Full Chip Leakage Estimation Considering Power Supply and Temperature Variations

Haihua Su Frank Liu Anirudh Devgan Emrah Acar Sani Nassif
IBM Corp.
11400 Burnet Rd.
Austin, TX 78758
{haihua,frankliu,devgan,emrah,nassif}@us.ibm.com

ABSTRACT

Leakage power is emerging as a key design challenge in current and future CMOS designs. Since leakage is critically dependent on operating temperature and power supply, we present a full chip leakage estimation technique which accurately accounts for power supply and temperature variations. State of the art techniques are used to compute the thermal and power supply profile of the entire chip. Closed-form models are presented which relate leakage to temperature and VDD variations. These models coupled with the thermal and VDD profile are used to generate an accurate full chip leakage estimation technique considering environmental variations. The results of this approach are demonstrated on large-scale industrial designs.

Categories and Subject Descriptors

I.6.3 [**Simulation and Modeling**]: Applications

General Terms

Verification, Algorithms

Keywords

leakage power, supply voltage variation, thermal analysis

1. INTRODUCTION

With continuous shrinking of minimal feature size, leakage current is expected to become a major challenge for future CMOS designs. Although leakage is about 10% of total chip power for the current generation of CMOS technologies, the number is expected to rise to 50% for next generation techniques [9]. The increasing leakage current not only poses a problem for battery-powered devices such as mobile and hand-held electronics, it is increasingly critical for active operation as it is becoming higher percentage of total power.

Most leakage estimation and reduction techniques have focused on subthreshold leakage (I_{sub}) due to the lowering of the power supply voltage and the accordingly reduction of the threshold voltage. With the reduction of the gate oxide thickness, the gate leakage current (I_{gate}) can no longer be ignored. Gate leakage is on trend to become comparable to the subthreshold leakage [9]. Full-chip leakage estimation is needed for both gate and sub-threshold leakage.

Some methods have been reported to estimate the full-chip leakage. For example, the authors of [6] use a linear regression model to estimate full-chip leakage based on the gate count in the ASIC environment. In [7], a method is proposed to include the effect of within-die process variation. It is well-known that the leakage current has strong dependency on the environmental factors, such as channel temperature, supply voltage and workload. As will be shown in this paper, the leakage has super-linear dependency on temperature: a 30 ℃ change in the temperature will affect the leakage by 30%. Its dependency on supply voltage is exponential, a 20% fluctuation on V_{dd} can affect the leakage by more than 2X. Even more importantly, in today's complex industrial designs, both temperature and V_{dd} fluctuations have very strong locality, i.e., they are nonuniform across the chip. The exact amount of the fluctuations at certain location depends on the distribution of the transistors and decoupling capacitors, the workload, as well as the quality of the power grid and package design. Assuming a uniform temperature and V_{dd} distribution in full-chip leakage estimation is too simplistic thus inaccurate.

In this paper, we present a full-chip leakage modeling technique with accurate consideration of both realistic temperature and V_{dd} fluctuations. To our knowledge, this is the first report on this topic. We use state-of-the-art numerical algorithms to calculate the full-chip V_{dd} and temperature profiles. The results are then coupled with a close-form model, which relates leakage to temperature and V_{dd} changes, to provide an accurate full-chip leakage estimation. Since the change of the leakage may in turn affects the V_{dd} and temperature profiles, iterations are performed to remove the pessimism. Although we focus on the average leakage and average dynamic power in this paper, the method can be extended to take into account the effect of workload, by using per-cycle dynamic and leakage power of each macro. The method has been implemented and applied on a contemporary industrial design on 0.13μm CMOS SOI technology.

ISLPED'03, August 25–27,2003,Seoul, Korea.
Copyright 2003 ACM 1-58113-682-X/03/0008 ...$5.00.

The results show that a simple assumption of uniform temperature and supply voltage variation can underestimate the full chip leakage by as much as 30%.

The rest of the paper is organized as follows. In Section 2, we present our chip-level leakage estimation flow. Full-chip power grid analysis and on-chip thermal analysis methodologies will be discussed in Section 3 and Section 4 respectively. We then present our leakage power model in Section 5 and dynamic power model in Section 6. Experimental results are shown in Section 7, followed by the conclusion in Section 8.

2. LPT: FULL CHIP LEAKAGE ESTIMATION CONSIDERING POWER SUPPLY NOISE AND TEMPERATURE VARIATIONS

Fig. 1 illustrates our leakage estimation flow considering power supply noise and temperature variations. We have developed a fast and efficient tool which is capable of performing full-chip power grid IR-drop as well as thermal analysis. Once we know the voltage drop and temperature variation at each macro, we can adjust the originally estimated power, both dynamic and leakage component. On the other hand, power grid voltage drop and temperature change also depends on the power consumption (both dynamic and leakage power) of the circuit, which is the source of the voltage and temperature variation. A complete analysis of this nonlinear coupling behavior often requires Newton-Raphson iteration, which is typically not practical for current large-scale integrated circuits. Instead, an iteration-based approach gives acceptable accuracy at significantly improved efficiency.

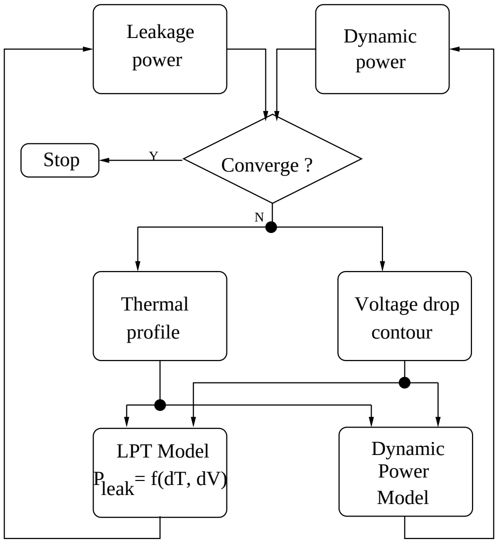

Figure 1: LPT flow.

3. POWER SUPPLY MODELING AND ANALYSIS

Various chip-level power grid methodologies published in the literature [2, 4, 10] de-couple the linear (power grid) and nonlinear portion (transistors) of the whole system as follows: First, the total power of each macro (or block) consisting of nonlinear devices is estimated assuming perfect power supply voltages (V_{dd} and G_{nd}). Usually average work load with reasonable switching factor is used to calculate the total power. An average leakage power can also be calculate for each block or macro. Next, these independent current sources (total power divided by V_{dd}) are applied to the power grid. Based on this scenario, a general power distribution network model looks as follows:

- The power grid is modeled as a resistive mesh with via resistors connecting metal layers.
- The loads (blocks or macros) are modeled as distributed independent current sources in parallel with parasitic capacitors connected between power and ground.
- The decoupling capacitors (decaps) are modeled as single lumped capacitors connected between power and ground.
- The top-level metal is connected to a package modeled with inductors or RL elements connected to ideal constant voltage sources.

In leakage power estimation, we are only interested in DC voltage drop across the whole chip. Therefore the entire network is reduced to multiple layers of close-coupled resistive meshes. If more accuracy is desired, a resistive package model can be attached between the top-level metal layer and ideal voltage sources modeling the voltage regulator. The network therefore becomes a large-scale linear circuit as shown in Fig. 2, in which the package, VDD grid and GND grid each stands for a large resistive mesh. The size of a typical power distribution circuit can have millions of nodes. Because of its size, traditional numerical analysis methods can easily run out of memory or extremely slowly.

In our implementation, the iterative algebraic multi-grid (AMG) solver is used. It works directly on matrix stamps and hierarchically creates a coarsened grid with a reduced number of nodes, whose exact solution can be obtained very efficiently. The solution at the coarsest grid is then mapped back to the fine grid, with a restricted number of iterative solve to reduce the high frequency error component produced during the reduction and interpolation process [1]. With AMG, we can solve power grid with multi-million nodes within a couple of minutes.

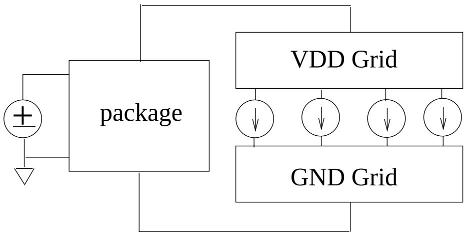

Figure 2: Power supply network modeling.

Given the voltage solution at every grid point and the set of supply and ground nodes a block is attached to, an average compression voltage drop between the supply and ground of this block can be obtained. The total leakage power of the block is updated according to this voltage drop value, based on our leakage model to be introduced in Section 5.

4. THERMAL MODELING AND TEMPERATURE SIMULATION

Similar to power grid analysis, the electrical (power) and thermal simulations can be de-coupled to compute the chip's thermal profile [3, 11].

A general 3D thermal analysis involves solving the heat conduction equation

$$\rho c_p \frac{\partial T(x,y,z,t)}{\partial t} = \nabla[k(x,y,z,T)\nabla T(x,y,z,t)] + g(x,y,z,t) \quad (1)$$

subject to the general boundary condition

$$k(x,y,z,T)\frac{\partial T(x,y,z,t)}{\partial n_i} + h_i T(x,y,z,t) = f_i(x,y,z) \quad (2)$$

where T is the temperature, g is the power density of heat sources, k is the thermal conductivity, ρ is the density of the material, c_p is the specific heat, h_i is heat transfer coefficient on the boundary, $f_i(x,y,z)$ is a function of position and n_i is the outward direction normal to surface i. In steady-state analysis, $\frac{\partial T}{\partial t} = 0$. Also, within the range of working temperature, the thermal conductivities k of various materials inside a chip (silicon, silicon dioxide, metals and ILDs) can be regarded as constants. Therefore, Eqn. (1) becomes

$$k\nabla^2 T(x,y,z) + g(x,y,z) = 0, \quad (3)$$

where $g(x,y,z)$ is the power density of devices at the surface of the silicon layer, including both the dynamic and leakage power.

Depending on the type of package (locations of heat sinks) and the surrounding environment, the following three types of chip boundary conditions (BC) can be derived from Eqn. (2) [3]:

1) Isothermal (Dirichlet) BC: $T = f_i(x,y,z)$, where $f_i(x,y,z)$ corresponds to temperatures at heat sinks. Generally the heat sink is attached to the back-side of the substrate.

2) Insulated (Neumann) BC: $\frac{\partial T}{\partial n_i} = 0$, where n_i corresponds to directions normal to the four side surfaces of the chip assuming they are perfectly insulated.

3) Convective (Robin) BC: $k_i \frac{\partial T}{\partial n_i} = h_i(T - T_a)$, where T_a is the ambient temperature. This condition is needed to derive an accurate heat sink and package thermal model.

Finite-difference technique is often applied to solve the above heat conduction equation (Eqn.(3)) with boundary conditions. Accordingly, an equivalent thermal resistive network can be constructed [5]. Assume the thermal conductivity is k, the typical thermal resistance of a cube with volume of $dx \cdot dy \cdot dz$ in the x direction is:

$$R_i = \frac{dx}{kdydz}, \quad (4)$$

and the resistance at the convective boundary with heat transfer coefficient h_b is

$$R_b = \frac{1}{h_b dydz} \quad (5)$$

Based on the above equation, a full-chip thermal model can be constructed which includes all the layers as well as the heat sinks and C4's. For a typical commercial chip, the size of the problem can also be quite big. Once again, AMG solver is a natural choice to solve it.

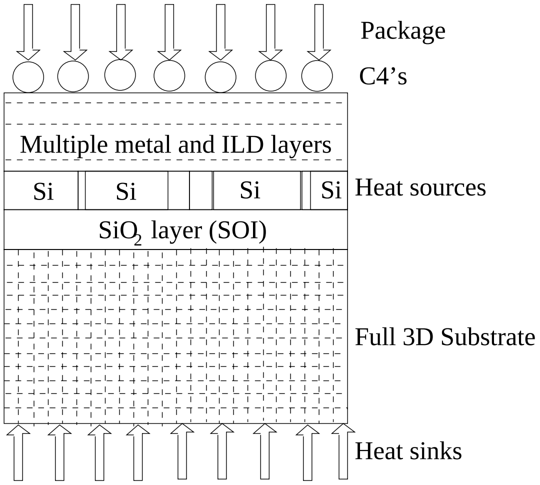

Figure 3: Thermal modeling of a chip with C4 package.

While a full 3D full chip model could result in a huge system of equations, various simplification techniques have been developed to simplify the analysis while still maintaining sufficient amount of accuracy to the temperature solution at the silicon (device) layer, where the temperature variation is to be used to estimate the circuit leakage power. A summary of the simplification techniques we have applied to our chip structure shown in Fig. 3 (cross-section view) is as follows:

1) Mixed 1D and 3D thermal modeling, similar to [3]. First, a full 3D substrate model is applied to increase the accuracy. Second, the package and heat sinks are treated as 1D thermal resistances.

2) Dense devices are assumed to occupy the entire silicon layer. According to this model, the thermal resistors in this layer are calculated using the thermal conductivity of silicon.

3) Equivalent thermal resistance modeling in the metal layers, ILD (inter-layer dielectric) layers and C4's. The equivalent thermal resistance in the metal is adjusted according to the metal density of the metal layer(s). Similarly, the equivalent thermal resistance in ILD is adjusted according to the via density between adjacent metal layers or the contact density between the lowest metal layer and devices. The thermal resistance at C4's can be estimated using the technique introduced in [8].

4) Ideal temperature is assumed within heat sinks and the package.

Similar to the case of power supply, given the temperature at each volume on the device layer, an average temperature variation among all volumes that a block is attached to can be obtained. The total leakage power of this block is to be updated according to this temperature value, based on our leakage model to be introduced in the next section.

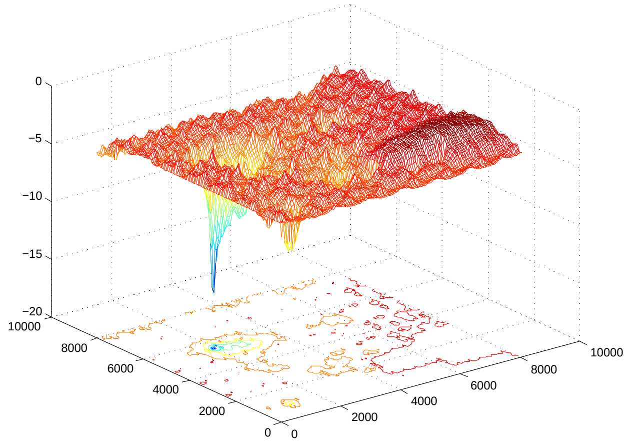

Figure 4: Voltage drop contour plot. Z-axis is the percentage change

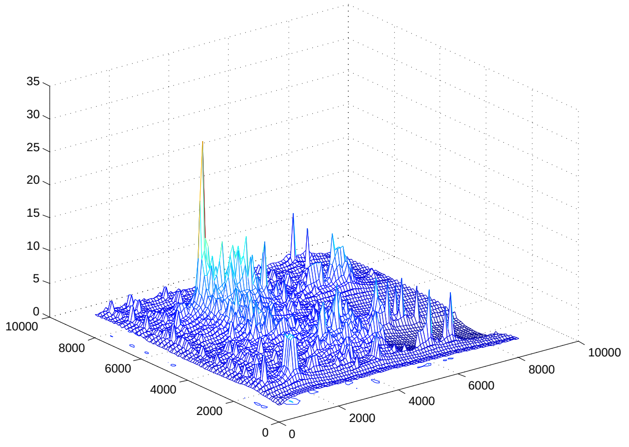

Figure 5: Full chip temperature increase profile.

5. LEAKAGE POWER MODELING

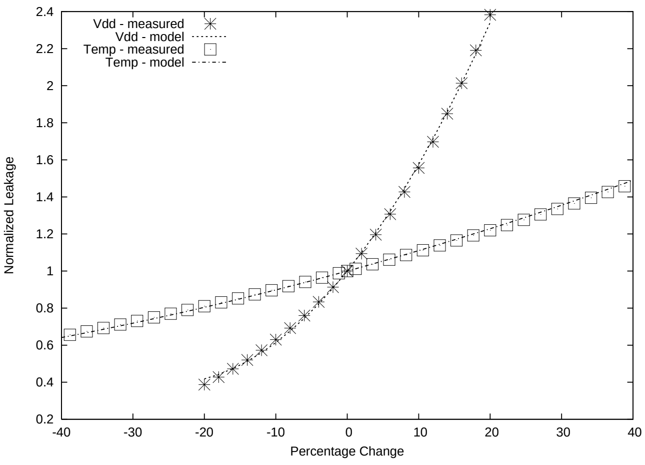

Figure 6: Temperature and Vdd leakage model compared with measured data.

To create an accurate leakage model with respect to temperature and supply voltage fluctuation, we use SPICE to simulate standard cells with accurate BSIM SOI device models. Both I_{sub} and I_{gate} are included in the simulation. Each cell is simulated at different temperatures and supply voltages. The average leakage at each temperature-V_{dd} node is calculated. Accurately, the V_{dd} dependency of the leakage is exponential and its temperature dependency is super-linear. However, since the power supply variations are typically no more than 20 or 30% of the nominal power supply, it can be modeled as a polynomial around its nominal value. We use a second-order polynomial to describe the dependencies. The coefficients of the polynomial is calculated by regression. The model is in the form of:

$$\frac{I_{leak}(\Delta T,\Delta V)}{I_{leak}(0,0)} = 1 + a_1 \cdot \Delta T + a_2 \cdot (\Delta T)^2 + b_1 \cdot \Delta V + \\ + b_2 \cdot (\Delta V)^2 + c_2 \cdot (\Delta T)(\Delta V) \qquad (6)$$

For different standard cells, the coefficients in the model are slightly different, but we observed that the difference is very small. The resulting model is verified using ISCAS benchmark C17, which is shown in Fig. 6. The figure clearly shows exponential and super-linear dependencies of the leakage on V_{dd} and temperature. It demonstrates that our model is quite accurate in the given range of fluctuations.

6. DYNAMIC POWER MODELING

Dynamic power also changes as the V_{dd} changes through the circuit. For this work, we assume the dynamic power is independent of the temperature and use the following simple model to update the dynamic power when V_{dd} changes:

$$P_s = P_{s0} \cdot (1 + \Delta V/V_{dd})^2 \qquad (7)$$

7. EXPERIMENTAL RESULTS

In this section we present the analysis result of several industry chips. We first demonstrate the performance of our power grid and thermal analyzer, then show the leakage estimation results on two industry designs. The leakage estimation flow, along with the power grid and thermal simulation engine, has been implemented in C++. All the experiments are run on an Intel Pentium-III 700MHz machine with 4GB memory, running Linux OS.

matrix size	# non-zeros	analysis type	CPU (sec)	Mem (GB)
0.17M	1.12M	TH	82.13	0.45
0.27M	1.76M	TH	139.17	0.61
0.63M	3.11M	IR	88.39	0.46
1.74M	8.89M	IR	293.58	1.3
2.73M	13.9M	IR	438.10	2.1

Table 1: Runtime performance and memory usage of our power grid and thermal analysis engine. IR is power grid analysis and TH represents thermal analysis.

The CPU time and memory usage of our power grid and thermal analysis tool on several chips are listed in Table 1. It can be seen that the AMG method discussed in Section 3 performs very well.

The next two designs are based on a 0.13 μm commercial CMOS SOI technology. The first chip (chip1) has approximately 160K macros, with the size around 8mm×8mm. The initial total chip power is 48Watts, out of which 9.6W (20%) is roughly estimated as leakage. The second example (chip2) is the CPU core of a microprocessor design. It occupies the area of 2.5mm×4.7mm with total power 5.6Watts, out of which approximately 1.12W is leakage. The supply voltage for each chip is 1.2V and 1.0V respectively. The total change of leakage power due to the temperature and voltage variation for both chips are listed in Table 2.

chip	ΔV (mV)	ΔT (℃)	$\Delta leakage$ (W)
1	min: -4 max: -184	min: -4.2 max: +25.3	-1.850
2	min: 0 max: -41	min: -9.5 max: +4.1	-0.136

Table 2: Leakage variation after one iteration.

Fig. 4 and Fig. 5 shows the supply voltage drop and the temperature distribution across chip1. The compression supply voltage between the power and ground plane is plotted. The figure shows the variation ranges from 3% to 15% of V_{dd}. Across the chip, the temperature variation (compared to the reference temperature at heat sinks) ranges from 0.8 to 30.3 ℃ . From the two plots it is easy to identify several "hot" spots in terms of both power supply voltage and temperature variations. They both correspond to the high power density regions.

Applying our LPT model on each functional block, we update the leakage power based on its average supply voltage and temperature changes. From our model, we observe that after one iteration, leakage power of each block becomes less than its initial value. Therefore, the ratio in the first iteration is less than one for each block; the farther it is away from 1, the larger the leakage varies.

Fig. 7, 8 and 9 show the distribution of blocks over the leakage variation ratio, defined by the updated leakage and the initial estimate, for the three leakage models: both voltage and temperature dependent (LPT), individual voltage dependent (LV) and individual temperature dependent (LT). A comparison of the three diagrams shows that the leakage depends more on the power supply voltage variation than on the temperature variation, which confirms our simple analytical estimation earlier.The dependency on both supply voltage and temperature variations is closely correlated and the overall effect brings the leakage down from the initial estimate.

Fig. 10 shows the leakage variation distribution across chip1. Clearly the large leakage variation regions correspond to the "hot" spots identified in the voltage and temperature profiles shown in Fig. 4 and Fig. 5 respectively. Similar observation can be made for chip2. The leakage variation distribution across chip2 is shown in Fig. 11. Note that although both chips are designed for the same technology, they have completely different leakage profiles. Simply counting the number of gates will not be able to capture such differences.

Table 3 shows total leakage using various estimation methods after one iteration of update from the initial value. In these methods, leakage of each block is updated based on the supply voltage and temperature variation of each individual block. For comparison reason, we list the results of the traditional method using uniform voltage and temperature profile. We assume a uniform 10% Vdd drop and a uniform 85℃ profile (zero spatial temperature variation) across the chip. The numbers are listed as "EMP" in the table. Because it blindly assumes a flat Vdd and temperature profile, it underestimates the full-chip leakage by 30%.

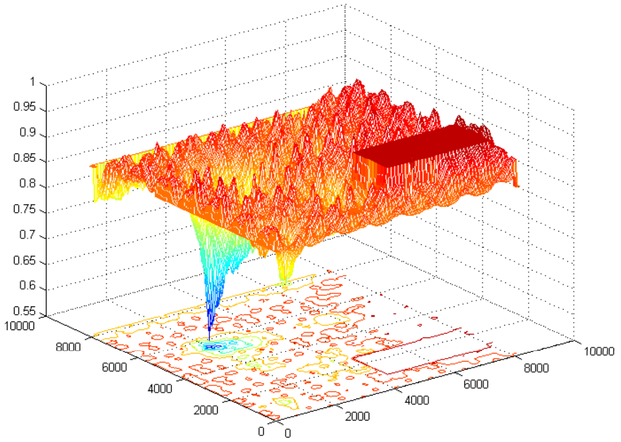

Figure 10: Leakage variation distribution of chip1.

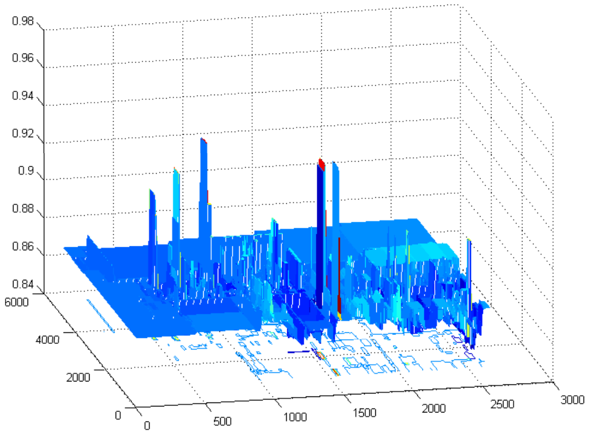

Figure 11: Leakage variation distribution of chip2.

Fig. 12 shows the update of leakage power for 5 iterations. The first iteration reduces leakage by 19.2%. After the first iteration, which corrects the leakage by 19.2%, the rest iterations only fine-tune the result within 0.5%. Therefore, usually one iteration can provide sufficiently accurate results.

8. CONCLUSION

We have proposed an accurate full-chip leakage estimation methodology accounting for both nonuniform supply voltage and temperature variations. An incremental leakage model has been proposed and successfully applied into our methodology. We have demonstrated the significance of voltage and temperature effects on leakage power on industry designs.

9. REFERENCES

[1] W. L. Briggs. *A Multigrid Tutorial.* SIAM, 1987.

[2] H. Chen and D. Ling. Power supply noise analysis

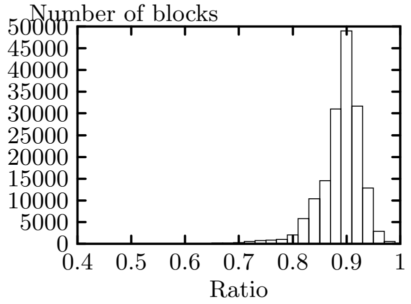

Figure 7: Histogram of $P_{leak} = f(\Delta V, \Delta T)$ variation ratio across chip1.

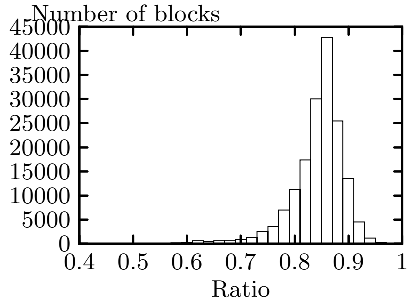

Figure 8: Histogram of $P_{leak} = f(\Delta V)$ variation ratio across chip1.

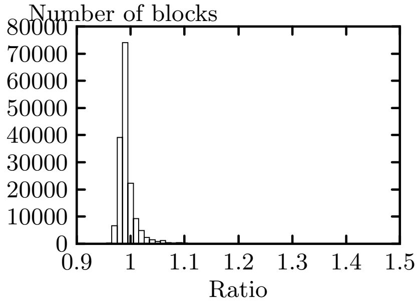

Figure 9: Histogram of $P_{leak} = f(\Delta T)$ variation ratio across chip1.

Methods	ΔV (mV)	ΔT (°C)	Total leakage(W)
LPT	min: -4 max: -184	min:-4.2 max:+25.3	7.75
LP	min: -4 max: -184	n/a	7.77
LT	n/a	min:-4.2 max:+25.3	9.63
EMP	-120	0	5.31

Initial leakage estimate = 9.6 (W)

Table 3: Comparison among various leakage estimation methods for chip1.

methodology for deep-submicron vlsi chip design. In *Proceedings of DAC*, 1997.

[3] Y. K. Cheng, C. C. Teng, A. Dharchoudhury, E. Rosenbaum, and S. M. Kang. A chip-level electrothermal simulator for temperature profile estimation of cmos vlsi chips. In *International Symposium on Circuits and Systems*, 1996.

[4] A. Dharchoudhury, R. Panda, D. Blaauw, and R. Vaidyanathan. Design and analysis of power distribution networks in powerpc microprocessors. In *Proceedings of DAC*, 1998.

[5] F. P. Incropera and D. P. Dewitt. *Fundamentals of Heat and Mass Transfer*. John Wiley & Sons, 2001.

[6] R. Kumar and C. P. Ravikumar. Leakage power estimation for deep submicron circuits in an asic design environment. In *International Conference on VLSI Design*, 2002.

[7] S. Narenda and *et al.* Full-chip sub-threshold leakage power prediction model for sub-0.18 μm cmos. In *Proceedings of ISLPED 2002*, 2002.

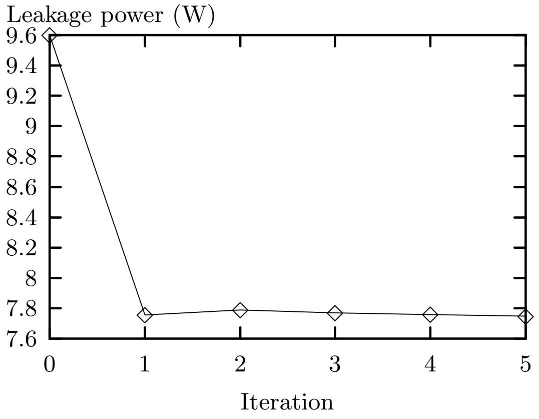

Figure 12: Leakage power per iteration for chip1.

[8] J. Parry, H. Rosten, and G. B. Kronmann. The development of component-level thermal compact models of a C4/CBGA interconnect technology: The Motorola PowerPC 603 and PowerPC 604 RISC microprocessors. *IEEE Trans. Components, Packaging, and Manufacturing Technology*, March 1998.

[9] Semiconductor Industry Association. *The International Technology Roadmap for Semiconductors*, 2001.

[10] G. Steele, D. Overhauser, S. Rochel, and S. Z. Hussain. Full-chip verification methods for dsm power distribution systems. In *Proceedings of DAC*, 1998.

[11] T.-Y. Wang and C. C.-P. Chen. Thermal-adi: A linear-time chip level dynamic thermal simulation algorithm based on alternating-direction-implicit (adi) method. In *International Symposium on Physical Design*, 2001.