between the gate terminal and the other three terminals (source, drain, and body). Gate leakage can be modeled as follows [13]:

$$I_{gate} = WLA_J \left(\frac{T_{oxr}}{T_{ox}}\right)^{nt} \frac{V_g V_{aux}}{T_{ox}^2} e^{-BT_{ox}(a-b|V_{ox}|)(1+c|V_{ox}|)} \quad (3)$$

- where A_J, B, a, b, and c are technology-dependent constants,
- nt is a fitting parameter with a default value of one,
- V_{ox} is the voltage across gate dielectric,
- T_{ox} is gate dielectric thickness,
- T_{oxr} is the reference oxide thickness,
- V_{aux} is an auxiliary function that approximates the density of tunneling carriers and available states, and
- V_g is the gate voltage.

II.B. Thermal Dependency Linearizion

Equations 1–3 demonstrate that subthreshold leakage depends primarily on temperature, supply voltage, and body bias voltage. Gate leakage, in contrast, is primarily affected by supply voltage and gate dielectric thickness, but is insensitive to temperature. Using the Taylor series expansion at a reference temperature T_{ref}, the total IC leakage current of a MOS device can be expressed as follows:

$$I_{leakage}(T) = I_{subthreshold} + I_{gate} \quad (4)$$

$$= A_s \frac{W}{L} \left(\frac{k}{q}\right)^2 T^2 e^{\frac{q(V_{GS}-V_{th})}{nkT}} + I_{gate} \quad (5)$$

$$= I_{linear}(T) + I_{high\_order}(T) \quad (6)$$

where the linear portion $I_{linear}(T)$ is

$$I_{linear}(T) = I_{gate} + A_s \frac{W}{L} \left(\frac{k}{q}\right)^2 e^{\frac{q(V_{GS}-V_{th})}{nkT_{ref}}} \times \left({T_{ref}}^2 + (2T_{ref} - \frac{q(V_{GS}-V_{th})}{nk})(T - T_{ref})\right) \quad (7)$$

and the high-order portion $I_{high\_order}(T)$ is

$$I_{high\_order}(T) = I''_{leakage}(T_{ref})(T - T_{ref})^2 + O((T - T_{ref})^3) \quad (8)$$

Therefore, the estimation error resulting from truncation of superlinear terms is bounded as follows:

$$Err = \left|\frac{I_{high\_order}(T)}{I_{leakage}(T)}\right| \quad (9)$$

Equations 8–9 demonstrate that the estimation error of the linear leakage power model is a function of $|T - T_{ref}|$, i.e., the difference between the actual circuit temperature T and the reference temperature T_{ref} at which the linear model is derived. Therefore, to minimize the estimation error, the linear leakage model should be derived as close as possible to the actual subcircuit temperature. This can be

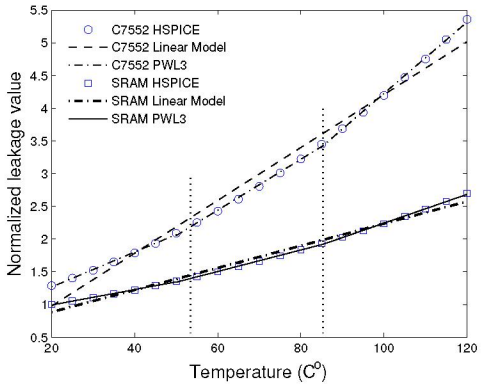

Fig. 2. Normalized leakages for HSPICE, piece-wise linear, and linear models using the 65 nm process for c7552 and SRAM.

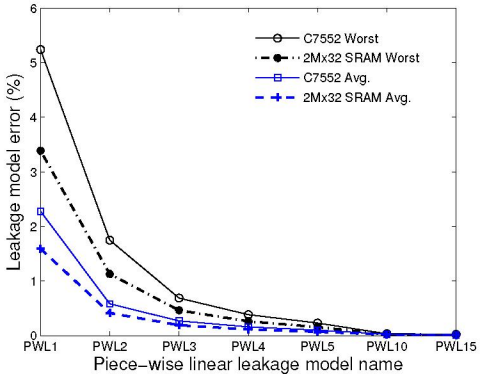

Fig. 3. Linear leakage model errors for c7552 and SRAM.

intuitively understood from Figure 2, which shows the normalized leakage power consumption of two circuits (a combinational circuit benchmark c7552 [14] and SRAM [15]) as a function of temperature. For each circuit, we can compare linear and three-segment piece-wise linear (PWL 3) models with HSPICE simulation results for the 65 nm predictive technology model [16]. Within the normal operating temperature ranges of many ICs, 55°C–85°C, even a linear model is fairly accurate. This accuracy can be further improved by using a piece-wise linear model. Accuracy improves with segment count although, in practice, only a few segments are needed. If a continuous leakage function is available, e.g., via curve fitting to measured or simulated results, the first and second terms of its Taylor series expansion at the average temperature of the IC or subcircuit of interest can be used to provide a derivative-based linear model at the reference temperature of interest.

Figure 3 shows average and maximum leakage model error as functions of piece-wise linear model segment count for the same two circuits considered in Figure 2. Comparisons with HSPICE simulation are used to compute error. Leakage was modeled in the IC temperature range of 25°C–120°C. Within each piece-wise linear region, a linear leakage model is derived at the average temperature of this region using Equation 7. The accuracy permitted by the piece-wise linear model is determined by the granularity of the regions. Figure 3 shows that modeling error decreases as the number of linear segments increases. For three or more segments, the maximum errors are less than 0.69% and 0.47% for c7552 and SRAM, respectively. These results demonstrate that coarse-grained piece-wise linear models permit good leakage estimation accuracy. Finer granularity or differentiation of curve fitted continuous functions will generally further improve accuracy, at the cost of increased complexity.

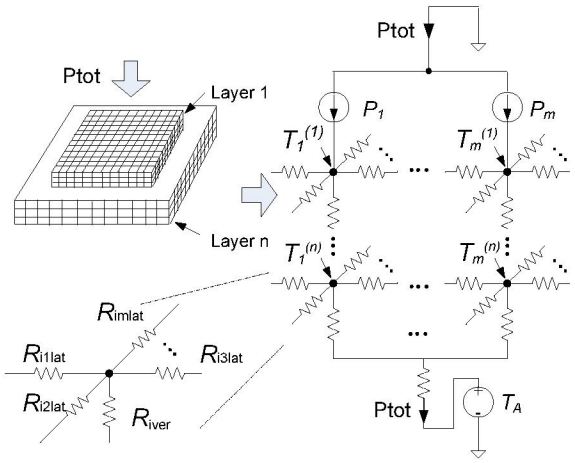

Fig. 4. Steady-state power distribution model.

III. Thermal Model and Properties

This section introduces the thermal model typically used in detailed temperature-aware IC leakage estimation and explains the properties of IC cooling solutions that permit use of the proposed leakage analysis technique.

III.A. Thermal Model Introduction

To conduct numerical thermal analysis, the IC chip and package are partitioned into numerous elements. This permits heat flow to be modeled in the same manner as electrical current in a distributed RC network [17], [18].

$$\mathbf{C}\frac{\mathrm{d}\vec{T}(t)}{\mathrm{d}t} = \mathbf{A}\vec{T}(t) - \vec{p}U(t) \tag{10}$$

where

- $\mathbf{C}$ is an $n \times n$ diagonal thermal capacitance matrix,
- $\mathbf{A}$ is an $n \times n$ thermal conductance matrix,
- $\vec{T}(t) = [T_1 - T_A, T_2 - T_A, \cdots, T_n - T_A]^{\mathrm{T}}$ is the temperature vector in which T_A is the ambient temperature,
- $\vec{p} = [p_1, p_2, \cdots, p_n]^{\mathrm{T}}$ is the power vector, and
- $U(t)$ is a step function.

In steady-state thermal analysis, the thermal profile does not vary with time. Therefore, we can denote $\lim_{t\to\inf}\vec{T}(t)$ as $\vec{T}$, allowing Equation 10 to be simplified as follows:

$$\vec{p} = \mathbf{A} \times \vec{T} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \times \vec{T}$$

The thermal resistance matrix $\mathbf{R}$ is the inversion of thermal conductance matrix, i.e., $\mathbf{R} = \mathbf{A}^{-1}$.

III.B. Insensitivity to Power Profile Claim and Proof

A typical IC thermal model is shown in Fig 4. In order to accurately model spatial temperature variation, several layers of thermal elements are generally necessary between the active layer and heat sink to permit accurate thermal analysis. Assuming an IC floorplan within which the active layer is divided into m isothermal blocks, blk_i, $i \in 1, 2, \cdots, m$, the temperature, area, and power consumption of blk_i are expressed as T_i, s_i, and p_i. The total power consumption of the chip is $P_{tot} = \sum_{i=1}^{m} p_i$. The matrix, $\mathbf{S}$, holds the values of vector $\vec{s}$, $[s_1, s_2, \cdots, s_m, \cdots, s_n]$ along its diagonal. We now prove that a useful property of IC cooling solutions permits use of the proposed leakage estimation technique.

Theorem 1 (Sum of Products Area-Temperature Conservation): For all IC cooling configurations, as long as the total power input is constant, the sum of the IC area-temperature product in the active layer, $\sum_{i=1}^{m} s_i T_i$, is constant if and only if each power source has the same impact on the average temperature of the active layer. That is, the subblock of area-weighted thermal resistance matrix $S \times R$ associated with the active layer should have the equal column sum property. The theorem can also be expressed as follows:

$$\sum_{i=1}^{m} s_i T_i \sim P_{tot} \iff \forall R_j,\ R_j = R_{const} \tag{11}$$

where $R_j = \sum_{i=1}^{m} s_i R_{ij}$. R_{ij} is the ith row and jth column item of the thermal resistance matrix R, and R_{const} is a constant decided by the material and thickness of the chip.

Proof: Assuming the following condition holds,

$$\forall R_j : R_j = R_{const} \tag{12}$$

the sufficiency of the theorem can be proven.

$$\sum_{i=1}^{m} s_i T_i = \sum_{j=1}^{m}\sum_{i=1}^{m} s_i R_{ij} p_j = \sum_{j=1}^{m} R_j p_j \tag{13}$$

According to Condition 12, Equation 13 can be rewritten as:

$$\sum_{i=1}^{m} s_i T_i = R_{const} P_{tot} \tag{14}$$

Therefore, if Condition 12 holds, the sum of each block's area-temperature product $\sum_{i=1}^{m} s_i T_i$ in the active layer keeps constant, as long as the total power input is constant. In particular the sum of area–temperature products, $\sum_{i=1}^{m} s_i T_i = S_{tot} T_{avg}$, i.e., the area-average temperature product of the IC, remains constant.

Next, we prove the necessity of the theorem. If Condition 12 does not hold, the sum of each block's area-temperature product $\sum_{i=1}^{m} s_i T_i$ in the active layer does not remain constant with changing power profile, even if total power consumption is constant. Assume, without loss of generality, there are regions with high and low thermal impact on the active layer: $R_{high} = \sum_{i=1}^{m} s_i R_{ij},\ j \in 1, 2, \cdots, q$, and $R_{low} = \sum_{i=1}^{m} s_i R_{ij},\ j \in q+1, \cdots, m$. The total power can be divided into two parts accordingly, $P_{tot} = P_{high} + P_{low}$. Thus, the sum of area-temperature product can be expressed as follows:

$$\sum_{i=1}^{m} s_i T_i = \sum_{j=1}^{q} R_{high} p_j + \sum_{j=q+1}^{m} R_{low} p_j = R_{high} P_{high} + R_{low} P_{low} \tag{15}$$

Even if P_{tot} is constant, it is clear that a differing ratio between P_{high} and P_{low} makes the sum of area-temperature products different. Necessity is proved. ■

We will show that for a typical multiple layer IC and cooling configuration, the sufficient and necessary conditions for the Theorem 1 are satisfied, based on the following assumptions:

1) All heat generated in the active layer flows eventually to the ambient through the top of the heatsink or the bottom of the package, i.e., no heat flows the sides of the silicon and
2) All layers either have the same area or are isothermal.

We will later demonstrate that these assumptions are well satisfied for a wide range of ICs. Due to space constraints, we can only summarize the proof that these assumptions permit the use of Theorem 1. However, this summary illustrates the reasons for the high accuracy indicated by the results in Section V. We first generate a thermal conductance matrix $\mathbf{A}_j$ for each layer j. $\mathbf{A}_j$ is clearly a real symmetric $m \times m$ matrix, in which the sum of items in the ith row (or column) equals $\frac{k_{con} \cdot s_i}{t_{die}}$, where k_{con} is the silicon thermal conductivity and t_{die} is the thickness of the layer. We transform $\mathbf{A}_j$ to $\mathbf{B}_j$ by factoring the area of each block s_i out of matrix $\mathbf{A}_j$ using matrix $\mathbf{S}_j$. We prove that matrix $\mathbf{B}_j$ has the equal column sum property and that the sum is $\frac{k_{con}}{t_{die}}$. For matrix $\mathbf{M}$, with the equal column sum property, it is easy to prove the following properties.

Given that ζ is an arbitrary set of matrices and β is the set of all matrices having the equal column sum property,

$$\zeta \subseteq \beta \Rightarrow \left(\sum_{\mathbf{M}\in\zeta} \mathbf{M}\right) \in \beta \wedge \left(\prod_{\mathbf{M}\in\zeta} \mathbf{M}\right) \in \beta \tag{16}$$

$$\forall_{\mathbf{M}\in\beta} : \exists \mathbf{M}^{-1} \Rightarrow \mathbf{M}^{-1} \in \beta \tag{17}$$

For the multiple layer case, we can prove that the subblock of area-weighted thermal resistance matrix $\mathbf{S} \times \mathbf{R}$ associated with the active layer can be expressed as a linear combination of matrices $\mathbf{B_j} \in \beta$ from each layer j. In this way, we prove that the Condition 12 is satisfied. We will further validate the sufficient and necessary conditions under realistic cooling configurations in Section V.

IV. Temperature-Aware Leakage Estimation

This section describes the approach conventionally used for temperature-aware leakage estimation and proposes a new accurate and fast technique.

IV.A. Conventional Approach

In the past, most attempts at temperature-aware leakage power consumption estimation used fine-grained thermal analysis to compute leakage power consumption [10], [11]. It can be surmised that this is due to the superlinear relationship between leakage and temperature. After partitioning the IC into thousands of thermal elements, the leakage current for each thermal element is computed based on the corresponding estimated temperature. The total leakage current is computed by taking the sum of the leakage of all thermal elements. Since the number of thermal elements is large, most computation time is spent estimating the detailed thermal profile in the conventional approach. This prevents efficient leakage estimation for many candidate solutions during synthesis or early design space exploration.

IV.B. Proposed Method

In this section, we propose a fast and accurate temperature-dependent leakage estimation method. Assume the IC is divided into n isothermal homogeneous grid elements, blk_i, $i \in 1, 2, \cdots, n$. The temperature, area, and power consumption of each element, blk_i, are expressed as T_i, s_i, and p_i, respectively. Using the linear leakage model developed in Section II, the leakage power of blk_i is expressed as follows:

$$p_{leak}^{blk_i}(T_i) \simeq V_{DD} I_{linear}^{blk_i}(T_i) \tag{18}$$

For a subcircuit with uniform design style, the leakage current is proportional to its area, i.e.,

$$I_{linear}^{blk_i}(T_i) \propto F_i s_i \tag{19}$$

yielding the following formula:

$$I_{linear}^{blk_i}(T_i) = F_i s_i (M_i T_i + N_i) \tag{20}$$

where F_i is the leakage current per unit area. This value depends on manufacturing technology, design style, supply voltage and input pattern. Since input vectors have a great influence on the leakage current, the leakage current should be an input vector probability weighted one. M_i and N_i are parameters obtained by curve fitting in the piece-wise linear model. Collectively, F_i, M_i, and N_i are referred to as *leakage coefficients*. If the derivative model is used, M_i and N_i are calculated at the estimated T_i using the Taylor series expansion technique developed in Section II.

Uniform Case: F_i, M_i, and N_i are decided only by the circuit design style, supply voltage and input pattern. For an IC with uniform design style and supply voltage, such as SRAM and field-programmable gate arrays (FPGAs), these values are the same under specific input patterns for all portions of the IC and can be denoted as F_{tech}, and M and N, respectively. Theorem 1 can be used to show that

$$\sum_{i=1}^{n} I_{linear}^{blk_i}(T_i) = M F_{tech} \sum_{i=1}^{n} (s_i T_i) + F_{tech} N \sum_{i=1}^{n} (s_i) \tag{21}$$

$$= F_{tech} S_{tot} (M T_{avg} + N) \tag{22}$$

Therefore, as long as the conditions necessary to use Theorem 1 are well satisfied, only a few thermal elements are needed for accurate leakage analysis of the entire IC. This permits highly-efficient leakage estimation.

Nonuniform Case: Many ICs are composed of regions with different design styles, e.g., logic and memory, or with different supply voltages. These regions have different F_i, M_i, and N_i values. In this case, we divide the chip into regions, within which the leakage coefficients are consistent. Therefore, the leakage current for region k is expressed as follows:

$$\sum_{blk_i \in reg_k} I_{linear}^{blk_i}(T_i) = M_k F_k \sum_{i=1}^{n} (s_i T_i) + F_k N_k \sum_{i=1}^{n} (s_i)$$

$$= F_k S_{tot} (M_k T_k^{reg} + N_k) \tag{23}$$

where T_k^{reg} is the average temperature of region k. By summing the leakage current of all regions, the total leakage current is obtained. The use of only one, or a few, thermal elements for each region allows extremely fast thermal and leakage analysis.

Multiple thermal elements may also be used in cases for which the IC leakage coefficients are uniform in order to increase estimation accuracy. Finer thermal model granularity implies smaller temperature variations within each thermal element. Recall that the estimation accuracy of a linear model depends on deviation between the actual temperature and the reference temperature at which the linear model was derived. Decreasing the size of a thermal element decreases the temperature variation within it. Therefore, decreasing thermal element size decreases the truncation error resulting from using a linear approximation of the superlinear leakage function. Our results in Section V indicate that, even given pathological power and temperature profiles, very few thermal elements are required for leakage estimation with less than 1% error.

Leakage power consumption influences temperature, which in turn influences leakage power consumption. This feedback can be handled by repeating thermal analysis until convergence. This usually requires only a few iterations for most ICs. More advanced techniques to model this feedback loop may also be devised, but are beyond the scope of this article.

V. Experimental Results

In this section, we evaluate the accuracy and efficiency of the proposed temperature-dependent leakage estimation technique, which consists of *piece-wise linear leakage modeling and coarse-grained thermal analysis*. We characterize the two sources of leakage estimation error introduced by this technique: truncation error as a result of using a linear leakage model and temperature error as a result of using a coarse-grained thermal model. The base case for comparison is conventional temperature-aware leakage estimation using *superlinear leakage model and fine-grained thermal analysis*. Our experiments demonstrate that for a set of FPGA, SRAM, microprocessor, and application specific integrated circuit (ASIC) benchmarks, the proposed leakage modeling technique is accurate and permits great increases in efficiency. All benchmarks were run on an AMD Athlon-based Linux PC with 1 GB of RAM.

V.A. Experimental Setup

We use the 65 nm predictive technology model [16], for leakage modeling. This model characterizes the impact of temperature on device leakage. We first derive the superlinear leakage model using

TABLE I
LEAKAGE ERROR FOR FPGA

T_{avg} (°C)	P_{tot} (W)	DM error Avg. (%)	DM error Max. (%)	CPU time SF (s)	CPU time DM (μs)	Speedup (million ×)
40	10	0.003	0.005	16.1	10	1.60
50	40	0.039	0.092	14.7	10	1.47
60	70	0.122	0.258	16.1	10	1.61
70	110	0.300	0.650	16.2	10	1.62
80	150	0.505	0.960	16.2	9	1.79
90	180	0.731	1.205	16.0	9	1.78

HSPICE simulation. The piece-wise linear leakage model is then derived using the method described in Section II: partitioning the temperature range into uniform segments and using least-squared error fitting for each segment. The derivative-based model is based on the first two terms of the Taylor series expansion of the superlinear leakage function around the reference temperature of interest.

We use HotSpot3.0 [19] for both coarse-grained and fine-grained steady-state thermal analysis. HotSpot3.0 supports both block-based coarse-grained and grid-based fine-grained stead-state thermal analysis. Previous work [20] demonstrated that the coarse-grained block-based method is fast. In contrast, fine-grained grid-based partitioning is slower but permits more accurate thermal analysis. In this work, coarse-grain thermal analysis is based on the block-based method, as only the average block temperature is required. For fine-grained thermal modeling, we partition the IC active layer into 100×100 elements. This resolution is necessary; decreasing resolution to 50×50 resulted in a 6 °C error in peak temperature for the Alpha 21264. A resolution of 100×100 elements is also sufficient for our benchmarks; we have used resolutions up to 1,000×1,000 to validate our results and have found that increasing resolution beyond 100×100 has little impact on temperature estimation accuracy.

V.B. Leakage Power Estimation

Table I shows the accuracy and speedup resulting from using the proposed leakage estimation technique on an FPGA [21]. We used six sets of 30 random power profiles. Six different total power consumptions (Column 2) resulting in different average temperatures (Column 1) were considered. Power profiles were generated by assigning uniformly-distributed random samples ranging from [0, 1] to each cell in a 5×5 array overlaying the IC and then adjusting the power values to reach the target total IC power while maintaining the ratios of power consumptions among cells.

In Section IV we show that the leakage power of an IC with uniform leakage coefficients depends only on total power consumption. To evaluate this claim, we compare the superlinear fine-grained model (SF) with the single-element linear derivative-based model (DM). At each total power setting, the average estimation error for the 30 randomized power profiles is shown in Column 3. As shown in Column 4, the maximum estimation error was never greater than 1.2%. As shown in Columns 5–7, the speedup permitted by our technique ranges from 1,470,000×–1,790,000×. This speedup results from a reduction in thermal model complexity that greatly accelerates the thermal analysis portion of leakage estimation.

In addition to considering modeling accuracy for uniform leakage coefficients in the presence of randomized power profiles, we designed a power profile to determine the error of the proposed technique under pathological conditions. In this configuration, all of the power in the IC is consumed by a corner block and other blocks consume no power. The total power input is set to 117 W, leading to an extremely unbalanced thermal profile. Temperatures ranged from 52.85 °C to 106.85 °C. This case goes well beyond what can be expected in practice, but serves to establish a bound on the estimation error of the proposed approach.

Figure 5 shows the leakage estimation error as a function of thermal modeling granularity for piece-wise linear thermal models with various numbers of segments and a linear model based on

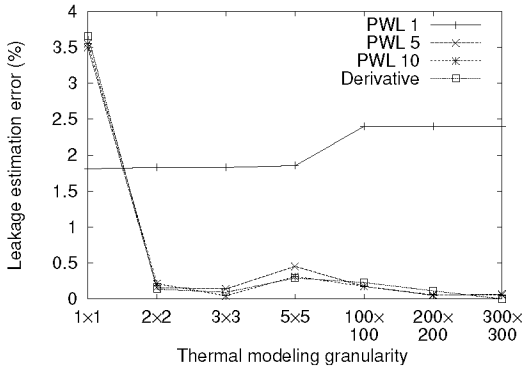

Fig. 5. Leakage estimation error of FPGA under worst-case power profile.

TABLE II
LEAKAGE ERROR FOR ALPHA 21264

Benchmark		gcc	equake	mesa	gzip	art	bzip2	twolf
Error (%)	PWL 5	0.52	0.71	0.53	0.42	0.34	0.45	0.65
	DM	0.54	0.64	0.51	0.48	0.56	0.47	0.57
Speedup (thousand ×)		59	67	65	81	66	67	66

the derivative of the continuous leakage function at the block's predicted temperature. Using the same one-segment linear model for all blocks (PWL 1) results in approximately 2% estimation error. However, piece-wise linear models with five or more segments, and the derivative-based model, all maintain errors of less than 0.5%, as long as at least four thermal elements are used. Note that the derivative based model is not identical to a piece-wise linear model in which the number of segments approaches infinity because the piece-wise linear model is fit to the leakage function using a least-squared error minimizer while the derivative based model is based on the Taylor series expansion around a single temperature. Therefore, it is possible for the piece-wise linear model to result in higher accuracy in some cases. From these data, we can conclude that even when faced with extreme power profiles, only a few thermal elements are necessary to permit high leakage power estimation accuracy.

In addition to considering ICs with uniform design styles, e.g., FPGAs, we have evaluated the proposed technique when used on the Alpha 21264 processor, an IC having regions with different sets of leakage coefficients, e.g., control logic, datapath, and memory. Power traces were generated using the Wattch power/performance simulator [22] running SPEC2000 programs. One thermal element is used for each functional unit in the processor. Table II shows results for five-segment piece-wise linear (PWL 5) and derivative-based (DM) leakage models. Row 4 shows that reducing thermal model complexity results in leakage estimation speedups ranging from 59,259×–80,965×. As Rows 2 and 3 show, derivative-based and piece-wise linear model leakage estimation errors are less than 1% for all benchmarks, compared with an HSPICE-based superlinear leakage model used with fine-grained thermal analysis. This small error has two components: truncation error resulting from coarse-grained thermal modeling and slight deviation of real cooling structures from the conditions stated in Theorem 1. We now discuss the conditions required by Theorem 1.

V.C. Thermal Model Error Breakdown

In Section III, we showed that the necessary and sufficient conditions for Theorem 1 hold under reasonable assumptions. IC cooling structures approximately conform to the assumptions required for sum of products area-temperature conservation to hold, e.g., much more heat leaves an IC and package through the heatsink than through the sides of the silicon die. However, they do not perfectly conform, e.g., some heat can leave the system through the sides of the die.

We now evaluate the error resulting from approximating the

TABLE III
$\sum_{i=1}^{n} s_i T_i$ WITH DIFFERENT POWER PROFILES

T_{avg} (°C)	FPGA SATP Error (%)		SRAM SATP Error (%)		EV6 SATP Error (%)		HP SATP Error (%)	
	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.
40	0.0016	0.0019	0.0013	0.0018	0.0002	0.0003	0.0202	0.5407
50	0.0057	0.0075	0.0097	0.0131	0.0099	0.0115	0.0085	0.1458
60	0.0099	0.0113	0.0189	0.0247	0.0180	0.0204	0.0139	0.2116
70	0.0145	0.0169	0.0280	0.0361	0.0263	0.0302	0.0168	0.2093
80	0.0178	0.0217	0.0337	0.0472	0.0338	0.0389	0.0177	0.1788
90	0.0224	0.0282	0.0424	0.0570	0.0421	0.0514	0.0215	0.1913

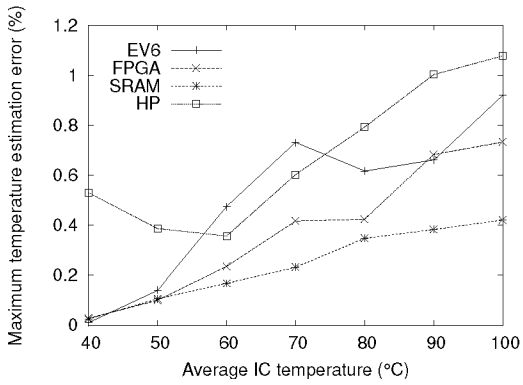

Fig. 6. Thermal error breakdown among different types of ICs.

conditions required to use Theorem 1. We use several ICs with differing floorplans: FPGA, SRAM [23], Alpha 21264, and HP, an ASIC benchmark from MCNC benchmark suite [24], to compare sum of area-temperature products (SATP) values given different power profiles. For each IC, SATP is calculated for 30 randomized power profiles, which are generated in same way as those for Table I. Each IC has a different area. Therefore, total power consumption values were chosen to produce each of the six reported average temperatures. Table III shows maximum and average differences between the SATP values for the random power profiles and the SATP value for a uniform power profile. From these results, we can conclude that the SATP error is less than 0.6% for all four benchmark ICs. We also computed SATP error for the unbalanced worst-case power FPGA profile used in Figure 5. The worst-case error is smaller than 0.015% for all thermal model granularities. We conclude that the conditions required to use Theorem 1 are well-satisfied for a wide range of ICs.

Although we have shown that the properties required to use Theorem 1 are well-approximated for a number of ICs, we have yet to show the implications of this observation upon temperature estimation accuracy. We partition the IC into blocks, each of which corresponds to a region with uniform leakage coefficients, and compare the average block temperatures with those calculated by using a fine-grained thermal model. Figure 6 shows the maximum temperature estimation error as a function of average IC temperature for the same set of benchmarks shown in Table III. Error is computed on the Kelvin scale. Figure 6 shows that the maximum temperature estimation error over all power profiles is less than 1.1%. For the Alpha 21264 processor we also calculated the temperature differences using power traces from SPEC2000 applications. In all cases, the average temperature difference is less than 0.61%. From this, we can conclude that using a coarse-grained thermal model is sufficient for IC leakage power consumption estimation.

VI. CONCLUSIONS

This article has presented an extremely fast and accurate method of estimating IC leakage power consumption during design and synthesis. This idea allows a speedup of 59,259×–1,790,000× while maintaining accuracy compared with a conventional temperature-aware leakage estimation technique using a detailed thermal model. The proposed technique's accuracy is proven based on two observations: (1) leakage may be accurately modeled as a linear function of temperature over the operating temperature ranges of real functional units and (2) given a fixed total power consumption, the average temperature of an IC active layer is mostly independent of the power distribution. Its accuracy is further validated via numerous comparisons with results from detailed thermal modeling. The proposed technique can easily be used in commercial or academic synthesis and design flows in order to accelerate accurate temperature-dependent leakage power consumption estimation.

REFERENCES

[1] "International Technology Roadmap for Semiconductors," 2005, http://public.itrs.net.

[2] S. Naffziger, et al., "The implementation of a 2-core, multi-threaded itanium family processor," *J. Solid-State Circuits*, vol. 41, no. 1, pp. 197–209, Jan. 2006.

[3] J. A. Butts and G. S. Sohi, "A static power model for architects," in *Proc. Int. Symp. Microarchitecture*, Dec. 2000, pp. 191–201.

[4] S. M. Martin, et al., "Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads," in *Proc. Int. Conf. Computer-Aided Design*, Nov. 2002, pp. 721–725.

[5] S. Narendra, et al., "Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18 CMOS," *J. Solid-State Circuits*, vol. 39, no. 2, pp. 501–510, Feb. 2004.

[6] Y. F. Tsai, et al., "Characterization and modeling of run-time techniques for leakage power reduction," *IEEE Trans. VLSI Systems*, vol. 12, no. 11, pp. 1221–1232, Nov. 2004.

[7] A. Abdollahi, F. Fallah, and M. Pedram, "Leakage current reduction in CMOS VLSI circuits by input vector control," *IEEE Trans. VLSI Systems*, vol. 12, no. 2, pp. 140–154, Feb. 2004.

[8] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[9] Y. Zhang, et al., "HotLeakage: A temperature-aware model of subthreshold and gate leakage for architects," Univ. of Virginia, Tech. Rep., May 2003, CS-2003-05.

[10] H. Su, et al., "Full chip leakage estimation considering power supply and temperature variations," in *Proc. Int. Symp. Low Power Electronics & Design*, Aug. 2003, pp. 78–83.

[11] W. P. Liao, L. He, and K. M. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 7, pp. 1042–1053, July 2005.

[12] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. IEEE Press, 2001.

[13] K. M. Cao, et al., "BSIM4 gate leakage model including source-drain partition," in *IEDM Technology Dig.*, Dec. 2000, pp. 815–818.

[14] "ISCAS85 benchmarks suite," http://www.visc.vt.edu/~mhsiao/iscas85.html.

[15] F. Zhang, "System-level leakage power modeling methodology," Dept. of Electronics Engg., Tsinghua University," Bachelor's Degree Thesis, July 2006.

[16] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *Proc. Int. Symp. Quality of Electronic Design*, Mar. 2006, pp. 585–590.

[17] G. S. Ohm, "The Galvanic circuit investigated mathematically," 1827.

[18] J. Fourier, *The Analytical Theory of Heat*, 1822.

[19] K. Skadron, et al., "Temperature-aware microarchitecture," in *Proc. Int. Symp. Computer Architecture*, June 2003, pp. 2–13.

[20] W. Huang, et al., "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. VLSI Systems*, vol. 14, no. 5, pp. 501–524, May 2006.

[21] I. C. Kuon, "Automated FPGA design verification and layout," Ph.D. dissertation, Dept. of Electrical and Computer Engg., University of Toronto, July 2004.

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

[23] "SRAM layout," SRAM link at http://www.eecs.umich.edu/UMichMP/Presentations.

[24] "MCNC benchmarks suite," http://www.cse.ucsc.edu/research/surf/GSRC/MCNCbench.html.