Conference version. Final version will be available in the PATMOS proceedings published by Springer in the LNCS series.

Statistical Modeling and Analysis of Static Leakage and Dynamic Switching Power

Howard Chen, Scott Neely, Jinjun Xiong,
Vladimir Zolotov, and Chandu Visweswariah

IBM Research Division, Thomas J. Watson Research Center,
1101 Kitchawan Road, Yorktown Heights, New York 10598-0218, U.S.A.
{haowei,sneely,jinjun,zolotov,chandu}@us.ibm.com
http://www.research.ibm.com/da

Abstract. As the device size continues to shrink and circuit complexity continues to grow, power has become the limiting factor in today's microprocessor design. Since the power dissipation is a function of many variables with uncertainty, the most accurate representation of chip power or macro power is a statistical distribution subject to process and workload variation, instead of a single number for the average or worst-case power. Unlike statistical timing models that can be represented as a linear canonical form of Gaussian distributions, the exponential dependency of leakage power on process variables, as well as the complex relationship between switching power and workload fluctuations, present unique challenges in statistical power analysis. This paper presents a comprehensive case study on the statistical distribution of dynamic switching power and static leakage power to demonstrate the characterization and correlation methods for macro-level and chip-level power analysis.

Key words: Statistical power analysis

1 Introduction

The advent of continued device scaling and increasing process variability has contributed to the growing popularity of statistical timing analysis, which not only replaces the traditional process-corner-based approach used in static timing analysis, but also revolutionalizes the way that chips are designed and verified today. In statistical timing analysis, all timing quantities such as gate delays, wire delays, arrival times, slews (rise/fall times) and slacks are represented by a canonical first-order delay model [1]:

$$a_0 + \sum_{i=1}^{n} a_i \Delta X_i + a_{n+1} \Delta R_a \ , \tag{1}$$

where a_0 is the mean or nominal value, ΔX_i represents the variation of the i^{th} global source of variation X_i from its nominal value, ΔR_a is the variation of an independent random variable R_a, and a_i is the sensitivity to each of the sources of variation. By scaling the sensitivity coefficients, the random variables ΔX_i and ΔR_a can be assumed to have a normalized Gaussian distribution $N(0,1)$. The capabilities of parameterized block-based statistical timing analysis in [1] have since been extended by [2] to handle non-Gaussian parameters and nonlinear delay functions.

Like static timing methodology, most traditional power analysis methodologies are deterministic and corner based, where only selected cases such as the nominal case, best (-3σ) case, and worst $(+3\sigma)$ case are analyzed. However, when the worst-case assumption is made for each random variable, the corner-based approach is inherently pessimistic. In order to avoid parametric yield prediction based on fully correlated and overly pessimistic corner points, statistical methods have been developed to model leakage power due to process variability. For example, an empirically-fit exponential quadratic equation is proposed in [3] to represent the subthreshold current as a function of channel length and estimate its probability density function. The mean and variance of the leakage current for the entire circuit can then be obtained by adding the lognormal distribution of leakage current from individual gates. A full-chip analysis of both the subthreshold leakage and gate tunneling leakage is described in [4] by considering spatial correlation due to intra-chip variations.

Although simplified lognormal models have been developed to estimate the leakage power distribution, the industry-standard BSIM [5] device models generally cannot be easily adapted to the analysis of process variability without loss of accuracy. Furthermore, the statistical characterization of power should include not only the static leakage power, but also the dynamic switching power. When the macro power is characterized deterministically by an average or worst-case number, designers often have little information about the true distribution of power that could potentially lead to thermal or yield problems. For example, a macro might consume $0.3W$ of power on average, but in any given cycle or state, this macro could operate with an idle power of $0.1W$ or a peak power of $0.5W$ under a wide range of switching factors (Fig. 1). Therefore it is important to look beyond one deterministic number that has traditionally characterized the power, and provide circuit designers an insight into the statistical distribution of power due to both process and workload variations. Such power analysis is useful for yield prediction, maximization of battery life, prediction of on-chip thermal gradients, power distribution design, decoupling capacitance allocation, eletromigration analysis, etc.

In addition, it has been shown that in leakage dominated technologies, the leakage power can cause the yield window to shrink by imposing a two-sided constraint on the window [6]. The correlation between power and performance due to their dependence on common process variables could have a significant impact on yield, especially in high-frequency bins [7]. In this paper, we will present a case study to analyze the statistical distribution of not only static leakage power, but also dynamic switching power, so that we can accurately estimate and optimize the parametric yield by finding the joint probability density function of both power and delay.

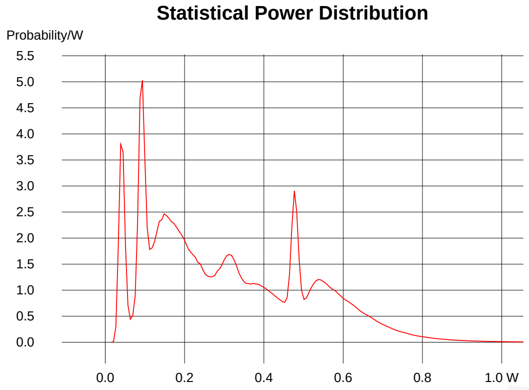

Fig. 1. Probability density function of normalized power.

2 Statistical Distribution of Leakage Power

The total standby Current I_{DDQ} of a CMOS transistor is comprised of two major components: the device subthreshold current I_{OFF} and the gate tunneling current I_{GATE}. Therefore the analysis of total leakage power must include both the channel leakage due to device subthreshold current I_{OFF} and the gate leakage due to tunneling current I_{GATE}. The leakage model described in this paper is a hardware-based model that has been extracted from various experiments in a pre-determined process window. After a model is fitted to the hardware measurements, it will accurately characteristize the leakage current of the corresponding device such as an NFET, a PFET, or an SRAM cell.

For our 45-nm SOI technology, the device subthreshold current I_{OFF} is an exponential function of the channel length L_P, the supply voltage V_{DD}, and the temperature T. The exponential function $I_{OFF}(L_P, V_{DD}, T)$ not only captures the charge-sharing and drain-induced barrier lowering effects, but also considers I_{OFF} variation due to threshold voltage (V_T) scattering and narrow channel effects. Fig. 2 shows the probability density function of subthreshold current due to channel length variation in a typical device. Although the variation of channel length L_P could be modeled as a Gaussian distribution, the corresponding subthreshold current variation is not a Gaussian distribution. For example, the long tail of Fig. 2 illustrates that I_{OFF} could increase by a factor of 4 due to channel length variation.

In addition to subthreshold current I_{OFF}, the gate tunneling current I_{GATE} for the thin-oxide devices could also be a significant contributor to the total leakage current on the chip. As depicted in Fig. 3, the gate tunneling current includes the current between gate and source/drain diffusion through the channel region (I_{gcs} and I_{gcd}), the current between gate and source/drain diffusion through the overlap region (I_{gos} and I_{god}), and the current between gate and body (I_{gb}).

For our 45-nm SOI technology, the gate tunneling current I_{GATE} is a linear function of the channel length L_P, but an exponential function of the gate oxide

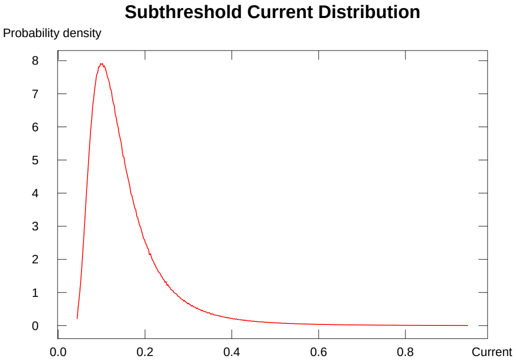

Fig. 2. Probability density function of normalized subthreshold current.

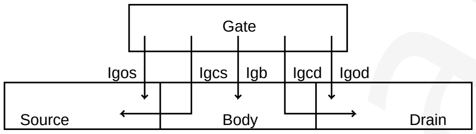

Fig. 3. Components of gate tunneling current.

thickness T_{OX}, the supply voltage V_{DD}, and the temperature T. Fig. 4(a) illustrates the probability density function of the normalized gate tunneling current due to channel length and oxide thickness variations in an $NFET$. Although the red curve in Fig. 4(a) shows that the gate tunneling current due to channel length (L_P) variation alone is a Gaussian distribution, the closely matched T_{OX} blue curve and $T_{OX} + L_P$ green curve clearly demonstrate that the statistical distribution of gate tunneling current is dominated by the variation of oxide thickness. As the T_{OX} decreases, I_{GATE} increases exponentially.

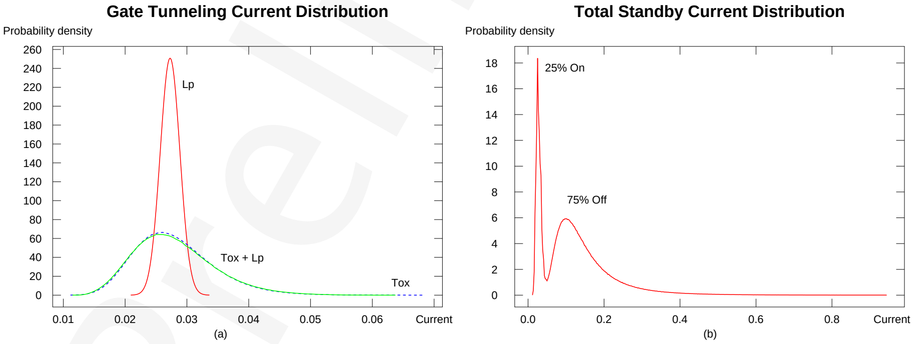

Fig. 4. Probability density function of gate tunneling current and total standby current.

Since gate leakage occurs when the gate is on and channel leakage occurs when the channel is off, Fig. 4(b) shows one statistical distribution of normalized I_{DDQ} when the gate is turned on 25% of the time and the channel is turned off 75% of the time. Based on the time-averaged channel and gate leakage state for each device reported by the circuit simulator, the probability of each leakage state not only reflects a device's average time in the on or off state, but also considers circuit topology effects such as device stacking.

3 Statistical Distribution of Switching Power

An accurate analysis of chip power is imperative for high-performance microprocessor design to understand system requirements and ensure system reliability. Therefore, in addition to the estimation of leakage power under process variation, the distribution of switching power due to workload variation should also be included in a statistical power analysis. The common power analysis methodology that we have developed to estimate switching power starts with transistor-level simulation of each on-chip macro. As a building block of the chip design hierarchy, the macro can be as simple as an I/O buffer, or as complex as a cache or multiplier. After generating the net list for each macro, adding input vectors and output loads, and capturing the current waveforms from circuit simulation, a comprehensive power analysis can be performed to extract device current models and specific circuit power characteristics.

Our circuit simulation is based on an event-driven circuit simulator [8], which allows current waveform integration on the fly and greatly reduces the output file size. Technology-dependent data such as device models, temperature and voltage parameters, and clock cycle time are used during circuit simulation. In addition, over 100 parameters are typically specified in a project file, which contains information such as signal timing and capacitive loads.

The process of creating a circuit net list can be run from both the schematic and layout views. The physical layout extraction not only identifies all the transistors and their parasitic capacitance, but also inserts current meters at the junction contacts between the transistors and power nets to collect detailed current distribution data under various operating conditions.

After the raw net list is extracted, appropriate input stimuli and output loads are added to generate the final net list. Voltage sources are applied at the primary inputs to represent the correct input vectors that satisfy all circuit and logic constraints and provide sufficient coverage for power, noise, and reliability analysis. The input vectors can be categorized into different states such as ramp-up, clock-gated, hold (idle), functional (average), and peak power. For example, the ramp-up cycles serve to flush random data through the circuit to initialize its state. After the circuit is initialized, all inputs are held constant for several cycles, except for the global clock signals, to allow the circuit to reach its inactive state where the idle or hold power can be determined. The hold power is a measure of the clock-related power, which is often regarded as the power of the most common state. Finally, workload-based random input vectors are applied to the data input nodes to measure the average functional power. A range of switching factors is applied to provide coverage for not only power calculation, but also noise and electromigration analysis.

The generation of the final net list is further controlled by a configuration file, which includes global circuit information such as voltage, temperature, clock cycle, signal timing, and output loads. The correct arrival time of input signals is extracted from the timing file and data-type constraints are assigned to the input nodes. Once all the macros have been simulated, the current data can be collected during the hold cycle, average-current cycle, and peak-current cycle for macro-level and chip-level power analysis. Fig. 5 shows the scatter plot of 141 data points for the switching current of a macro with 79,082 NFETs and 79,921 PFETs. The relationship between switching current and input switching factor can be approximated by a linear, quadratic, or higher-order polynomial regression. In Fig. 5, about 80% of the data points are scattered between the upper and lower regression lines, which represent the 90th and 10th percentile of switching current respectively.

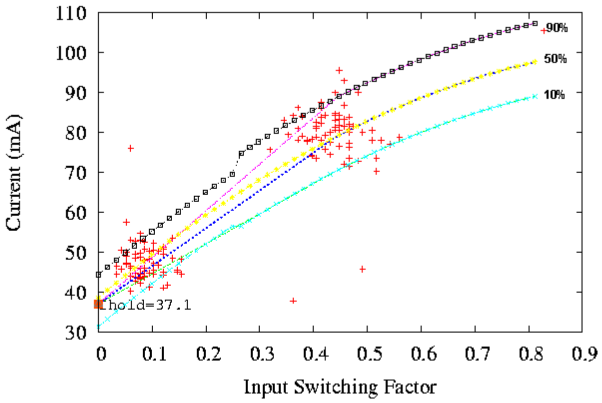

Fig. 5. Regression model of switching current as a function of switching factor.

Fig. 6 shows the probability density function of macro power when there is a 10% probability that the macro operates with a switching factor about 0.5, a 10% probability that the macro operates with a switching factor about 0.1, a 30% probability that the macro is idle ($SF = 0$ with clock running), and a 50% probability that the macro is clock-gated (with leakage only). For the two clusters of data points where $SF \neq 0$, their input switching factors are assumed to follow two relatively narrow Gaussian distributions $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, where the means μ_1 and μ_2, and standard deviations σ_1 and σ_2 are determined by the data in Fig. 5. Similarly, the variations of channel length and gate oxide thickness are assumed to have Gaussian distributions during leakage calculation.

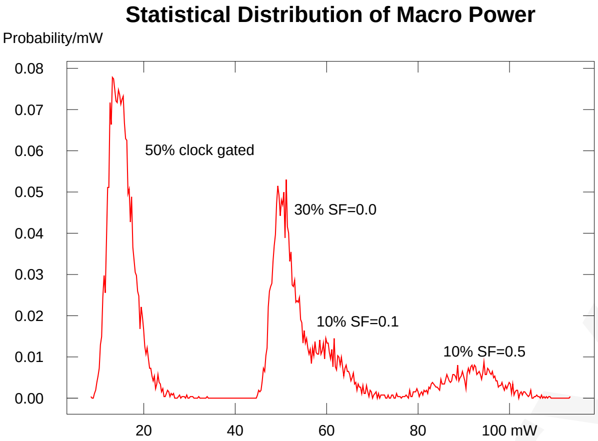

Fig. 6. Probability density function of total macro power.

4 Experimental Results

A statistical power analysis has been performed on selected benchmark macros in our 45-nm technology. For gate tunneling leakage calculation, the oxide thickness T_{OX} is assumed to have a Gaussian distribution $N(\mu_T, \sigma_T)$, where μ_T is the mean and σ_T is the standard deviation of gate oxide thickness. To take into account the inter-chip and intra-chip variation, the oxide thickness is modeled as $T_{OX} = \mu_T + \sigma_T(\alpha \Delta T_{global} + \beta \Delta T_{local})$, where ΔT_{global} is a normalized Gaussian distribution due to global chip-to-chip variation, ΔT_{local} is a normalized Gaussian distribution due to local intra-chip variation, and $\alpha^2 + \beta^2 = 1$. In our case study below, α is set to 0.8 and β is set to 0.6.

Similarly, process variations such as gate lithography, etch bias, and lateral source/drain diffusion, result in channel length variation. For subthreshold leakage calculation, the channel length L_P is assumed to have a Gaussian distribution $N(\mu_L, \sigma_L)$, where μ_L is the mean and σ_L is the standard deviation of physical channel length. To take into account the inter-chip and intra-chip variation, the total channel length variation is modeled as $L_P = \mu_L + \sigma_{CHIP} \cdot \Delta L_{CHIP} + \sigma_{ACLV} \cdot \Delta L_{ACLV}$, where ΔL_{CHIP} is a normalized Gaussian distribution due to chip mean variation, ΔL_{ACLV} is a normalized Gaussian distribution due to across-chip line-width variation, and $\sigma^2_{CHIP} + \sigma^2_{ACLV} = \sigma^2_L$. Fig. 7 shows the probability density functions of gate leakage, subthreshold leakage, and total leakage power for a macro with 1725 NFETs and 1586 PFETs. The global and local variations of oxide thickness and channel length are considered at the transistor level and included in the respective leakage power distribution.

To model the correlated and independent randomness of switching power, the statistical power distribution of each macro must first be characterized by its probability density function PDF and cumulative distribution function CDF. The power of macro i can then be determined by the inverse function of CDF_i, where $P_i = CDF_i^{-1}(X_i)$, and $0 \le X_i \le 1$. Depending on how the switching

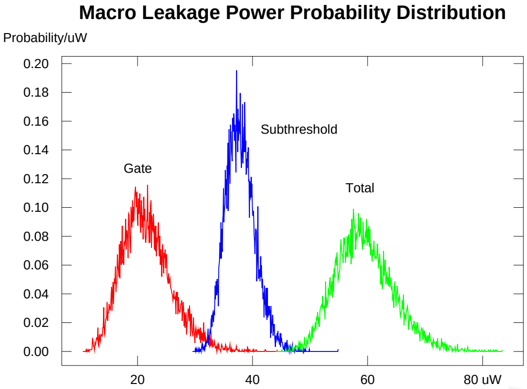

Fig. 7. Probability density function of total leakage power.

activities of macro i are associated with workload j, the random variable X_i will be assigned a CDF value based on the equation:

$$X_i = \begin{cases} \sum_{j=1}^{n} a_{ij} \cdot W_j + b \cdot M_i & \text{if } \mathrm{a_{ij}} > 0 \text{ (positive correlation)} \\ \sum_{j=1}^{n} |a_{ij}| \cdot (1 - W_j) + b \cdot M_i & \text{if } \mathrm{a_{ij}} < 0 \text{ (negative correlation)} \end{cases} \quad (2)$$

where W_j is the CDF value that corresponds to the power fluctuation of workload j, a_{ij} is the sensitivity of the dynamic power of macro i to workload j, M_i is the CDF value that corresponds to the non-workload-dependent power variation of macro i, and b is the sensitivity of the dynamic power of macro i to its own random variation, subject to the constraints that $b \geq 0$ and $(\sum_{j=1}^{n} |a_{ij}|) + b = 1$. Since CDF values are assigned to W_j and M_i, both variables assume a standard uniform distribution $U(0,1)$.

In order to provide maximum generality and flexibility to model macros with different patterns of switching activities, Monte Carlo simulation with a sample size of 10,000 is used for statistical power analysis. It takes about 10 CPU hours to simulate a large chip with 43 million transistors. Fig. 8 shows the cumulative distribution function of macro power for 10 macros with a total of 159,003 MOSFETs. Two extreme cases where the switching activities of different macros are either completely independent or perfectly correlated, and one nominal case where the switching activities of different macros are 60% correlated to a common workload, are used to illustrate how the correlation of switching activities affects the overall power distribution.

Statistical power analysis can be further combined with statistical timing analysis to make better yield predictions. Fig. 9 shows the joint probability density function of both power and delay for a benchmark macro. By integrating the statistical power distribution with the statistical timing distribution, this three-dimensional yield versus power and performance plot provides a more comprehensive means for designers to define the corners, improve the yield, determine bin splits, and optimize other design variables.

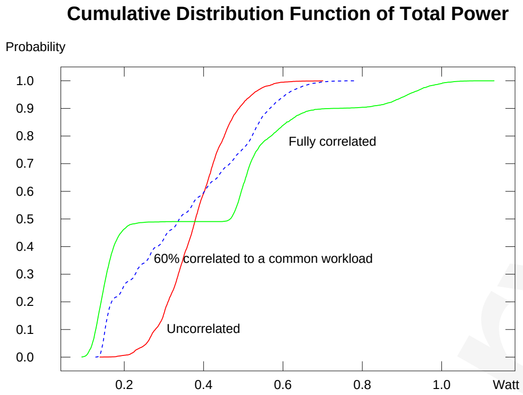

Fig. 8. Cumulative distribution of total power for 10 macros.

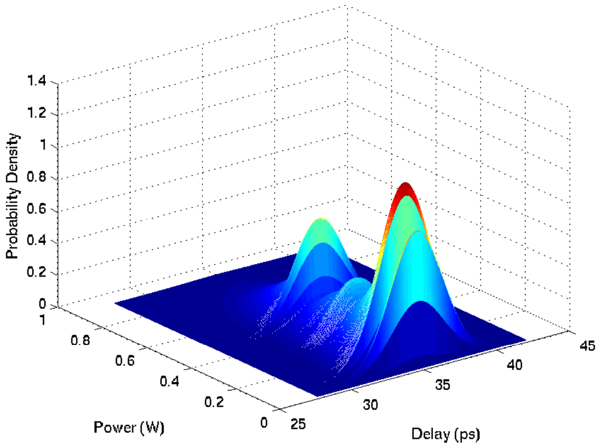

Fig. 9. Joint probability density function of power and delay.

5 Conclusions

Although the statistical distribution of leakage power due to process variation has been extensively studied in the literature, the statistical analysis of switching power due to workload variation remains a difficult challenge. This paper presents a first study on the combined analysis of leakage power and switching power that takes both global correlation and local randomness into account. Leakage power due to process variables such as oxide thickness and channel length is modeled and correlated at the transistor or gate level, while switching power due to workload-related activities is modeled and correlated at the macro or block level. In order to provide a general framework to handle non-Gaussian and multiple-peak distributions, a CDF-based Monte-Carlo simulation is performed to analyze the statistical distribution of macro and chip power. Based on these benchmark results, we not only demonstrate the feasibility of a general statistical analysis for both leakage and switching power, but also develop a design methodology where the statistical power distribution of each macro is characterized by its PDF and CDF functions in the circuit library.

References

1. Visweswariah, C., Ravindran, K., Kalafala K., Walker, S., Narayan S.: First-Order Incremental Block-Based Statistical Timing Analysis. In: 41st Design Automation Conference, pp. 331–336. ACM, New York (2004)
2. Chang, H., Zolotov, V., Narayan S., Visweswariah, C.: Parameterized Block-Based Statistical Timing Analysis with Non-Gaussian Parameters, Nonlinear Delay Equations. In: 42nd Design Automation Conference, pp. 71–76. ACM, New York (2005)
3. Rao, R., Srivastava, A., Blaauw, D., Sylvester, D.: Statistical Estimation of Leakage Current Considering Inter- and Intra-die Process Variation. In: International Symposium on Low Power Electronics and Design, pp. 84–89. ACM, New York (2003)
4. Chang, H., Sapatnekar S.: Full-Chip Analysis of Leakage Power under Process Variations, Including Spatial Correlations. In: 42nd Design Automation Conference, pp. 523–528. ACM, New York (2005)
5. U.C. Berkeley Device Group, `http://www-device.eecs.berkeley.edu/~bsim3`
6. Rao, R., Agarwal, K., Devgan, A., Nowka, K., Sylvester, D., Brown, R.: Parametric Yield Analysis and Constraint-Based Supply Voltage Optimization. In: 6th International Symposium on Quality of Electronic Design, pp. 284–290. IEEE Computer Society, Los Alamitos (2005)
7. Srivastava, A., Kaviraj, C., Shah, S., Sylvester, D., Blaauw, D.: A Novel Approach to Perform Gate-Level Yield Analysis and Optimization Considering Correlated Variations in Power and Performance. IEEE Trans. Computer-Aided Design, vol. 27, no. 2, pp. 272–285 (Feb 2008)
8. Devgan, A., Rohrer, R.: Adaptively Controlled Explicit Simulation. IEEE Trans. Computer-Aided Design, vol. 13, no. 6, pp. 746–762 (Jun 1994)