Statistical Analysis of Full-Chip Leakage Power Considering Junction Tunneling Leakage

Tao Li
Institute of Microelectronics
Tsinghua University
Beijing 100084, China
litaofrank99@mails.tsinghua.edu.cn

Zhiping Yu
Institute of Microelectronics
Tsinghua University
Beijing 100084, China
yuzhip@tsinghua.edu.cn

ABSTRACT

In this paper we address the the growing issue of junction tunneling leakage ($I_{\rm junc}$) at the circuit level. Specifically, we develop a fast approach to analyze the state-dependent total leakage power of a large circuit block, considering $I_{\rm junc}$, subthreshold leakage ($I_{\rm sub}$), and gate oxide leakage ($I_{\rm gate}$). We then propose our algorithm to estimate the full-chip leakage power with consideration of both Gaussian and non-Gaussian parameter distributions, capturing spatial correlations using a grid-based model. Experiments on ISCAS85 benchmarks demonstrate that the estimated results are very accurate and efficient. For a circuit with G gates, the complexity of our approach is $O(G)$.

Categories and Subject Descriptors

B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

General Terms

Algorithm, Design, Performance, Reliability

Keywords

Statistical analysis, junction tunneling leakage, Gaussian and non-Gaussian parameter distributions

1. INTRODUCTION

Aggressive scaling of CMOS devices in each technology generation has resulted in higher integration and performance, however, the off-state leakage has increased significantly with technology scaling [1]. There has been extensive studies on the analysis of $I_{\rm sub}$ and $I_{\rm gate}$ as they poses a fundamental scaling limit to traditional CMOS design [2]. However, scaled devices require the use of the higher substrate doping density and the application of the "halo" profiles to reduce the depletion region width of the drain/source-substrate junctions, which can cause significantly large tunneling current [1]. As an added complication for full-chip leakage analysis, process variability has grown in recent technologies due to random effects in small devices, the patterning of features smaller than the wavelength of the optical lithography system and related trends.

Many works have been developed to do a full-chip leakage analysis considering the process variations [3, 4, 5, 6], they may be based on a first order model or a quadratic model, may incorporate spatial correlation effects or not, may consider two parameter variations (e.g., $L_{\rm eff}$ and $T_{\rm ox}$) or more, etc. However, none of these have included junction tunneling leakage, and a statistical full-chip leakage analysis method, which can handel the case where the underlying process parameters are correlated non-Gaussian distributions, is still in need.

In this paper, we make two primary contributions. First is the development of a fast approach for total leakage current analysis that considers $I_{\rm junc}$, $I_{\rm sub}$ and $I_{\rm gate}$. The second contribution is that we propose a novel algorithm to do the full-chip leakage analysis that can handle the case where the underlying process parameters may be spatially correlated non-Gaussian as well as Gaussian distributions. As a preprocessing step, we employ independent component analysis (ICA) to transform the set of correlated non-Gaussian parameters to a basis set of parameters that are statistically independent, and principal component analysis (PCA) to orthogonalize the Gaussian parameters. Together, we refer to this set of independent variables as the *basis set*. Next, we use some mathematical manipulations to represent the full-chip leakage as a linear canonical function of the basis set. From the moments of the basis set, we compute the moments of the full-chip leakage variables and translate them into a probability distribution function (PDF).

2. LEAKAGE ANALYSIS METHOD

The states dependence of $I_{\rm sub}$ and $I_{\rm gate}$ have been effectively studied in [2]. While in this paper, we focus on the junction leakage on circuit behavior. To examine the state dependence of $I_{\rm junc}$, we first consider a simple inverter shown in Fig. 1 (a). For a low input state, the NMOS $I_{\rm junc}$ combines with the NMOS $I_{\rm sub}$ and each can be computed independently and then added to obtain the total leakage current. For a high input state, the total leakage can be modeled as a sum of PMOS $I_{\rm sub}$, PMOS $I_{\rm junc}$, and NMOS $I_{\rm gate}$, these three components can also be generated independently. We next consider a multi-input gate with an NMOS transistor stack. If all inputs have a high state, the analysis is again similar to that of the inverter. For input states where at least one input is low and the gate output is

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
DAC 2007, June 4–8, 2007, San Diego, California, USA.
Copyright 2007 ACM 978-1-59593-627-1/07/0006 ...$5.00.

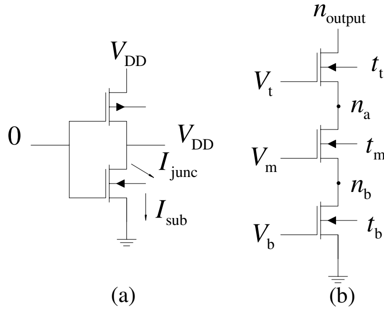

Figure 1: Circuits with junction tunneling leakage

high, $I_{\rm sub}$ through turned-off transistors and $I_{\rm gate}$ through turned-on transistors may combine with the junction tunneling leakage at internal stack nodes. These three leakage components are therefore interdependent in these cases, and must be analyzed simultaneously. Based on this observation, We now consider the junction tunneling leakage in six distinct scenarios for a 3-transistor stack (Fig. 1 (b)), while the complementary PMOS transistors are omitted for clarity. Our analysis method can be easily extended to include PMOS-based $I_{\rm junc}$. We now discuss each scenario in more detail:

1. $V_{\rm t} = 0$, $V_{\rm m} = V_{\rm b} = V_{\rm DD}$. In this case, the internal nodes $n_{\rm a}$ and $n_{\rm b}$ have a conducting path to the ground node and are at nominal 0V. The $I_{\rm junc}$ of output node and $I_{\rm gate}$ of transistor $t_{\rm m}$ and $t_{\rm b}$ are added to the $I_{\rm sub}$ of the stack to obtain the total leakage.

2. $V_{\rm t} = V_{\rm m} = 0$, $V_{\rm b} = V_{\rm DD}$. In this case, node $n_{\rm b}$ has a conducting path to ground and is at 0V. Based on SPICE simulation, node $n_{\rm a}$ has a voltage in a range of 100-200mV, and $I_{{\rm junc},n_{\rm a}}$ under such a low bias is over one order of magnitude smaller than scenario 1 and can be safely ignored, leaving the $I_{\rm sub}$ relatively unchanged. Therefore, $I_{\rm sub}$, $I_{\rm junc,output}$ and $I_{{\rm gate},n_{\rm b}}$ can be computed independently and be added up to obtain the total leakage.

3. $V_{\rm t} = V_{\rm b} = 0$, $V_{\rm m} = V_{\rm DD}$. In this case, based on SPICE simulation, internal nodes $n_{\rm a}$ and $n_{\rm b}$ have a voltage in the range of 100-200mV. Based on the discussion in scenario 2, we can safely ignore $I_{{\rm junc},n_{\rm a}}$ and $I_{{\rm junc},n_{\rm b}}$ for their small magnitudes. From the analysis in [2], we find the sum of $I_{\rm sub}$ and $I_{\rm gate}$ by computing each of the two components separately and set the total current to their maximum. This result is then added to the $I_{\rm junc,output}$ of the stack to obtain the total leakage.

4. $V_{\rm t} = V_{\rm DD}$, $V_{\rm m} = V_{\rm b} = 0$. In this case, the internal node $n_{\rm a}$ has a conducting path to the output node and is held at $(V_{\rm DD} - V_{\rm th})$ (with body effect), while $n_{\rm b}$ has a voltage in a range of 100-200mV. Based on the discussion in scenario 2, only $I_{\rm junction,output}$ and $I_{{\rm junc},n_{\rm a}}$ need to be considered for the total leakage.

5. $V_{\rm t} = V_{\rm m} = V_{\rm DD}$, $V_{\rm b} = 0$. In this case, the total leakage can be computed with the method described in scenario 4.

6. $V_{\rm m} = 0$, $V_{\rm t} = V_{\rm b} = V_{\rm DD}$. For the internal node $n_{\rm a}$ has a conducting path to the output node and is held at $(V_{\rm DD} - V_{\rm th})$, the internal node $n_{\rm b}$ is held at 0V. For the computation of junction tunneling leakage, only $I_{\rm junction,output}$ and $I_{{\rm junc},n_{\rm a}}$ need be considered.

Based on the six scenarios, we find that the junction leakage for a transistor stack can be computed independently with the computation of $I_{\rm sub}$ and $I_{\rm gate}$. Junction leakage current has a state dependency and a simple look-up table can be used to include this effect. By keeping only dominant states for $I_{\rm junc}$, i.e., the number of "on" transistors connected to the output node in a series transistor stack, the size of the table for a k-input can be greatly reduced from 2^k to k while maintaining a reasonable accuracy. For the fast approach to compute $I_{\rm sub}$ and $I_{\rm gate}$, the reader is referred to [2] for details.

To demonstrate the accuracy of the proposed leakage estimation method, we show the analysis results for a 4-input NAND gate under all possible input states in Fig. 2. The analytical model for junction leakage is given in [1], and it is cooperated with a 65nm technology file to study the impact of $I_{\rm junc}$ on circuit behavior. It has a $T_{\rm ox}$ of 17Å, $L_{\rm eff}$ of 50nm. $V_{\rm th}$ is approximately 400mV, $V_{\rm DD}$ is equal to 1.2V, and all results are for the room temperature. Compared with the leakage current results obtained from SPICE simulation, our scheme exhibits an average absolute error of 1.5% over all input states, while the maximum error occurs for state 1110 and is 4.5%.

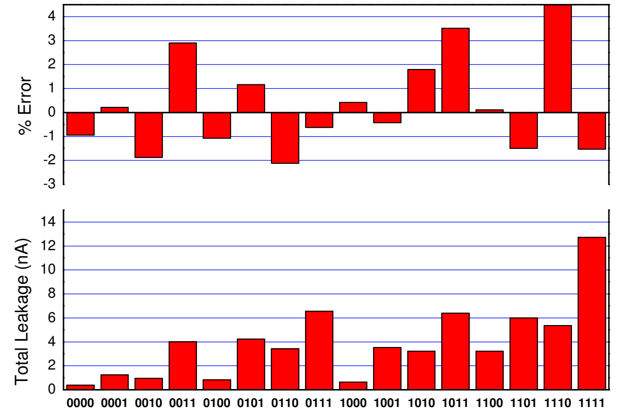

Figure 2: Leakage estimation for 4-input NAND gate

3. MODELING PROCESS VARIATIONS

To incorporate the effects of both Gaussian and non-Gaussian parameters of distribution in our leakage analysis framework, the overall chip area is divided into grids as in [3], and the process variations, (e.g., $\Delta V_{\rm th}$, $\Delta T_{\rm ox}$, and ΔL) of each logic grid, are pretreated to have zero mean and unit variance. If the components of random vector X were correlated Gaussian random variables with a covariance matrix $\sum$, PCA can be applied to decompose correlated Normal distributions into independent ones [7]. After PCA, the prescaled process variations can be modeled as:

$$X_{{\rm grid}i} = A_{{\rm grid}i} S \qquad (1)$$

$X_{{\rm grid}i} = [X_{{\rm grid}i1} \quad X_{{\rm grid}i2} \cdots]^T$ denotes the Gaussian random parameter variables of i-th logic grid, $S = [s_1 s_2 \cdots s_N]^T$

is extracted by PCA, its components are all independent and satisfy the standard Normal distribution. N is the total number of Gaussian random variables of the entire die, and it is typically large (e.g. $10^3 \sim 10^6$) for practical industry designs. $A_{\text{grid}i}$ captures the correlations among the random variables.

For the correlated non-Gaussian variables of i-th grid, PCA transformation would not guarantee statistical independence for the correlated non-Gaussian variables. However, ICA is a mathematical technique that precisely accomplishes the desired goal [7]. The procedure can be written as:

$$Y_{\text{grid}i} = B_{\text{grid}i} R \tag{2}$$

where $Y_{\text{grid}i} = [Y_{\text{grid}i1} \quad Y_{\text{grid}i2} \cdots]^T$ denotes the non-Gaussian random parameter variables of i-th logic grid. The dimension of R is M, and M is the total number of non-Gaussian random variables of the entire die, which is also very large.

We then substitute Eq. 1 and Eq. 2 to arrive at the following parameter model:

$$Z_{\text{grid}i} = C_{\text{grid}i} \cdot T \tag{3}$$

where $Z_{\text{grid}i} = [X^T_{\text{grid}i} \quad Y^T_{\text{grid}i}]^T$, $T = [S^T \quad R^T]^T$, $C_{\text{grid}i} = \begin{bmatrix} A_{\text{grid}i} & 0 \\ 0 & B_{\text{grid}i} \end{bmatrix}$.

4. FULL-CHIP LEAKAGE ANALYSIS

4.1 Full-chip Leakage Modeling

Statistical leakage analysis typically starts from the leakage modeling for logic grids. Based on the fast approach proposed in Section 2, we approximate the logarithm of the grid leakage current by a linear model (More details for this procedure are not included in this paper due to the limited number of available pages, however, the reader is referred to [3] for details):

$$\log(I_{\text{grid}i}) = \widetilde{V}^T_{\text{grid}i} \cdot Z_{\text{grid}i} + W_{\text{grid}i} \tag{4}$$

where $I_{\text{grid}i}$ denotes the total leakage current (including junction tunneling leakage, subthreshold leakage and gate tunneling leakage) of the i-th grid, $\widetilde{V}^T_{\text{grid}i}$ is the sensitivity vector of the leakage with respect to the zero-mean randomly varying parameters $Z_{\text{grid}i}$, and $e^{(W_{\text{grid}i})}$ is the nominal leakage of the i-th grid. The grid leakage in Eq. 4 can be either the leakage current for a fixed input state or the average leakage current over all input states. Substituting Eq. 3 into Eq. 4 yields:

$$\log(I_{\text{grid}i}) = V^T_{\text{grid}i} \cdot T + W_{\text{grid}i} \tag{5}$$

where $V_{\text{grid}i} = C^T_{\text{gridi}} \cdot \widetilde{V}_{\text{grid}i}$, $V_{\text{grid}i} \in R^{M+N}$, and $M+N$ denotes the total number of random variables for the entire die.

For simplifying the notation, we define the following symbols to represent all grid leakage models in a matrix form:

$$\begin{aligned} \log(I_{\text{grid}}) &= [\log(I_{\text{grid}1}) \quad \log(I_{\text{grid}2}) \quad \cdots \quad \log(I_{\text{grid}P})]^T \\ V_{\text{grid}} &= [V_{\text{grid}1} \quad V_{\text{grid}2} \quad \cdots \quad V_{\text{grid}P}] \\ W_{\text{grid}} &= [W_{\text{grid}1} \quad W_{\text{grid}2} \quad \cdots \quad W_{\text{grid}P}]^T \end{aligned} \tag{6}$$

where P is the total number of logic grids in a chip. Comparing Eq. 6 with Eq. 5, it is easy to verify that:

$$\log(I_{\text{grid}}) = V^T_{\text{grid}} \cdot T + W_{\text{grid}} \tag{7}$$

We next develop the algorithm to efficiently extract the model of the full-chip leakage current. As the equation shown below:

$$I_{\text{chip}} = I_{\text{grid}1} + I_{\text{grid}2} + \cdots + I_{\text{grid}P} \tag{8}$$

the full-chip leakage current is the sum of all grid leakage currents. Applying the log transform to both sides of Eq. 8 yields:

$$\log(I_{\text{chip}}) = \log(e^{\log(I_{\text{grid}1})} + e^{\log(I_{\text{grid}2})} + \cdots + e^{\log(I_{\text{grid}P})}) \tag{9}$$

Substitute Eq. 7 into Eq. 9,we have:

$$\log(I_{\text{chip}}) = \log(\sum_{i=1}^{P} e^{(V^T_{\text{grid}i} \cdot T + W_{\text{grid}i})}) \tag{10}$$

Since the parameter variations are in general around 10-20% [3], we employ a second-order Taylor expansion at the nominal values of $e^{(W_{\text{grid}i})}$, after some mathematical manipulations we obtain the full-chip leakage:

$$\log(I_{\text{chip}}) = V^T_{\text{chip}} \cdot T + W_{\text{chip}} \tag{11}$$

where the model coefficients are given by:

$$W_{\text{chip}} = \log(\frac{1}{\alpha}) \tag{12}$$

$$V_{\text{chip}} = \alpha \cdot V_{\text{grid}} \cdot \beta \tag{13}$$

$$\alpha = \frac{1}{e^{W_{\text{grid}1}} + e^{W_{\text{grid}2}} + \cdots + e^{W_{\text{grid}P}}} \tag{14}$$

$$\beta = [e^{W_{\text{grid}1}} \quad e^{W_{\text{grid}2}} \quad \cdots \quad e^{W_{\text{grid}P}}] \tag{15}$$

The values of α and β in Eq. 14 and Eq. 15 can be computed with linear computational complexity.

4.2 Full-chip Leakage PDF Evaluation

The inputs required for our full-chip leakage analysis correspond to the moments of parameters of variation. Given the moments of the independent components, $t_1, t_2, \cdots, t_{M+N}$, which can be generated by the binomial moment evaluation scheme from the moments of $Z_{\text{gridi}}, i = 1, 2, \cdots, P$ [7], as inputs to the APEX algorithm [8]. The PDF/CDF of $\log(I_{\text{chip}})$ can be extracted from Eq. 11. After that, the PDF/CDF of I_{chip} can be easily computed by a simple nonlinear transform [9].

4.3 Computational Complexity

Considering the preprocessing steps including the ICA and PCA transforms, and the moments generation of the independent components $t_1, \cdots, t_{M+N}$ as a one time precharacterization cost, the full-chip leakage analysis procedure consists of the following main steps: generation of the linear model expressed in Eq. 4 for all grid leakage currents and the computation of full chip leakage using Eq. 13. Based on the analysis in [3], for a circuit with G gates, the computational complexity for the generation of all the grid leakage currents is $O(G)$. Based on the previous discussion, the dimension of matrix V_{grid} is $(M+N) \times P$, where $M+N$ is the total number of the random variables, and P is the number grids. Since M and N are both $O(G)$, the computation

Table 1: Comparison of the proposed method results with Monte Carlo simulation

Circuit Name	Gate Number	Grid Number	Our Method $\mu(\mu A)$	Our Method $\sigma(\mu A)$	Error($\frac{Our-MC}{MC}$)% μ	σ	95%Pt	5%Pt	Error($\frac{MCG-MC}{MC}$)% μ	σ	95%Pt	5%Pt
C7552	5235	64	28.53	10.74	-1.63	3.02	3.84	3.91	6.32	23.44	24.66	4.56
C5315	3768	64	21.44	8.12	-1.07	-2.82	-4.09	-3.68	5.69	17.56	20.31	4.89
C6288	2552	16	17.05	6.70	-1.15	-2.14	3.52	3.61	5.98	14.63	14.89	3.11
C3540	2491	16	13.71	4.58	0.71	1.56	2.97	2.88	4.96	10.23	15.34	-3.16
C2670	1854	16	9.22	3.87	-0.81	1.34	2.90	2.77	4.78	8.84	11.13	2.34
C1908	1197	16	5.14	2.01	-0.64	-0.98	-2.45	2.12	3.45	8.02	8.98	4.34
C880	556	4	2.56	0.89	-0.23	-0.59	-1.26	-1.32	2.12	6.14	9.32	1.23
C432	273	4	1.15	0.34	-0.07	-0.23	-0.98	-0.84	1.29	5.99	4.14	-2.01

of the matrix-vector product $V_{\text{grid}} \cdot \beta$ has a complexity of $O(G \cdot P)$. In general, the number of grids, is substantially smaller than the number of gates in the circuit and can be regarded as a constant number. Therefore, the time complexity for our methodology is $O(G)$.

5. RESULTS

Our methodology for statistical modeling of full-chip leakage dissipation was implemented and tested with 8 ISCAS85 circuits. The technology parameters that were used correspond to a 65nm commercial technology model, and the 3σ value of parameter variations for L, T_{ox} and N_{d} were set to 20% of the nominal parameter values, of which inter-die variations constitute 50% and intra-die variations, 50%. The parameters L and T_{ox} are modeled as correlated sources of variations, and the dopant concentration, N_{d} is modeled as an independent source of variation. The same framework can be easily extended to include other parameters of variations. We model the gate length L of gates in each grid as non-Gaussian parameters, and T_{ox} of gates in each grid as Gaussian parameters. For the correlated non-Gaussian L parameters, we randomly assign to L in each grid a uniform distribution. The independent parameter N_{d} is assumed to follow a Poisson distribution.

For comparison purposes, we performed Monte Carlo (MC) simulations with 10,000 runs on the benchmarks, and the results of the comparison are shown in Table 1. We compare the mean (μ), the standard deviation (σ), the 95% and the 5% quantile points of the full-chip leakage current distribution obtained from our scheme with those generated from the MC simulations as the metrics of accuracy. As seen in Table 1, the results of our scheme are quite close to that of MC simulations. These errors are reasonably small as compared to the accuracy penalty paid by assuming the incorrect normal distribution modeling of parameters. Columns ten to thirteen of Table 1 show the error incurred when modeling the non-Gaussian L parameters as normally distributed random variables and performing MC simulations, termed as MCG, for each benchmark circuit. For instance, for the largest benchmark circuit C7552, when assuming that the non-Gaussian L parameters follow Gaussian distributions, the error observed is 6.32% for μ, 23.44% for σ, 24.66 % for the 95% point and 4.56% for the 5% point. Thus, modeling the non-Gaussian parameters as normally distributed ones leads to significant inaccuracy.

6. CONCLUSIONS

We developed a fast approach to compute the total leakage current in circuit blocks considering three leakage components: I_{sub}, I_{gate}, and I_{junc}. The proposed approach accurately accounts for the complex interaction of these leakage components in stacked MOS configurations and is based on pre-characterized tables for only dominant input states. Based on the proposed analysis method, we propose an efficient statistical full-chip leakage analysis algorithm incorporating both Gaussian and non-Gaussian parameter distributions, capturing spatial correlations using a grid-based model. We have also shown that the correlated non-Gaussian parameters must be considered appropriately in order to predict the full-chip leakage distribution correctly.

7. ACKNOWLEDGMENTS

This work is supported by a grant from China's Ministry of Science and Technology (973 project #2006CB302700). The collaboration with SMIC (Dr. Hanming Wu) and Cadence Berkeley Labs (Drs. Andreas Keuhlmann and Zhenhai Zhu) are greatly appreciated.

8. REFERENCES

[1] S. Mukhopadhyay, S. Member, A. Raychowdhury, S. Member, and K. Roy. Accurate estimation of total leakage in nanometer-scale bulk cmos circuits based on device geometry and doping profile. *IEEE Trans. CAD*, 24(3):363–381, March 2005.

[2] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester. Analysis and minimization techniques for total leakage considering gate oxide leakage. In *Proc. DAC*, pages 175–180, 2003.

[3] H. Chang and S. Sapatnekar. Full-chip analysis of leakage power under process variations, including spatial correlations. In *Proc. DAC*, pages 523–530, 2005.

[4] X. Li, J. Le, and L. T. Pleggi. Projection-based statistical analysis of full-chip leakage power with non-log-normal distributions. In *Proc. DAC*, pages 103–108, 2006.

[5] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester. Statistical analysis of subthreshold leakage current for VLSI circuits. *IEEE Trans. VLSI SYST.*, 12(2):131–139, February 2004.

[6] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, and S. Director. Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance. In *Proc. DAC*, 2005.

[7] J. Singh and S. Sapatnekar. Statistical timing analysis with correlated non-Gaussian parameters using independent component analysis. In *Proc. DAC*, pages 155–162, 2006.

[8] X. Li, J. Le, P. Gopalakrishnan, and L. T. Pleggi. Asymptotic probability extraction for non-normal distributions of circuit performance. In *Proc. ICCAD*, pages 2–9, 2004.

[9] A. Papoulis and S. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2001.