Interdependency Study of Process and Design Parameter Scaling for Power Optimization of Nano-CMOS circuits under Process Variation

Saraju P. Mohanty
smohanty@unt.edu

E. Kougianos
eliask@unt.edu

Dhruva Ghai
dvg0010@unt.edu

VLSI Design and CAD Laboratory
Dept. of Computer Science and Engineering
University of North Texas, Denton, TX 76203.

Priyadarsan Patra
Priyadarsan.patra@intel.com

Microprocessor Technology Labs
Intel Corporation
Hillsboro, OR 97124.

Abstract

*In sub-*65nm *CMOS technology, switching power and gate as well as subthreshold leakage power are the major components of total power dissipation. To achieve power-performance trade-offs one varies different process parameters* (T_{ox}, K, $V_{th,}$) *and design parameters* (V_{DD}, W). *Techniques for (i) dual-*K *and dual-*T_{ox} *have been proposed to reduce gate leakage, (ii) dual (multiple)-*V_{th} *has been introduced to minimize subthreshold leakage, and (iii) dual (multiple)-*V_{DD} *has been used to optimize dynamic power. In view of the optimization regime above, the following question arises: I*f T_{ox}, L, V_{th}, and V_{DD}, etc. are scaled simultaneously, will one obtain a power-performance optimal circuit that has minimal gate leakage, minimal subthreshold leakage, and minimal dynamic power consumption? *The objective of this paper is to essentially answer this question. Assuming dual values of* T_{ox}, V_{th}, *and* V_{DD} *for a particular* K, *we estimate the values for gate leakage, subthreshold leakage, dynamic power, and performance in architectural units such as adder, multiplexor, multiplier, etc. while accounting for the process variation. Statistical variations in the parameters, each assumed to be Gaussian, are explicitly taken into account by using Monte Carlo simulations while characterizing the architectural units. The paper then analyzes the proportion of values of gate and subthreshold leakage and dynamic power in the total power consumption of these units. This in essence gives a relative and integrated perspective of various power-performance tradeoffs against the nominal case, thus serving as a guideline to help designers take appropriate decisions.*

1. Introduction

Power dissipation is an important design constraint in high performance processor systems and system-on-chip (SOC) design, in addition to the case of traditional low power targeted applications, such as portable systems. To meet the increasing demand of low power VLSI circuits with high performance, VLSI design engineers are implementing aggressive scaling in process as well as design parameters of CMOS transistors. Both dynamic and static power are significant fractions of total power dissipation in a nanoscale CMOS circuit. Each one of them has several forms and origins; they flow between different terminals and in different operating conditions of a transistor as shown in Fig. 1.

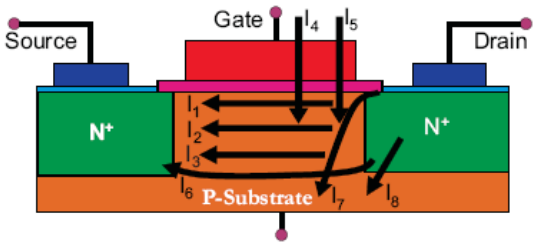


Figure 1. Current flow paths in a nanoscale CMOS transistor during power dissipation in different states of its operation [24, 16, 4]: I_1 - drain to source active current (ON state), I_2 - drain to source short circuit current (ON state), I_3 - subthreshold leakage (OFF state), I_4 - gate leakage (both ON and OFF states), I_5 - gate current due to hot carrier injection (both ON and OFF states), I_6 - channel punch through current (OFF state), I_7 - gate induced drain leakage (OFF state), I_8 - reverse bias PN junction leakage (both ON and OFF states).

In short channel nano-CMOS transistors, several short channel effects (SCE) become significant, such as drain induced barrier lowering (DIBL), large V_{th} roll-off, diminishing on-to-off current ratio and band-to-band tunneling (BTBT.) As a result, there has been a drastic change in the leakage components of the device both in the inactive as well as active mode of operation. The

leakage current in short channel nanometer transistors has diverse forms, such as reverse biased diode leakage, subthreshold leakage, SiO_2 tunneling current (leading to gate leakage), hot carrier gate current, gate induced drain leakage, and channel punch through current [24, 16, 4]. While biased diode leakage and SiO_2 tunnel currents flow during both active and sleep mode of the circuit, the other currents flow during the sleep mode only. The ITRS prediction of major sources of power dissipation, such as dynamic current, subthreshold leakage and gate leakage is presented in Fig. 2 [8, 1].

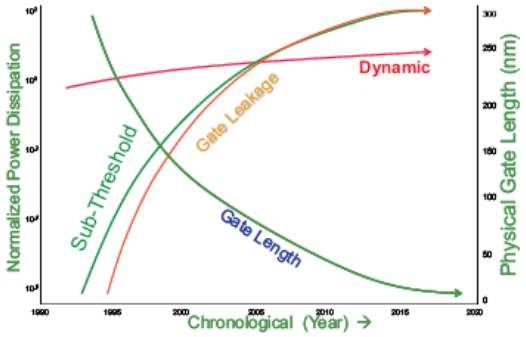


Figure 2. Prediction of trends of major sources of power dissipation in nanoscale CMOS transistors and circuits [8, 1].

It has been observed that the principal power components (dynamic current, I_{dyn}) and leakage components (gate leakage, I_{gate}, and subthreshold leakage I_{sub}) predominantly depend on the gate oxide thickness (T_{ox}), threshold voltage (V_{th}) and supply voltage (V_{DD}). Hence any methodology for power optimization must focus on the variation of these process and design parameters.

The rest of the paper is organized as follows: in Section 2 we outline the novelty and contributions of this paper. In Section 3, we summarize relevant research considering power optimization in nano-CMOS technologies. In Section 4, a hierarchical methodology to characterize architectural level units for gate leakage, subthreshold leakage, and dynamic power, as well as their propagation delay is presented. In Section 5, we investigate the effects of process variation on scaling. In Section 6, we examine the effects of scaling on individual power components. The paper concludes in Section 7.

2. Contributions of this Paper

The contributions of this paper are in many forms. First, we develop a methodology to characterize nano-CMOS based architectural components for gate leakage, subthreshold leakage, and dynamic power while simultaneously accounting for process variation. We consider several input design corners and translate them to output power performance corners, which can be used by design engineers for design space exploration. We provide statistically characterized models, which were simulated at transistor level for obtaining the mean (μ) and standard deviation (σ) of gate leakage, subthreshold leakage, dynamic current and propagation delay with simultaneous variation of all process and design parameters. These characterization data can be eventually used in a power optimization framework. We analyze the interdependency of T_{ox}, V_{th}, and V_{DD} scaling on various power (current) components with and without process variation. In particular, seven different cases, such as (1) only T_{ox} scaling, (2) only V_{th} scaling, (3) only V_{DD} scaling, (4) simultaneous T_{ox} and V_{th} scaling, (5) simultaneous T_{ox} and V_{DD} scaling, (6) simultaneous V_{th} and V_{DD}, and (7) simultaneous T_{ox} and V_{th} and V_{DD} scaling, are analyzed for optimization of various forms of power; at the same time, the performance penalty has been tabulated.

3. Related Research

The current literature contains many research contributions considering various form of scaling for power optimization. These include mainly dual-V_{th}, dual-T_{ox}, and dual-V_{DD} designs. However, these scaling methods have been applied relatively independently. For example, researchers apply dual-V_{th} to reduce subthreshold leakage without studying its impact on gate leakage and dynamic power. On the other hand, the goal of this research is to study the interdependency of these scaling methods from power (current) and performance (delay) point of view.

3.1. Dual V_{th} Works

Multiple threshold CMOS have been used by Pant *et al.*. [21] as well as Rao *et al.* [23] for subthreshold current reduction. Khouri and Jha [10] proposed a dual-V_{th} technique for subthreshold leakage analysis and reduction during behavioral synthesis, targeting the least used modules as the candidates for leakage optimization. Gopalakrishnan and Katkoori in [6, 7] also use the multi-threshold CMOS approach

for reduction of subthreshold current during high level synthesis and propose binding algorithms for power, delay, and area trade-off. They used a clique partitioning approach in [7] and a knapsack based binding algorithm in [6]. In [14], Liu *et al.* have applied probabilistic analysis to V_{th} variation. The analysis of dual V_{th} design methodology is done in the presence of large variations in threshold voltage. In [2], a dual V_{th} and dual T_{ox} technique is applied to SRAMs in order to reduce leakage. In [12], a dual V_{th} FPGA architecture has been proposed in which the logic elements are used for dual V_{th} assignment. In [28], Wei *et al.* have tried to reduce the leakage power by using high V_{th} transistors in the non-critical paths, and low V_{th} transistors in the critical paths.

3.2. Dual T_{ox} Related Works

In [19] Mukherjee *et al.* have proposed a gate oxide leakage minimization approach using dual-Tox and dual-K. Mohanty *et al.* in [17] have presented analytical models and a datapath scheduling algorithm for reduction of tunneling current. The heuristic assigns higher thickness resources to more leaky nodes (multipliers), but does not address the area overhead. In [13], Lee *et al.* developed a method for analyzing gate oxide leakage current in logic gates and suggested utilizing pin reordering to reduce gate leakage. Sultania *et al.* in [27], developed an algorithm to optimize the total leakage power by assigning dual T_{ox} values to transistors in a given circuit. In [26] Sirisantana and Roy use multiple channel lengths and multiple gate oxide thickness for reduction of leakage. In [20] Mukhopadhyay *et al.* have carried out extensive modeling and estimation of total leakage current of CMOS devices considering the effect of parameter variation.

3.3. Dual V_{DD} Related Works

The research using this technique is quite mature and several approaches have been proposed in the literature over the last several years [18, 15, 9, 25, 5]. Certain types of circuitry called Level Converters are used for this purpose. The transistors on critical paths are operated on a higher supply voltage (VDDH), whereas transistors on the non-critical paths are operated on a lower supply voltage (VDDL) [11] and [22].

4. Process Variation Aware Power and Performance Characterization Methodology for Architectural Units

In this section we present a hierarchical methodology to characterize architectural level units for gate leakage, subthreshold leakage, and dynamic power, as well as their propagation delay as shown in Fig. 3. Initially a 2 input NAND gate was designed and layout tested for functional correctness. Via Monte Carlo simulations, we translated the process and design variations (inputs) into gate leakage, dynamic and subthreshold current and delay probability density distributions (outputs). The input process and design variations are assumed to be Gaussian in nature. This is demonstrated in Fig. 4.

While state dependent data are obtained at the logic level, at the architectural level we followed a state independent approach. This was done by using the state averaged data derived from the characterized NAND gate. In order to account for the lognormal distribution of the currents at the gate level, we used the Central Limit Theorem. Since a typical functional unit is comprised of hundreds of NAND gates, according to the theorem the leakage, dynamic and subthreshold currents for the total unit will be normally distributed even though the same currents are lognormally distributed for each individual gate.

Following the above discussion, we can model the currents and the delays of the functional units (typically synthesized into a network of 2-input NAND gates) by utilizing the characterized data for a 2-input NAND gate. Since the total current in the functional unit can be defined as the sum of currents in the individual NAND gates comprising the unit, and the distributions for each gate are statistically independent of each other, the mean and variance of the currents can be derived as:

$$\begin{aligned} \mu_{FU} &= N\mu_{NAND} \\ \sigma_{FU} &= \sqrt{N}\sigma_{NAND} \end{aligned} \tag{1}$$

where there are N NAND gates in the implementation of the functional unit. From the above equations the mean and the variance of I_{gate},

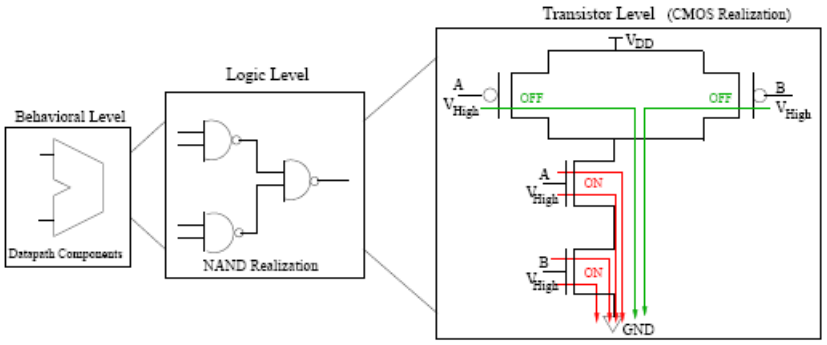


Figure 3. The Three Levels Of Hierarchy

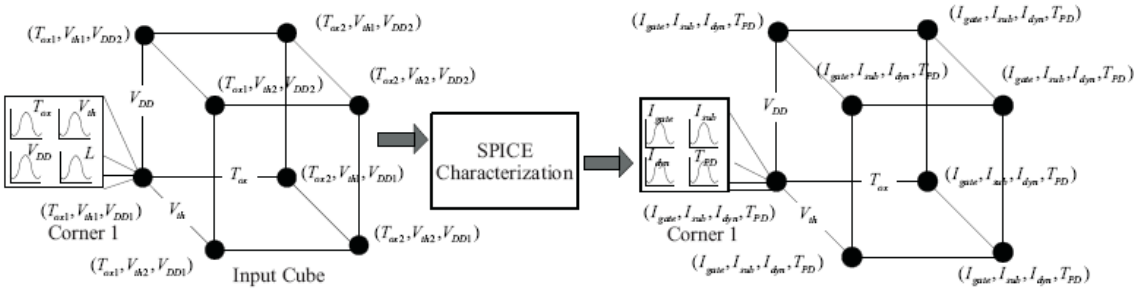


Figure 4. Simulation Methodology accounting for statistical process variation for power optimization.

I_{sub} and I_{dyn} for each of the functional units was calculated. The calculation of the mean and variance for the delay T_{PD} also was performed in a similar manner. At the end of this procedure, a complete process and design variation aware cell library was obtained and used in the subsequent optimization procedure.

5. Accounting for Process Variation with Scaling

In this section we describe the methodology by which the statistical information regarding process and design parameter variability is translated into statistical information about power consumption and performance, as shown schematically in Fig. 4.

The SPICE characterization engine considers the combination of the dual values of the three input parameters (T_{ox}, V_{th}, and V_{DD}) as eight corners of a design cube. The engine then processes each corner of the input cube and generates a corresponding output cube. The output consists of eight power and propagation delay (T_{PD}) probability density functions, each set corresponding to a particular design corner of the input cube. The provided statistical information for each input corner is used to obtain sets of current (Igate, Isub and Idyn) and propagation delay (T_{PD}) probability density functions, each set corresponding to a particular design corner of the input cube. The provided statistical information for each input corner is used to generate N = 1000 Monte Carlo runs (per corner) which provides the statistical distributions of the output parameter.

It was observed that with normally distributed input parameters, the distribution of the output currents was lognormal (as expected from their exponential dependence on the inputs) while that of the propagation delay was Gaussian. Sample distributions of the logarithms of the currents (which are normally distributed) and the delay are shown in Fig. 5.

6. Analysis of Effects of Scaling on Individual Power Components

Characterization data for various corners are shown in Fig. 6. For the analysis, we consider the 8 corners of the cube as nominal corners. Corner 1 is considered as the baseline corner having values of V_{DD}, T_{ox}, and V_{th} as the typical values for a 45nm nano-CMOS technology. The values in the other nominal corners are varied with respect to

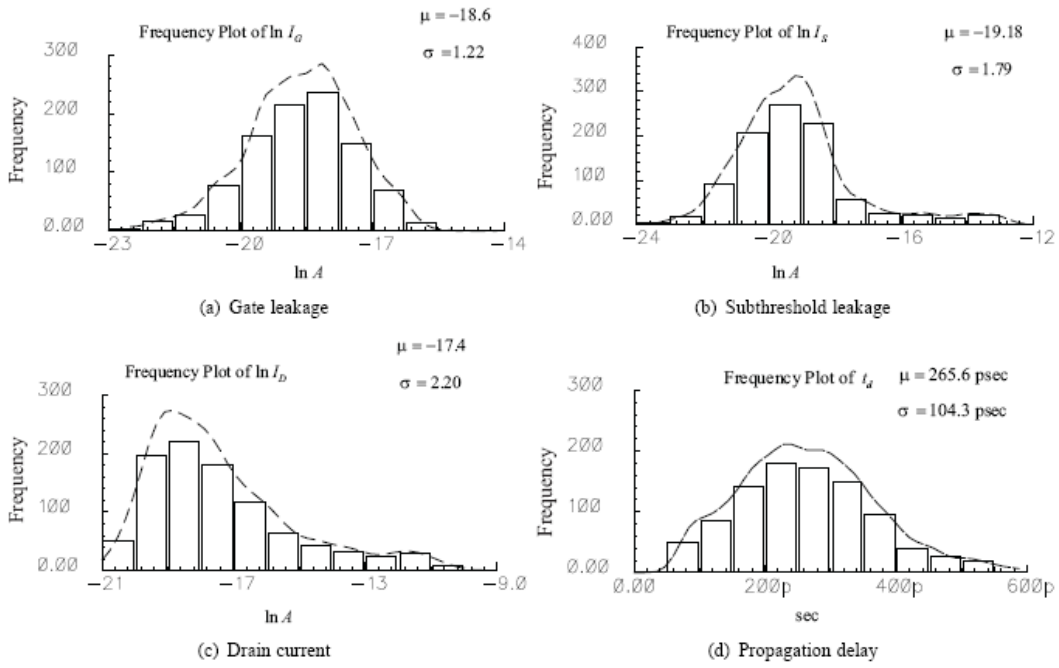


Figure 5. Effects of statistical process variation on gate leakage, subthreshold leakage, dynamic power and propagation delay in a 2-input NAND gate.

this corner. Then these nominal corners are processed through SPICE, and the outputs obtained are the values of I_{gate}, I_{sub}, I_{dyn}, and T_{pd}.

For the purposes of this analysis, we are considering the case of a divider only. But the trend is the same for other datapath components as well. In total we had 7 cases. These are nominal results without considering statistical distributions. In the next section we will consider the results with the distribution. It should also be pointed out that, in this discussion, we refer to "scaling" as the process of reduction of power. In that sense, scaling V_{DD} implies reduction in its value but scaling T_{ox} and V_{th} implies an *increase* in their values.

6.1. Only Tox scaling

This case may arise when T_{ox1} changes to T_{ox2}, e.g. (T_{ox1}, V_{th1}, V_{DD1}) versus (T_{ox2},V_{th1},V_{DD1}). In this case we observe that all power components are reduced with an overall reduction of 91.6% achieved. As expected, the increase in oxide thickness results in a 65.8% delay penalty.

6.2. Only V_{th} scaling

Scaling V_{th} only ((T_{ox1},V_{th1},V_{DD1}) versus (T_{ox1},V_{th2},V_{DD1})):

In this case we observe that total power dissipation decreases by 46.1% while the delay penalty is only 17%.

6.3. Only V_{DD} scaling

Scaling V_{DD} only ((T_{ox1},V_{th1},V_{DD1}) versus (T_{ox1},V_{th1},V_{DD2})):

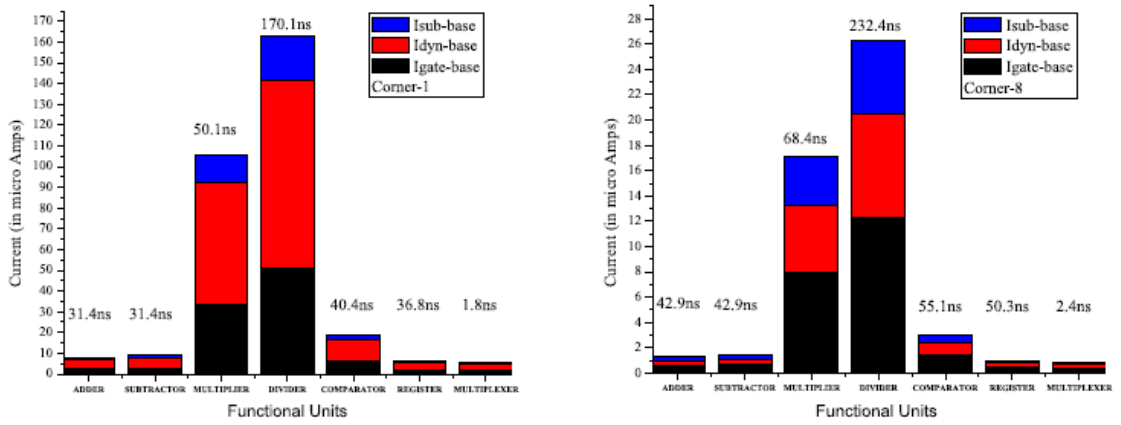
In this case we observe that total power dissipation decreases by 57.8% with a modest 21.4% delay penalty.

6.4. Simultaneous T_{ox} and V_{th} Scaling

((T_{ox1},V_{th1},V_{DD1}) versus (T_{ox2},V_{th2},V_{DD1})):

In this case the combined effect of T_{ox} and V_{th} increase results in 94.8% reduction in power but a very significant 100% delay penalty. This is due to the inverse relation of the delay to *both* T_{ox} and V_{th} [3]:

$$t_d \approx \frac{1}{T_{ox}(V_{DD} - V_{th})^2} \qquad (2)$$



(a) Design corner 1: $T_{ox} = 1.4nm$, $V_T = 0.22V$, and $V_{DD} = 0.7V$

(b) Design corner 8: $T_{ox} = 1.7nm$, $V_T = 0.25V$, and $V_{DD} = 0.9V$

Figure 6. Nominal results showing individual components of power consumption for different output corners. Corner 1 is considered as baseline while in corner 8 all parameters have been scaled. It may be noted that not only the total current values has reduced but also their proportion in the total current.

[The caption in Fig6. does not make sense! The table results need some clarification. Is it averaged over all functional units? Why is corner 8 not optimal but others are?]

Table 1. Percentage reduction in current dissipation and increase in propagation delay

	Parameter varied						
Power Component	T_{ox}	V_{th}	V_{DD}	$T_{ox}+V_{th}$	$T_{ox}+V_{DD}$	$V_{th}+V_{DD}$	$T_{ox}+V_{th}+V_{DD}$
Dynamic	88.3	52.6	70.1	94.3	93.4	73.5	96.8
Gate Leakage	96.2	11.3	69.2	97.7	98.8	72.7	99.3
Subthreshold Leakage	94.3	10.5	63.4	89.8	97.9	59.6	96.3
Total Current	91.6	46.1	57.8	94.8	96.4	71.1	97.8
Propagation Delay	65.8	17.0	21.4	100.0	101.4	42.1	142.9

6.5. Simultaneous T_{ox} and V_{DD} Scaling

((T_{ox1},V_{th1},V_{DD1}) versus (T_{ox2},V_{th1},V_{DD2})):

As anticipated from Eq. 2, the delay penalty is again significant (101.4%) with similar reduction in power as in the previous case (93.4%).

6.6. Simultaneous V_{th} and V_{DD} Scaling

((T_{ox1},V_{th1},V_{DD1}) versus (T_{ox1},V_{th2},V_{DD2})):

In this case we observe that since both V_{th} and V_{DD} have been scaled, by Eq. 2 we anticipate a more pronounced delay: a 42.1% increase. The overall power reduction is not as large as when T_{ox} is scaled (due to the exponential dependence of gate leakage on T_{ox}): 71.1%.

6.7. Simultaneous T_{ox}, V_{th} and V_{DD} Scaling

We note that the effect of scaling all the parameters cannot be easily or accurately obtained from the responses (output results) of varying a single or a few parameter(s).

((T_{ox1},V_{th1},V_{DD1}) versus (T_{ox2},V_{th2},V_{DD2})):

Finally, when all three parameters are scaled simultaneously, we obtain the power reduction of 97.8% and the worst delay penalty of 142.9% when averaged over all units. These performance results, indicated in Fig. 6, are not easily anticipated from simple analysis of the prior 6 cases (corners 2 thru 7). This is hard because of parameter interdependency and variation statistics wherein comes the usefulness of our quick statistical library models and analysis methodology..

7. Conclusion

In this work a novel process variation aware power characterization methodology was presented. An exhaustive functional unit model library was created by considering the individual

and combined variations of T_{ox}, V_{th}, and V_{DD} via transistor level Monte Carlo SPICE simulations. The statistical variation of process and device parameters (assumed known) are thus transformed into a resulting characterization consisting of the mean and S.D. of I_{gate}, I_{dyn}, I_{sub} and delay of the functional units. The effect of scaling of three parameters, T_{ox}, V_{th}, and V_{DD} on various power (current) components was studied. It was observed that simultaneous scaling of all three may not result in expected power-performance tradeoff, with the expectation based on the effect of individual parameter variations. Hence, power optimization techniques in circuit or process design, which resort to parameter selection/assignment techniques, need to do so judiciously.

References

[1] Semiconductor Industry Association, International Technology Roadmap for Semiconductors. http://public.itrs.net.

[2] B. Ameliford, F. Fallah, and M. Pedram. Reducing the Subthreshold and Gate-tunneling Leakage of SRAM Cells using Dual-Vt and Dual-Tox Assignment. In *Proceedings of Design,Automation and Test in Europe*, pages 1–6, March 2006.

[3] R. J. Baker, H. W. Li, and D. E. Boyce. *CMOS: Circuit Design, layout, and Simulation*. IEEE Press, 1998.

[4] J. A. Butts and G. S. Sohi. A Static Power Model for Architects. In *Proc. of the 33rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-33)*, pages 191–201, 2000.

[5] J. M. Chang and M. Pedram. Energy Minimization using Multiple Supply Voltages. *IEEE Trans on VLSI Systems*, 5(4):436–443, Dec 1997.

[6] C. Gopalakrishnan and S. Katkoori. Knapbind: an area efficient binding algorithm for low-leakage datapaths. In *Proc. of 21st International Conference on Computer Design*, pages 430–435, 2003.

[7] C. Gopalakrishnan and S. Katkoori. Resource allocation and binding approach for low leakage power. In *Proceedings of 16th International Conference on VLSI Design*, pages 297– 302, 2003.

[8] J. G. Hansen. Design of CMOS Cell Libraries for Minimal Leakage Currents. Master's thesis, Dept. of Informatics and Mathematical Modelling, Computer Science and Engineering Technical University of Denmark, Fall, 2004.

[9] M. Johnson and K. Roy. Datapath Scheduling with Multiple Supply Voltages and Level Converters. *ACM Transactions on Design Automation of Electronic Systems*, 2(3):227–248, July 1997.

[10] K. S. Khouri and N. K. Jha. Leakage power analysis and reduction during behavioral synthesis. In *Proceedings of International Conference on Computer Design*, pages 561–564, 2000.

[11] S. H. Kulkarni and D. Sylvester. High Performance level Conversion for Dual VDD Design. *IEEE Trans on VLSI Systems*, 12(9):926–936, Sept 2004.

[12] A. Kumar and M. Anis. Dual-Vt Design of FPGAs for Subthreshold Leakage Tolerance. In *Proceedings of International Symposium on Quality Electronic Design*, March 2006.

[13] D. Lee, D. Blaauw, and D. Sylvester. Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits. *IEEE Transactions on VLSI Systems*, 12(2):155–166, February 2004.

[14] M. Liu,W.-S.Wang, and M. Orshansky. Leakage Power Reduction by Dual-Vth Designs Under Probabilistic Analysis of Vth Variation. In *Proc. of International Symposium on Low Power Electronics and Design*, pages 2–7, Aug 2004.

[15] R. S. Martin and J. P. Knight. Using Spice and Behavioral Synthesis Tools to Optimize ASICs' Peak Power Consmpution. In *Proceedings of the 38th Midwest Symposium on Circuits and Systems*, pages 1209–1212, 1996.

[16] S. P. Mohanty and E. Kougianos. Modeling and Reduction of Gate Leakage during Behavioural Synthesis of NanoCMOS Circuits. In *Proc of International Conf on VLSI Design*, Jan 2006.

[17] S. P.Mohanty, V. Mukherjee, and R. Velagapudi. Analytical Modeling and Reduction of Direct Tunneling Current during Behavioral Synthesis of Nanometer CMOS Circuits. In *Proceedings of the 14th ACM/IEEE International Workshop on Logic and Synthesis (IWLS)*, pages 249–256, 2005.

[18] S. P.Mohanty and N. Ranganathan. Energy Efficient Datapath Scheduling using Multiple Voltages and Dynamic Clocking. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 10(2):330–353, April 2005.

[19] V. Mukherjee, S. P. Mohanty, and E. Kougianos. A Dual Dielectric Approach for Performance Aware Gate Tunneling Reduction in Combinational Circuits. In *Proceedings of the IEEE International Conference on Computer Design (ICCD)*, pages 441–443, 2005.

[20] S. Mukhopadhay and K. Roy. Modeling and estimation of total leakage current in nano-scaled CMOS devices considering the effect of parameter variation. In *Proc. of the IEEE International Symp. on Low Power Design*, pages 172–175, 2003.

[21] P. Pant, R. K. Roy, and A. Chattejee. Dual-Threshold Voltage Assignment with Transistor Sizing for Low Power CMOS Circuits. *IEEE Trans. on VLSI Systems*, 9(2):390–394, April 2001.

[22] C. Ping-Yuan and Y. Chien-Cheng. A Voltage Level Converter Circuit Design with Low Power Consumption. In *Proc. of the 6th International Conference on ASIC*, pages 358–359, Oct 2005.

[23] R.M. Rao, J. L. Burns, and R. B. Brown. Circuit Techniques for Gate and Sub-Threshold Leakage Minimization in Future CMOS Technologies. In *European Solid-State Circuits Conference*, pages 313–316, 2003.

[24] K. Roy, S. Mukhopadhyay, and H. M. Meimand. Leakage Current Mechanisms and Leakage

Reduction Techniques in Deep-Submicrometer CMOS Circuits. *Proceedings of the IEEE*, 91(2):305–327, February 2003.

[25] W. T. Shiue and C. Chakrabarti. Low-Power Scheduling with Resources Operating atMultiple Voltages. *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, 47(6):536–543, June 2000.

[26] N. Sirisantana and K. Roy. Low-power Design using Multiple Channel Lengths and Oxide Thicknesses. *IEEE Design and Test of Computers*, 21(1):56–63, Jan-Feb 2004.

[27] A. K. Sultania, D. Sylvester, and S. S. Sapatnekar. Tradeoffs Between Gate Oxide Leakage and Delay for Dual Tox Circuits. In *Proceedings of Design Automation Conference*, pages 761–766, 2004.

[28] L. Wei and et.al. Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications. *IEEE Transactions on VLSI Systems*, 7(1):16–24, March 1999.