A Fast Full-Chip Static Power Estimation Method

Jiachun Wan*, Hai Wang*, Sheldon X.-D. Tan†, Chi Zhang*, Yuan Yuan‡, Keheng Huang§, and Zhenghong Zhang§
*School of Microelectronics & Solid-State Electronics,
University of Electronic Science & Technology of China, Chengdu, 610054 China
†Department of Electrical Engineering, University of California, Riverside, CA 92521 USA
‡School of Automation Engineering, University of Electronic Science & Technology of China, Chengdu, 610054 China
§Southwest China Research Institute of Electronic Equipment, Chengdu, 610036 China

***Abstract*—As IC technology advances, leakage current induced static power has become the major obstacle for chip to achieve high performance. Fast estimation of full-chip static power is difficult because static power depends nonlinearly on temperature. In this paper, we propose a new fast full-chip static power estimation method. The new method uses Taylor expansion based approximation to linearize the nonlinear leakage-temperature dependency locally. A new linear thermal model is formulated based on the local linearization for easy full-chip transient simulation of temperature and static power. Experiments show that the new method achieved up to $10\times$ speedup against traditional iteration based method with high estimation accuracy.**

I. Introduction

At the early years of IC technology advancing, Dennard scaling is the key to the success of Moore's law [1], which states that power density of the chip stays constant as transistor size scales because current and voltage scale along with transistor size. However, Dennard scaling has broken down since around the year of 2006 with increased power density. The major reason behind this is the static power issue caused by leakage: leakage does not scale with transistor size and its resulting static power becomes significant in the nanometer technology years. What is even worse is that leakage is highly dependent on temperature, so the static power will cause the chip to heat up and further increase static power itself. As a result, static power is one of the most important limiting factors of chip performance today.

Estimating static power is especially important for dynamic thermal management (DTM), which makes power adjustment decisions in order to enhance chip performance under thermal constraint. While dynamic power distribution can be easily estimated by performance counter based methods, estimating static power at runtime is challenging, because leakage's complex dependence on temperature makes the static power-temperature relation nonlinear. Iterative methods are proposed to handle such nonlinearity. Although these methods are considered to be accurate, they perform temperature calculation using thermal model multiple times in the iteration, which makes them slow, especially when full-chip power estimation is needed.

In this paper, we propose a fast full-chip static power estimation method. The new method employs thermal sensor assisted Taylor expansion technology to handle the nonlinearity problem without iteration. Through testing on multi-core chip, the new method has similar accuracy against the standard iteration based method.

This work is supported in part by National Natural Science Foundation of China under grant No. 61404024, in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, in part by the Open Foundation of State Key Laboratory of Electronic Thin Films and Integrated Devices under grant No. KFJJ201409.

II. Background

A. Static power basics

Power of chip p is composed of dynamic power p_d and static power p_s. Static power p_s, caused by leakage current I_{leak} as $p_s = V_{dd} I_{leak}$ and is independent of the activity of the chip, is hard to be estimated. This is because the subthreshold current I_{sub}, which is the main component of leakage current, has a complex relation with temperature, modeled in BSIM 4 model [2] as (also apply $V_{DS} \gg v_T$ [3])

$$I_{sub} = K {v_T}^2 e^{\frac{V_{GS}-V_{th}}{\eta v_T}} (1 - e^{\frac{-V_{DS}}{v_T}}) \approx K {v_T}^2 e^{\frac{V_{GS}-V_{th}}{\eta v_T}}, \quad (1)$$

where $v_T = \frac{kT_p}{q}$ is the thermal voltage and T_p is a scalar representing temperature at one place, K and η are process related parameters, and V_{th} is the threshold voltage.

From (1), it is easy to see that the static power distribution depends mainly on the distribution of temperature of the chip. Since temperature also depends on power, in order to view the whole picture, thermal model is used to describe temperature's dependency on power as shown next.

B. Thermal model: from power to temperature

Thermal model of the chip is used to calculate full-chip temperature by providing power distribution information, and it is usually formed as the following equations as

$$\begin{aligned} GT(t) + C\frac{dT(t)}{dt} &= B_c P(T,t), \\ Y(t) &= LT(t), \end{aligned} \quad (2)$$

where $T(t) \in \mathbb{R}^n$ is the temperature vector (distinguished from T_p, a scalar representing temperature at only one place), representing temperatures at n places (called *thermal nodes*) of the chip and package; $G \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{n\times n}$ contain

978-1-4673-9713-6/$31.00 ©2015 IEEE

equivalent thermal resistance and capacitance information respectively; $B_c \in \mathbb{R}^{n\times l}$ stores the information of how powers are injected into the thermal nodes; $P(T,t) \in \mathbb{R}^l$ is the power vector, which contains power consumptions of l components on chip, including dynamic power P_s and static power P_d, i.e., $P(T,t) = P_s(T,t) + P_d(t)$, reminding that static power $P_s(T,t)$ is actually a function of temperature T; $Y(t) \in \mathbb{R}^m$ is the output temperature vector, containing only temperatures of thermal nodes that is interested by the user, for example, thermal nodes on the chip only (excluding package thermal nodes); $L \in \mathbb{R}^{m\times n}$ is the corresponding selecting matrix.

In order to calculate transient temperature (static power) using (2), we can discretize it, for example using backward Eular's method with time step h as

$$(\frac{C}{h}+G)T(t+h) = \frac{C}{h}T(t)+B_c(P_d(t+h)+P_s(T,t+h)), \quad (3)$$

so $T(t+h)$ and $P_s(T,t+h)$ need to be calculated provided $T(t)$ at the previous time point is calculated.

C. Iteration based static power estimation

Because static power depends on temperature, (3) is a nonlinear equation, and as a result, $T(t+h)$ cannot be calculated directly. Traditionally, iterative method is used to solve such equation. First, based on the process technology used, an initial guess of $P_s(T,t+h)$, $P_s^0(T,t+h)$, is provided, and temperature $T(t+h)^0$ is calculated using (3) with such initial guess. Then, the static power is updated as $P_s^1(T,t+h)$ using (1) with $T(t+h)^0$, and the iteration goes on until $\|P_s^{i-1}(T,t+h) - P_s^i(T,t+h)\| < \epsilon$, where ϵ is the tolerance.

The major problem of the iteration based method is the computing time. For full-chip static power estimation, system in (3) is large, and solving (3) many times at each time step makes the simulation time to be long.

III. Fast Full-Chip Static Power Estimation

In this work, we propose a novel non-iteration based fast full-chip static power estimation method, which resolves the long computing time problem of the iteration based method.

A. Local linearization of subthreshold current

Since the major difficulty of calculating static power comes from the nonlinearity relation of subthreshold current with temperature, as shown in (1), the basic idea of the new method is to linearize (1) locally to avoid iteration. Specifically, we perform Taylor expansion of (1) at the temperature point T_{p_0}, and ignore the terms with order higher than two, resulting in the linearized I_{sub}, denoted as I_{lin}:

$$\begin{aligned} I_{lin} =& K(\frac{k}{q})^2 e^{\frac{q(V_{GS}-V_{th})}{\eta k T_{p_0}}} \\ & \times ({T_{p_0}}^2 + (2T_{p_0} - \frac{q(V_{GS}-V_{th})}{\eta k})(T_p - T_{p_0})). \end{aligned} \quad (4)$$

It is obvious that I_{lin} is a linear function of T_p, and it is an approximation of I_{sub}. The approximation is good when T_p is close to T_{p_0}. From previous research, it has been shown that such approximation has high accuracy for common temperature ranges of chip (from $55°$C to $85°$C) [3], [4].

B. Formulating the linear thermal model with static power

Since we have already linearized the relation of subthreshold voltage and temperature, we can rewrite the power and temperature relation in a linear form. In order to do that, we need to integrate (4) into (2). Please note that (4) is in scalar form but (2) is in vector/matrix form. So we first rewrite (4) in vector/matrix form by collecting and accumulating scalars I_{lin} and T_p at multiple positions of the chip into vectors, then change the current variables to powers by multiplying voltage V_{dd}. Rewriting from (4), the resulting linearized static power representation in vector/matrix form is

$$P_s = P_0 + A_s T, \quad (5)$$

where $P_0 \in \mathbb{R}^l$ is a known vector, with each element formed by terms not associated with T_p in (4) at each position of the chip. $A_s \in \mathbb{R}^{l\times n}$ is a known rectangular diagonal matrix (the left $l \times l$ block matrix is diagonal representing the chip, and the right $l \times (n-l)$ block matrix is all zeros representing the package), with each diagonal element formed by the coefficient associated with T_p in (4) at each position of the chip.

Integrating (5) into (2), we have

$$(G - B_c A_s)T(t) + C\frac{dT(t)}{dt} = B_c(P_d(t) + P_0). \quad (6)$$

It is obvious that we have successfully obtained a linear thermal model considering static power and eliminated the nonlinear relation of static power and temperature. Now, transient simulation of (6) is as straightforward as in (3) by viewing "$G - B_c A_s$" as the new "G" matrix, and $P_d(t) + P_0$ as the new "$P(T,t)$" vector.

C. Selecting the proper Taylor expansion point

Since both P_0 and A_s are formed by the Taylor expansion point T_{p_0}, we now discuss how to determine T_{p_0}. It is well known that as a property of Taylor expansion, linear approximation (4) (also the equivalent (5) and (6)) is accurate if the nonlinearity around the expansion point T_{p_0} is weak and the actual temperature T_p (or T in vector form) is close enough to T_{p_0}. As a result, in order to ensure the approximation accuracy, we want the expansion point T_{p_0} to be close to the actual T_p.

However, if we update T_{p_0} every time T_p changes, the computing cost will rise (but still much faster than iteration based method as shown in the experiments). This is because transient simulation of the linear thermal model needs the LU factorization of $(G - B_c A_s)$. Since matrix A_s depends on the Taylor expansion point T_{p_0}, LU factorization has to be re-performed if Taylor expansion point changes.

In order to balance the accuracy and computing cost, we divide the common chip temperature range into multiple regions, with each region length determined by the strength of the nonlinearity in that region. One expansion point is assigned to each region at the middle point. If T_p is inside the ith region, then the corresponding expansion point is used as T_{p_0}. This means if current T_p is in the same region as the previous one,

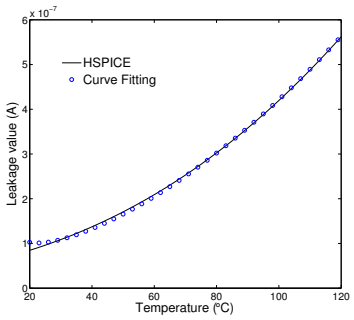


Fig. 1. Leakage of a TSMC 65 nm process MOSFET from HSPICE and it's curve fitting using (1).

then there is no need to re-perform the LU decomposition. Of course, T_p is an unknown variable, so its value is estimated by the thermal sensor reading from the sensor near the T_p position.

IV. Experimental Results

First, we use TSMC 65 nm process model to characterize the impact of temperature on device leakage through HSPICE simulation. Using the simulation data, we obtain the parameters of model (1) through curve fitting with results shown in Fig. 1. Then, we combine the leakage model (1) with thermal model (2) extracted from HotSpot for transient thermal/power analysis with SPEC bechmarks [5]. The simulation time step h is chosen to be 0.01s. The chip tested has dimension of $10mm \times 10mm$ with 16 cores, and there is one thermal sensor placed on each core. We also partition each core into 5×5 thermal blocks for fine-grained analysis.

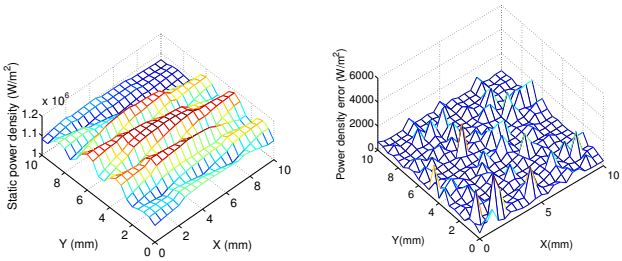
Table I shows the accuracy of estimated static power and temperature using the proposed method compared with those using the traditional iteration method. Since accuracy and computing time differ with how temperature regions are predetermined for Taylor expansion points, we test the lower and the upper bounds by using two extreme case for temperature regions: one is to use infinite number of temperature regions (LU decomposition is updated for each time point, best accuracy, but worst computing time), another is by using only one region (one LU decomposition for all time points, worst accuracy, best computing time) for all temperatures. What is more, in order to consider the impact of the error caused by thermal sensors, two kinds of sensor noise are tested: one has standard normal distribution $N(0,1)$, which models unbiased sensor measurements (marked as "Normal" in Table I), another one has distribution $N(10,0)$, which models the biased sensor measurement (marked as "Worst" in Table I). From the table, it is clear that even for the slowest case (region # as ∞), the new method achieved $3.4\times$ speedup with excellent accuracy. While for the fastest case (region # as 1), our new method is $10.4\times$ faster still with reasonable accuracy. In addition, the new method keeps a high accuracy while a seriously biased sensor measurement (as in "Worst") is used.

The full-chip static power density map snapshot taken from one transient time point is given in Fig. 2. We take the

TABLE I
Accuracy and speed comparison. "Ite" means iteration method, and "new" means the proposed method. Time is reported for computing every 100 time steps.

Region #	Temp Err (°C)				Power Err (%)				Time (s)		Speedup
	Worst		Normal		Worst		Normal		Ite	New	
	Max	Avg	Max	Avg	Max	Avg	Max	Avg			
∞	0.98	0.66	0.22	0.01	2.92	2.20	0.26	0.04	2.18	0.65	$3.4\times$
1	1.34	0.91	1.34	0.91	3.91	3.20	3.91	3.20		0.21	$10.4\times$

results from the traditional iteration-based method (shown in Fig. 2(a)) as standard, and plot the error map of the new method in Fig. 2(b). It is clear that static power estimation by the new method is accurate all across the chip.



(a) Static power density map calculated by the iteration method, which serves as golden reference.

(b) Static power density error of the proposed method.

Fig. 2. Comparison of static power density map given by two methods.

V. Conclusion

In this paper, we have demonstrated the new fast full-chip static power estimation method. The new method uses Taylor expansion based local linearization technology to avoid the time consuming iterations. A new linear thermal model is also formulated for easy transient simulation of temperature and static power. We have tested the new method on multi-core chip with SPEC benchmarks, and the results show that the new method is up to $10\times$ faster than traditional iterative method with high accuracy.

References

[1] H. Esmaeilzadeh *et al.*, "Dark silicon and the end of multicore scaling," *IEEE Micro*, vol. 32, no. 3, pp. 122–134, May 2012.

[2] W. Liu *et al.*, "BSIM 4.0.0 technical notes," EECS Department, University of California, Berkeley, Tech. Rep. UCB/ERL M00/39, 2000. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2000/3863.html

[3] Y. Liu *et al.*, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *Proc. European Design and Test Conf. (DATE)*, 2007, pp. 1–6.

[4] S. Sarangi, G. Ananthanarayanan, and M. Balakrishnan, "LightSim: A leakage aware ultrafast temperature simulator," in *Proc. Asia South Pacific Design Automation Conf. (ASP-DAC)*, 2014, pp. 855–860.

[5] J. L. Henning, "SPEC CPU 2000: Measuring CPU performance in the new millennium," *IEEE computer*, vol. 1, no. 7, pp. 28–35, July 2000.