Reducing the Sub-threshold and Gate-tunneling Leakage of SRAM Cells using Dual-V_t and Dual-T_{ox} Assignment

Behnam Amelifard
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-9481
amelifar@usc.edu

Farzan Fallah
Fujitsu Labs of America
Sunnyvale, CA
(408) 530-4544
farzan@fla.fujitsu.com

Massoud Pedram
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-4458
pedram@ceng.usc.edu

***Abstract*:** Aggressive CMOS scaling results in low threshold voltage and thin oxide thickness for transistors manufactured in very deep submicron regime. As a result, reducing the subthreshold and gate-tunneling leakage currents has become one of the most important criteria in the design of VLSI circuits. This paper presents a method based on dual-V_t and dual-T_{ox} assignment to reduce the total leakage power dissipation of SRAMs while maintaining their performance. The proposed method is based on the observation that the read and write delays of a memory cell in an SRAM block depend on the physical distance of the cell from the sense amplifier and the decoder. Thus, the idea is to deploy different types of six-transistor SRAM cells corresponding to different threshold voltage and oxide thickness assignments for the transistors. Unlike other techniques for low-leakage SRAM design, the proposed technique incurs neither area nor delay overhead. In addition, it results in a minor change in the SRAM design flow. Simulation results with a 65nm process demonstrate that this technique can reduce the total leakage power dissipation of a 64Kb SRAM by more than 50%.

I. Introduction

CMOS scaling beyond 100nm technology node requires not only very low threshold voltages to retain the device switching speeds, but also ultra-thin gate oxides to maintain the current drive and keep threshold voltage variations under control when dealing with short-channel effects [1]. Low threshold voltage results in an exponential increase in the subthreshold leakage current, whereas ultra-thin oxide causes an exponential increase in the gate leakage current. The leakage power dissipation is roughly proportional to the area of a circuit. Since in many processors caches occupy about 50% of the chip area [2], the leakage power of caches is one of the major sources of power consumption in high performance microprocessors.

While one way of reducing the subthreshold leakage is to use higher threshold voltages in some parts of a design, to reduce the gate leakage, it is necessary to use multiple oxide thickness. There are different ways to achieve a higher threshold voltage [3], among them are adjusting the channel doping concentration and applying a body bias. To achieve multiple oxide thicknesses, on the other hand, Arsenic implantation into the silicon substrate before thermal oxidation can be used [4].

In the past, much research has been conducted to address the problem of leakage in SRAMs [5-9]. In [5], for example, the authors used a dynamically controlled sleep transistor to reduce the leakage power dissipation of a large on-chip SRAM. In [6], Kim *et al.* proposed a dynamic threshold voltage method to reduce the leakage power in SRAMs. In their technique, the threshold voltage of the transistors of each cache line is controlled separately by using forward body biasing. By deploying an extra diode in parallel with a sleep transistor connected between the source of NMOS transistors and the ground in an SRAM cell, the authors of [7] reduced both gate and subthreshold leakage currents. In [8], on the other hand, by observing the fact that in ordinary programs most of the bits in data-cache and instruction-cache are zero, the authors proposed using asymmetric SRAM cells to reduce the subthreshold leakage. By including the device-level optimization into circuit-level techniques, reference [9] presented a forward body-biasing technique for active and standby leakage power reduction in cache memories.

Most proposed techniques have hardware overhead and hence increase area of the SRAMs. Furthermore, they try to reduce the subthreshold leakage current only, whereas for sub-100nm technology node, the gate tunneling leakage is comparable to the subthreshold leakage. In this paper we present a method for reducing both subthreshold and gate tunneling leakage current of an SRAM by using different threshold voltages and oxide thicknesses for transistors in an SRAM cell. The proposed technique in this paper has several main advantages over previous techniques:

- it reduces both subthreshold and gate tunneling leakage current,

- it does not involve any hardware overhead,
- it does not have any delay overhead,
- it requires only a minor change in the SRAM design flow, and
- it improves the static noise margin under process variation.

The remainder of this paper is organized as follows. In Section II the structure of an SRAM block is discussed. Section III briefly describes the leakage components. Our idea for reducing the leakage power dissipation is presented in Section IV. Section V shows the experimental results, while Section VI concludes the paper.

II. SRAM Architecture

A typical SRAM, shown in Figure 1, consists of several blocks: cell arrays, address decoder, column multiplexers, sense amplifiers, I/O, and a control circuitry. The functionality and design of every component of an SRAM block can be found in [10]. Figure 2 shows a 6-transistor (6T) SRAM cell. The bit value stored in the cell is preserved as long as the cell is connected to a supply voltage whose value is greater than the Data Retention Voltage (DRV) [11]. This feature, which is due to the presence of cross-coupled inverters inside the 6T SRAM, holds independent of the amount of leakage current. In an SRAM cell, the pull-down NMOS transistors and the pass-transistors reside in the read path. To achieve a high read stability, the pull down transistors are made stronger than the pass-transistors. The pull-up PMOS transistors and the pass-transistors, on the other hand, are in the write path. Although using strong PMOS transistors improves the read stability, it degrades the write-margin. A proper sizing of pass-transistors is required to achieve an adequate write margin [5].

Traditionally all cells used in an SRAM block are identical (i.e., they have the same width, threshold voltage, and oxide thickness for equivalent transistors) which results in identical leakage characteristic for all cells. However, as we will show in this paper, by using non-identical cells, but still with the same layout footprint, one can achieve more power efficient designs.

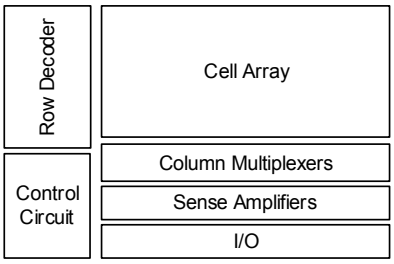

Figure 1. An SRAM block.

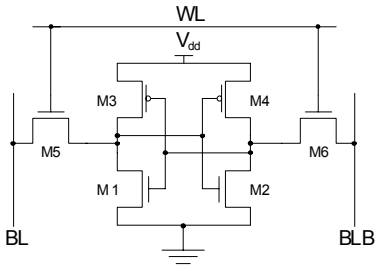

Figure 2. A 6T SRAM cell.

III. Leakage Components

The leakage current of a deep submicron CMOS transistor consists of three major components: junction tunneling current, subthreshold current, and gate tunneling current [12]. In the following, each of these three factors is briefly discussed.

A. Junction Tunneling Leakage

The reversed biased p-n junction leakage has two main components: one is minority carriers' diffusion near the edge of the depletion region and the other is due to electron-hole pair generation in the depletion region of the reverse biased junction [12]. The junction tunneling current is an exponential function of junction doping and reverse bias voltage across the junction.

Since junction tunneling current is a minimal contributor to the total leakage current [12], in this paper we do not attempt to reduce this component of leakage in an SRAM; however, it should be noticed that by applying a forward substrate biasing, junction tunneling current can be reduced [14].

B. Subthreshold Leakage

Subthreshold leakage is the drain-source current of a transistor when the gate-source voltage is less than the threshold voltage. More precisely, subthreshold leakage happens when the transistor is operating in the weak inversion region. The subthreshold current depends exponentially on threshold voltage, which results in large subthreshold current in short channel devices.

To reduce the subthreshold leakage of an SRAM cell, one can increase the threshold voltage of all or some of the transistors in the cell. The drawback of this technique is an increase in read/write delay of the cell. If the threshold voltage of the pull up PMOS transistors is increased, the write delay increases whereas the effect on the read delay would be negligible. On the other hand, if

the threshold voltage of the pull down NMOS transistors is increased, the read delay increases whereas the effect on the write delay would be marginal. By increasing the threshold voltage of the pass transistors both read and write delays increase. Due to the delay of sense amplifiers and output buffers in a read path, the write delay of an SRAM cell tends to be smaller than its read delay. Therefore, one can think of reducing the subthreshold leakage by increasing the threshold voltage of the PMOS transistors as long as the write delay is less than the read delay.

B. Gate Tunneling Leakage

Electrons (holes) tunneling from the bulk silicon through the gate oxide into the gate results in gate tunneling current in an NMOS (PMOS) transistor. Gate tunneling current is composed of three major components: (1) gate to source and gate to drain overlap current, (2) gate to channel current, part of which goes to source and the rest goes to drain, and (3) gate to substrate current. In bulk CMOS technology, the gate to substrate leakage current is several orders of magnitude lower than the overlap tunneling current and gate to channel current [15]. On the other hand, while the overlap tunneling current dominates the gate leakage in the OFF state, gate to channel tunneling dictates the gate current in the ON condition. Since the gate to source and gate to drain overlap regions are much smaller than the channel region, the gate tunneling current in the OFF state is much smaller than gate tunneling in the ON state [15].

If SiO_2 is used for the gate oxide, PMOS transistors will have about one order of magnitude smaller gate leakage than NMOS transistors [15, 16]. Therefore, in an SRAM cell, the power saving achieved by increasing the oxide thickness of the PMOS transistors is marginal.

The subthreshold and gate tunneling leakage currents of an SRAM cell storing "0" are shown in Figure 3.

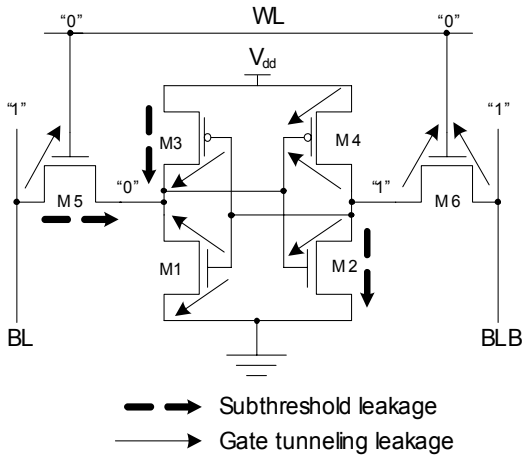

Figure 3. Subthreshold and gate tunneling leakage in an SRAM cell storing "0".

IV. Hybrid Cell SRAM

Due to the non-zero delay of the interconnects of the address decoder, word-lines, bit-lines, and the column multiplexer, read and write delays of cells in an SRAM block are different. Simulations show that for a typical SRAM block, depending on the number or rows and columns, the read time of the closest cell to the address decoder and the column multiplexer may be 5-15% less than that for the furthest cell. This gives an opportunity to reduce the leakage power consumption of an SRAM by increasing the threshold voltage or oxide thickness of some of the transistors in the SRAM cells.

It is known that each additional threshold voltage or oxide thickness needs one more mask layer in the fabrication process, which increases the fabrication cost [17]. As a result, in many cases, only two threshold voltages and two different oxide thicknesses are utilized in circuits. So, in the remainder of this paper we concentrate on the problem of low-leakage SRAM design in a dual-V_t and dual-T_{ox} technology. However, it is possible to extend the results to handle more than two threshold voltages and two oxide thicknesses.

The simulation results in this paper are obtained by using a 65nm technology using HSPICE [21] simulation with BSIM4 model [18], which accurately models subthreshold and gate leakage current. The value of low threshold voltage in this technology is 0.20V, while the high threshold voltage is 0.25V. The thin oxide thickness is 17A^o while the thick oxide is 19A^o. The supply voltage of this technology is 1.0V and all simulations are done at 100°C.

A. SRAM Cell Configurations

To reduce the subthreshold leakage power consumption of a cell, the threshold voltage of all or some of the transistors of the cell can be increased. When the threshold voltages of all transistors within a cell are increased, the subthreshold leakage reduction is the highest. However, since this scenario has the worst effect on the read delay of the cell, the number of memory cells that can be changed is low. Thus, we consider other configurations which have smaller subthreshold leakage reductions, but lower delay penalties. On the other hand, as mentioned in Section III, to reduce the gate tunneling leakage of an SRAM cell, only the oxide thickness of the pull down NMOS transistors and pass-transistors need to be increased. Although this is seemingly desirable from a

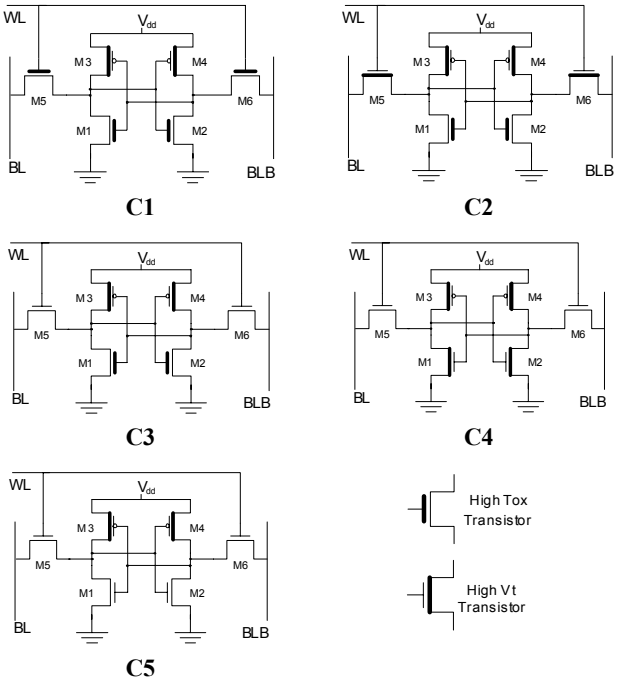

Figure 5. Non-dominated configurations

low power point of view, it is not applicable for all cells in the cell array; thin oxide needs to be used in the cells far from the address decoder and sense amplifiers. It should be emphasized that increasing the oxide thickness also increases the threshold voltage, resulting in a decrease in the subthreshold leakage. In the following, high V_t transistors refer to those transistors whose threshold voltage have been modified by e.g., increasing the channel doping, not the ones whose threshold voltage has been boosted as a result of increasing the oxide thickness. To make the memory cells more manufacturable, unlike [8], we use a symmetric cell configuration, which means the symmetrically located transistors within an SRAM cell have the same threshold voltages and oxide thicknesses. Thus, there are 32 different possibilities for assigning high and low threshold voltages and oxide thickness to the transistors within a cell. Since increasing the oxide thickness also increases the threshold voltage of a transistor, we do not increase the oxide thickness and threshold voltage of a transistor at the same time because the delay penalty will be too high. Therefore, the number of different configurations is reduced to eighteen (there are two choices for the pair of PMOS transistors and three choices for each of the pull-down NMOS pair and pass-transistor pair). Each configuration has a different effect on read and write delays of cells. By simulating all configurations, the dominated ones, i.e., the ones with higher leakage and longer read/write delay than at least

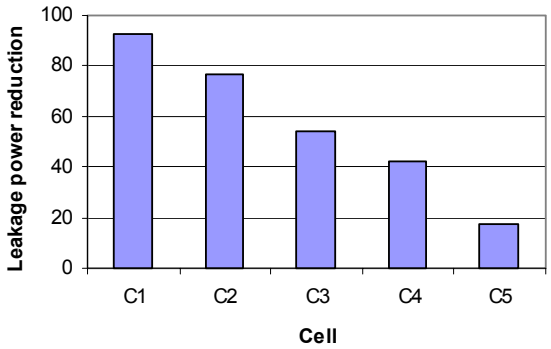

Figure 6: Leakage power reduction of each configuration

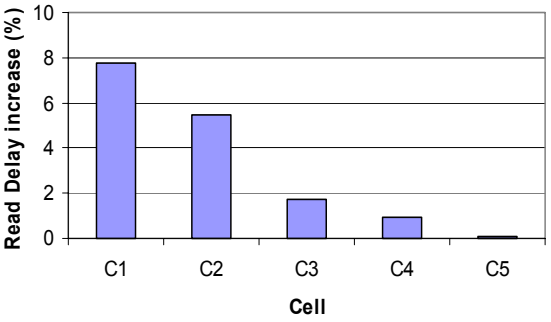

Figure 7: Read delay increase for each configuration

one other configuration, are eliminated. Five configurations remain as shown in Figure 5.

The configurations shown in Figure 5 have different leakage power consumptions. The decrease in leakage power consumption of each configuration, compared to the initial configuration where all threshold voltages are low and all oxide thicknesses are thin, is shown in Figure 6. One can see that the C1 cell, for which all four NMOS transistors have thick-T_{ox} and the PMOS transistors have a high threshold voltage, exhibits 90% lower leakage compared to the initial cell C0, for which all transistors have low-T_{ox} and low-V_t.

Figure 7 shows the effect of each configuration on the read delay of a cell. While the first configuration, C1, has a large read delays, C5 has almost the same delay as C0.

B. Static noise margin

The static noise margin (SNM) of a CMOS SRAM cell is defined as the minimum DC noise voltage necessary to flip the state of a cell [19]. SRAM cells are especially sensitive to noise during a read operation because the "0" storage node rises to a voltage higher than ground due to a resistive voltage divider comprised of the pull-down NMOS transistor and the pass transistor. If this voltage is high enough, it can change the cell's value. We have simulated the SNM of all cell configurations during a

read operation. These simulations show that compared to the original cell, the SNM of four out of five new cells, improves and only in C5 (i.e., when only PMOS transistors are high threshold), it degrades by 10%. The numbers in Table 1 show the increase of SNM for each cell configuration compared to the SNM of the original cell, C0.

Table 1: SNM improvement of different cells

Cell Type	SNM Improvement
C1	43.8%
C2	28.8%
C3	3.75%
C4	5.0%
C5	-10.0%

Since using C5 degrades SNM, we do not use this configuration in the design of the low-power SRAM. Thus, the only valid configurations that can be used instead of C0 are C1, C2, C3, and C4. Note that since all eliminated configurations are dominated by C1, C2, C3, or C4, by removing C5 from the list of configurations, we do not need to consider any other configuration.

C. *Hybrid Cell Assignment*

Starting from a pre-designed SRAM with all low-Vt and low-T_{ox} cells (C0 case), to design a hybrid-cell SRAM, we need to find out the slowest read and write delays Next, considering the configurations shown in Figure 1 and the fact that C1 has the least leakage power consumption among all configurations, we replace as many C0 cells as possible with C1 cells in such a way that the access delay of the replaced cells will not be larger than the slowest access delay in the original SRAM design. After that, we try to replace the remaining C0 cells with other configurations in descending order of the leakage saving, i.e., C2, C3, and C4. Since modifying V_t and T_{ox} does not change the footprint of a cell, the hybrid cell assignment does not change the layout of the cell array and can be performed without affecting the overall SRAM module floorplan.

It is noteworthy that using C1 cells, whose pass transistors have thick gate oxides, decreases the word-line and (to some extent) the bit-line capacitances, and thereby, reduces the delay of the word-line and bit-line. Notice that if the control signals of the SRAM, i.e., pre-charge, read-mux, write-mux, and sense-enable, have not been properly designed i.e., they cannot tolerate this small decrease in the delay, then the control circuitry needs to be modified. The required modification will, however, be minor.

V. Simulation Results

To study the efficiency of the proposed technique, a 1GHz, 64Kb SRAM with a 64-bit word has been designed and simulated in a predictive 65nm CMOS technology with 1.0V for the supply voltage and 0.2V for the low threshold voltage and 17A^o as the gate oxide thickness. All local and global interconnects, including bit and bit-bar lines, word line, and decoder wires have been modeled as distributed RC circuits. The SRAM module consists of two cell arrays, each of which has 32 rows and 1024 columns. For optimizing the delay of the decoder, the pre-decoding

Table 2: The utilization frequency (in percentage) of each cell in the final solution.

Cell Type	Utilization Frequency (%)	
	$V_{t,high}$=0.25V	$V_{t,high}$=0.30V
C0	20%	30%
C1	37%	37%
C2	9%	0%
C3	24%	33%
C4	10%	0%

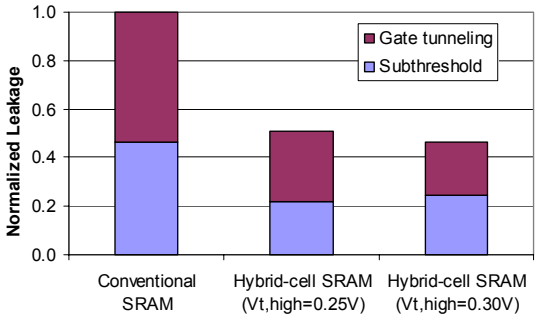

Figure 8. The contribution of subthreshold and gate tunneling leakage

scheme has been used as described in [12]. After designing the SRAM module, the hybrid cell assignment has been applied to the design as described in Section IV. To achieve an order of magnitude reduction in gate tunneling leakage, the thicker oxide is assumed to be 2A^o more than the thin oxide [20], i.e., it is 19A^o. On the other hand, to consider the tradeoff between leakage current reduction of a cell and increase in its access time

as a result of increasing the threshold voltage, two different values have been considered for the high threshold voltage, 0.25V and 0.30V. All simulations on the SRAM have been done at 100°C. Table 2 shows the utilization frequency of each cell configuration in the final low-power SRAM. Figure 8 shows the contribution of the subthreshold and gate tunneling in the leakage power dissipation of the conventional SRAM and hybrid-cell SRAM. As demonstrated in this figure, when $V_{t,high}$=0.25V the leakage power reduction of the SRAM is 49.2%, while $V_{t,high}$=0.30V results in 53.5% leakage power reduction.

VI. Conclusions

In this paper we have presented a novel technique for low-leakage SRAM design. Our technique is based on the fact that due to the non-zero delay of interconnects of the address decoder, word-line, bit-line and the column multiplexers, cells of an SRAM have different access delays. Thus, the threshold voltage or the thickness of gate oxide of some transistors of cells can be increased without degrading the performance. By using five different configurations for the SRAM cells, we have achieved a low-leakage SRAM without scarifying performance and area. By applying the proposed technique to a 64Kb SRAM in 65nm technology node, the total leakage power dissipation of the SRAM has been reduced by 53.5%.

References

[1] Y. Taur, "CMOS scaling and issues in sub-0.25 μm systems," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, pp. 27–45..

[2] C. Molina, C. Aliagas, M. Garcia, A. Gonzalez, and J. Tubella, "Non redundant data cache," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2003, pp. 274–277.

[3] N. Sirisantana, L. Wei, and K. Roy, "High performance low power CMOS circuits using multiple channel length and multiple oxide thickness," in *Proc. Int. Conf. on Computer Design: VLSI in Computers and Processors*, 2000 , pp. 227-232.

[4] M. Togo, K. Noda, and T. Tanigawa, "Multiple-thickness gate oxide and dual-gate technologies for high-performance logic embedded DRAMs," in *IEDM Tech. Dig.*, 1998, pp. 347–350.

[5] K. Zhang *et al*., "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, Apr. 2005, pp. 895-901.

[6] C. Kim and K. Roy, "Dynamic Vt SRAM: a leakage tolerant cache memory for low voltage microprocessor," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2002, pp. 251–254.

[7] A. Agarwal and K. Roy, "A noise tolerant cache design to reduce gate and sub-threshold leakage in the nanometer regime," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2003, pp. 18–21.

[8] N. Azizi, F. Najm, and A. Moshovos, "Low-leakage asymmetric-cell SRAM," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 4, Aug. 2003, pp. 701-715.

[9] C. H. Kim, J. Kim, S. Mukhopadhyay, and K. Roy, "A forward body-biased low-leakage SRAM cache: device, circuit and architecture considerations," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 3, Mar. 2005, pp. 349-357.

[10] R. Preston, "Register files and caches," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, pp. 285-308.

[11] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. Int. Symposium on Quality Electronic Design*, Mar. 2004.

[12] V. De *et al*., "Techniques for leakage power reduction," in *Design of High-Performance Microprocessor Circuit, Circuits*, A. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, pp. 285-308.

[13] D. Weiss, J. Wuu, and V. Chin, "The on-chip 3MB subarray based 3rd level cache on an Itanium microprocessor," in *Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2002, pp. 112–113.

[14] A. Agarwal, C. Kim, S. Mukhopadhyay, and K. Roy, "Leakage in nano-scale technologies: mechanisms, impact and design considerations," in *Proc. of Design Automation Conf.*, 2004, pp. 6-11.

[15] D. Lee, D. Blaauw, and D. Sylvester, "Gate oxide leakage current analysis and reduction for VLSI circuits," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 2, Feb. 2004, pp. 155-166.

[16] F. Hamzaoglu and M. Stan, "Circuit-level techniques to control gate leakage for sub-100nm CMOS," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2002, pp. 60–63.

[17] A. Sirvastava, "Simultaneous Vt selection and assignment for leakage optimization," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2003, pp. 146-151.

[18] http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html

[19] A. J. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits,* vol. 36, no. 4, Apr. 2001, pp. 658–665.

[20] A. Sultania, D. Sylvester, and S. Sapatnekar, "Tradeoffs between gate oxide leakage and delay for dual Tox circuits," in *Proc. Design Automation Conf.*, 2004, pp. 761-766.

[21] http:// synopsys.com/products/mixedsignal/hspice/hspice.html