

# Ultra-Low-Voltage Robust Design Issues in Deep-Submicron CMOS

Andrei Vladimirescu<sup>1,2</sup>,

Yu Cao<sup>1</sup>, Olivier Thomas<sup>2</sup>, Huifang Qin<sup>1</sup>, Dejan Markovic<sup>1</sup>, Alexandre Valentin<sup>2</sup>, Razvan Ionita<sup>2</sup>, Jan Rabaey<sup>1</sup>, Amara Amara<sup>2</sup>

<sup>1</sup>University of California, Berkeley

<sup>2</sup>Institut Supérieur d'Electronique de Paris

**Abstract-** Design challenges for operating CMOS circuits fabricated in 0.13 $\mu\text{m}$  and finer technologies at ultra-low-voltages are analyzed. The design goal consists in minimizing energy by reducing  $V_{\text{DD}}$  while maintaining delay and yield at acceptable levels in the presence of increasing variability of process parameters. First, an estimation model developed to accurately predict operation of bulk- and SOI-CMOS in subthreshold is described. The relation between yield, energy, delay and device parameter distributions is examined next along with tradeoffs necessary to achieve the desired performance point. The main objective of minimizing energy is explored for SRAM cells by predicting the minimum  $V_{\text{DD}}$  based on the data-retention voltage, DRV, and, acceptable signal-to-noise margins, SNM. Experimental data from a 4kB-SRAM test chip in 0.13 $\mu\text{m}$  CMOS are presented demonstrating a 90% leakage reduction potential in standby under reduced bias of 250mV.

## I. INTRODUCTION

The rapid scaling of silicon technology has enabled a dramatic increase in functionality and complexity in integrated circuits (ICs) design. One negative side effect of technology scaling is that leakage power increases significantly from one technology generation to the next and represents one of the main challenges in future system-on-a-chip (SoC) integration. Another important development in state-of-the-art ICs is that an increasing chip area is dedicated to embedded memory which also contributes an important percentage of leakage. In battery-supported applications with low duty-cycles, such as the Pico-Radio wireless sensor nodes [1], cellular phones, PDAs, or, medical devices, leakage power dominates system power consumption and limits battery life.

Leakage in CMOS circuits has been addressed by reducing the supply voltage,  $V_{\text{DD}}$ , and, by operating circuits in the subthreshold region. Performance evaluation of various circuit topologies, logic gates, flip-flops, memory cells, etc., require the availability of an estimation model for a MOS transistor in the subthreshold region much simpler than the BSIM3 simulation model. A simple yet realistic physics-based model [2] which describes the subthreshold drain current of a MOSFET taking into account the body- and drain-voltage dependencies in addition to the commonly modeled gate-voltage dependence, is presented in Sec. II.

Another level of complexity in ULV design is added by the fact that precise control of chip manufacturing becomes increasingly difficult and expensive in the nanometer domain. As a result, circuit

performance exhibits much wider variability, leading to increasing yield degradation in successive technology generations. Tradeoffs among energy, delay and yield, and, the key variables are described in Sec. III.

Embedded SRAM design is one critical area for current and future technology generations; over 50% of the transistors on a SOC today are used for memory implementation and this percentage will increase to 80% according to the ITRS road map. Therefore, suppressing leakage current in memories is critical in low-power design. By scaling down the standby supply voltage  $V_{\text{DD}}$  to the lowest theoretical limit for preservation of the information in standby, the Data Retention Voltage (DRV), leakage power can be substantially reduced. The presence of noise from various sources must be taken into account by allowing for a static-noise margin (SNM). An exploration of the lower limits of DRV [5] and SNM [6] for an SRAM is presented in Sec. IV corroborated by measurement results from an SRAM test chip.

## II. SUBTHRESHOLD ESTIMATION MODEL

This model adequately describes the pseudo-triode and pseudo-saturation regions of MOS transistors operated below  $V_{\text{T0}}$  observed in measurements, see Fig. 1, but not included in commonly used models. This model can be applied for predicting bulk- or SOI-CMOS circuit operation.

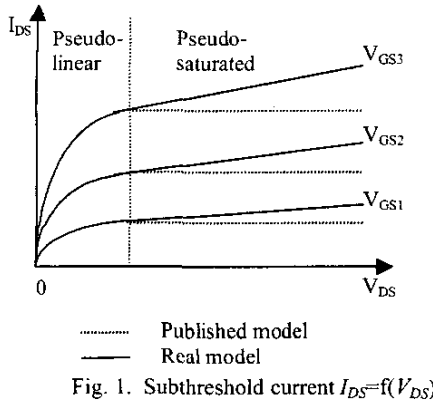
The complete  $I_{\text{DS}}$  equation [2] is

$$I_{\text{DS}}(V_{\text{GS}}) = W \cdot \left( \frac{I_0}{W_0} \right) \cdot 10^{\left( \frac{|V_{\text{GS}}| - |V_{\text{T0}}|}{S} \right)} \cdot \left( 1 - \exp\left(-\frac{V_{\text{DS}} \cdot m}{V_{\text{TH}}}\right) \right) \cdot (a + \lambda \cdot V_{\text{DS}}) \quad (1)$$

$$V_{\text{T0}} = V_{\text{T0}} + \gamma \left( \sqrt{2\psi_F - V_{\text{BS}}(1-\alpha)} - \sqrt{2\psi_F} \right)$$

where  $V_{\text{T0}}$  is the constant-current pseudo-threshold voltage corresponding to a current  $I_0$  of a reference transistor of width  $W_0$ ,  $S$  is the subthreshold slope and  $m$ ,  $\alpha$  and  $\lambda$  are fitting parameters.

Eq. 1 is used in Sec. IV to derive analytical expressions for the logic switching threshold, DRV and SNM to predict the performance of CMOS inverters and SRAM cells operated below  $V_{\text{T0}}$ .

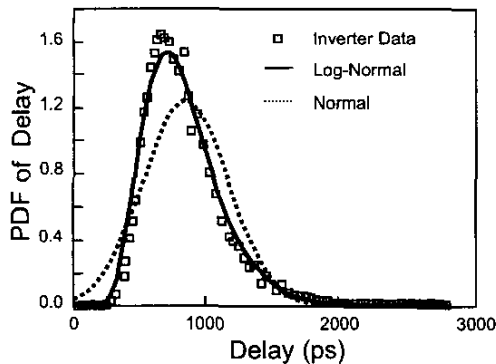
Fig. 1. Subthreshold current  $I_{DS}=f(V_{DS})$ 

### III. YIELD OPTIMIZATION WITH ENERGY-DELAY CONSTRAINTS

#### A. Statistical Model

As circuit parametric variations aggravate in advanced technology, yield emerges as an important figure-of-merit in circuit design. Energy, a major constraint in ULV design, can be effectively reduced by tuning supply voltage ( $V_{DD}$ ), threshold voltage ( $V_{T0}$ ), and device width ( $W$ ), at the expense of circuit yield which degrades during this process. Delay variability on the other hand is most sensitive to fluctuations in  $L$ ,  $V_{DD}$ , and  $V_{T0}$ , making these parameters the focus of this study [4].

The detailed technology specifications are summarized in Table I. Yield-energy-delay information is extracted from Monte-Carlo SPICE simulations of digital circuits. At low supply voltages (or low  $V_{DD}/V_{T0}$ ), the shape of the circuit performance distribution deviates from a Gaussian shape, commonly used for circuit performance, due to its nonlinear dependence on process and voltage parameters. Fig. 2 illustrates the delay distribution of an inverter chain implemented in 130nm technology. The probability density function (PDF) of 5000 Monte-Carlo simulation points has an asymmetrical shape, skewed toward the lower end of the data range. In order to account for this asymmetry, the lognormal model is used in this study providing a very

Fig. 2 Log-normal statistical model ( $V_{DD}=400\text{mV}$ ,  $V_{T0}=130\text{mV}$ )

good fit to the inverter data as seen in Fig. 2.

The main test circuit studied is a five-stage delay-optimized inverter chain driving a 1pF capacitor load. Conventional variation-tolerant design strategies generally rely on optimization of the design

at nominal conditions while reserving adequate margins at design corners. However, as performance variability is exacerbated, these margins widen to the extent that this methodology becomes inaccurate. Furthermore, from experiments and simulations, it is observed that performance variations are not only sensitive to circuit parametric fluctuations due to within-die and die-to-die process variations, but also to operation conditions, such as  $V_{DD}$  and temperature.

#### B. $V_{DD}$ and $V_{T0}$ Optimization

Monte-Carlo simulations confirm the dependence of delay, normalized delay variation ( $3\sigma/\mu$ ) and switching energy of the inverter chain on  $V_{DD}$  and  $V_{T0}$ . These results also show that the delay variability changes at a similar rate to the nominal delay with  $V_{DD}$  and  $V_{T0}$  tuning, and, as a result, circuit yield degrades when  $V_{DD}$  is lowered or  $V_{T0}$  is increased. The sensitivity of delay on parameter variations increases with decreasing  $V_{DD}$  or  $V_{DD}/V_{T0}$ , leading to wider variability. Therefore, tuning  $V_{DD}$  and  $V_{T0}$  is the most effective technique for balancing the reduction of active power consumption and leakage with sacrifices in performance.

TABLE I. Technology Specifications

|                             | Mean  | $3\sigma/\text{mean}$ |
|-----------------------------|-------|-----------------------|
| $L_{\text{eff\_nmos}}$ (nm) | 71    | 15%                   |
| $L_{\text{eff\_pmos}}$ (nm) | 80    | 15%                   |
| $V_{T0\_nmos}$ (V)*         | 0.24  | 15%                   |
| $V_{T0\_pmos}$ (V)*         | -0.34 | 15%                   |
| $V_{DD}$ (V)                | 1.2   | 10%                   |

\* This is for minimal sized inverter:  $W_n=0.585\mu\text{m}$ .

Improved circuit yield however, requires higher  $V_{DD}$  and lower  $V_{T0}$ . The results show that yield degrades at a much slower rate with lower  $V_{DD}$  compared to the sharp reduction of switching energy. For example, when  $V_{DD}$  is reduced from 1.2V to 0.4V (at  $V_{T0}=0.241\text{V}$ ), yield degrades from 90% to 60%, while energy is saved by 80%, see Fig. 3. This relationship indicates that the energy-yield tradeoff is favorable for low power design.

#### C. Device Width ( $W$ ) Tuning Effect

Besides  $V_{DD}/V_{T0}$  tuning, device sizing is another effective technique to optimize energy consumption under delay constraint. In the case of a delay-optimized inverter chain, the switching power consumption can be effectively reduced with a comparatively small delay penalty by systematically shrinking the tapering factor along the path:

$$W_{i+1}/W_i = W_i/W_{i-1} + u \cdot W_i \quad (2)$$

where  $W_i$  is the device width at stage  $i$ , and  $u$  is a sizing factor. When  $u=0$ , we get the delay-optimal condition where the sizes of all stages are uniformly scaled. Energy is reduced when  $u<0$  and successive inverter sizes increase at a slower rate. Due to the dependence of  $V_{T0}$  variation on  $W$ , a device with smaller  $W$  suffers larger  $V_{T0}$  fluctuations and thus a worse yield. When  $u$  goes from 0 to -0.2, the size of the last inverter in the chain is 14 times smaller and the

switching energy is reduced by another 10%; meanwhile, yield may degrade by more than 5%, depending on operation conditions. Due to the yield concern, it is preferred to use large values of  $W$  in low power design.

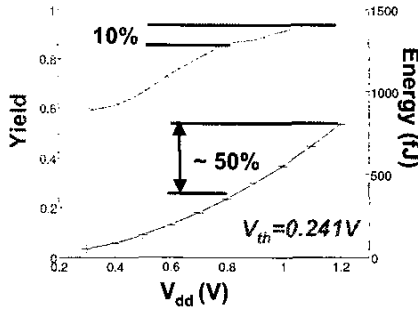


Fig. 3 Energy-Yield Tradeoff

It has been observed that performance variability is relatively insensitive to circuit topology and device length ( $L$ ). Therefore future yield-aware designs need to consider mainly dependence on  $V_{DD}$ ,  $V_{T0}$  and to a certain degree  $W$ .

#### IV. SRAM CELL ANALYSIS AT MINIMUM $V_{DD}$

##### A. Data-Retention Voltage

An analytical model for DRV as a function of process and design parameters is derived for design space explorations. This model is verified using simulations as well as measurements from a 4KB SRAM IP in a  $0.13\mu\text{m}$  technology.

In a standard SRAM cell, see Fig. 4, when  $V_{DD}$  scales down to DRV, the voltage transfer curves (VTC) of the internal inverters degrade to such a level that noise margin of the SRAM cell degrades to zero, as illustrated in Fig. 5. Using the notations of Fig. 4, this condition is given by:

$$\left. \frac{\partial V_1}{\partial V_2} \right|_{\text{Left inverter}} = \left. \frac{\partial V_2}{\partial V_1} \right|_{\text{Right inverter}}, \text{ when } V_{DD} = \text{DRV} \quad (3)$$

If  $V_{DD}$  is reduced below DRV, the inverters flip to the biased state determined by the deteriorated VTC and lose the capability to preserve the stored data.

Assuming that during standby

$$V_1 \approx 0 \text{ and } V_2 \approx V_{DD}, \quad (4)$$

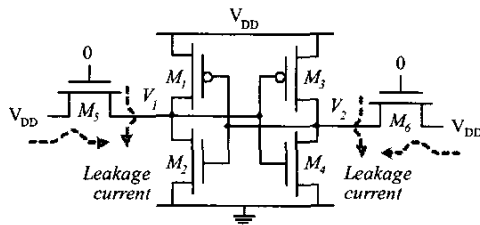


Fig. 4 6-T SRAM Cell

and assuming that the bit-lines are set to  $V_{DD}$ , we first estimate the initial value of DRV,  $DRV_1$ , using the approximations in Eq. (4):

$$DRV_1 = \frac{1}{S_p^{-1} + S_n^{-1}} \cdot \log \left[ \frac{A_4}{A_2 A_3} \left( \frac{A_5}{S_n} + \frac{A_1}{(S_p^{-1} + S_n^{-1})^{-1}} \right) \right], \quad (5)$$

where  $S_p$  and  $S_n$  are the subthreshold slopes for the p and n transistors, respectively, and,

$$A_i = W_i \left( \frac{I_0}{W_0} \right) 10^{-\frac{I_0}{S_i}} \quad (6)$$

Based on this first expression of DRV accurate expressions for  $V_1$  and  $V_2$  can be obtained leading to an exact formulation for DRV rather than applying the approximation in Eq. (4).

The above DRV formula only relies on the values of  $A_i$  and  $S_i$ , which can be easily extracted from transistor characterizations either by simulation or measurement. In addition, it captures the dependence of DRV on process variations in transistor characteristics,  $V_{T0}$ , and sizing  $W_i$ , and, chip temperature  $T$  through  $I_0$ . Based on Eqs. 5 and 6 the impact of these process and design factors on DRV can be formulated as:

$$\Delta DRV = DRV_0 + \sum_i a_i \frac{\Delta(W/L)_i}{(W/L)_i} + \sum_i b_i \Delta V_{T0} + c \Delta T \quad (7)$$

where  $DRV_0$  is the nominal value at room temperature;  $a_i$ ,  $b_i$ , and  $c$  are coefficients. With the  $0.13\mu\text{m}$  technology used and considering an industrial SRAM cell sized for optimal performance, the DRV formula predicts  $DRV_0 = 77\text{mV}$  at perfect matching and  $DRV_0 = 169\text{mV}$  with  $3\sigma$  variations in  $V_{T0}$  and channel length. These analytical results match well with SPICE simulated values of  $78\text{mV}$  and  $170\text{mV}$ , respectively. The model coefficients  $a_i$ 's are extracted from simulations:  $a_1 = 10\text{mV}$ ,  $a_3 = -41\text{mV}$ ,  $a_4 = 11\text{mV}$  ( $a_2$  is negligible). Temperature coefficient  $c$  is extracted as  $0.169\text{mV}/^\circ\text{C}$ , which predicts an increase of  $12.3\text{mV}$  in DRV when  $T$  rises from  $27^\circ\text{C}$  to  $100^\circ\text{C}$ .

##### B. Signal-to-Noise Margin

In an actual SRAM implementation reducing the supply voltage all

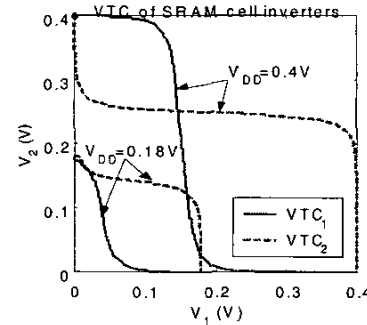


Fig. 5 DRV and SNM Definition

the way down to the DRV is not really an option, as other mechanisms may disrupt the state of the memory cell. Noise on the supply rail as well as radiation particles require an appropriate noise margin be provided. The SNM is defined as the smallest diagonal of the two maximum squares that can be fit into the cross section of the VTC diagrams of the cross-coupled inverters [6], see Fig. 5.

One can estimate the SNM starting from Eq. 3 for a given  $V_{DD}$  and solve for the smallest diagonal. MATLAB has been used [6] to obtain values of SNM for different transistor ratios in the SRAM cell supplied at a set  $V_{DD}$ . The SNM variation is shown in Fig. 6 for a 6-T

cell implemented in SOI-CMOS. The upper curve is comparable to bulk-CMOS with the floating-body effect of the SOI-CMOS cell not considered; if the effect of coupling to the floating body is taken into account SOI-CMOS results in a 30mV reduction in SNM. On the other hand SOI-CMOS has a lower DRV compared to bulk-CMOS due to its higher subthreshold slope according to Eq. 5. The results also show that for reducing DRV and therefore  $V_{DD}$  a high  $W_p/W_n$  needs to be coupled with a low  $W_A/W_D$ , where  $W_A$  and  $W_D$  are the width of the access and driver transistors, respectively. For PD-SOI implementation a 4-T cell may have advantages over 6-T [7].

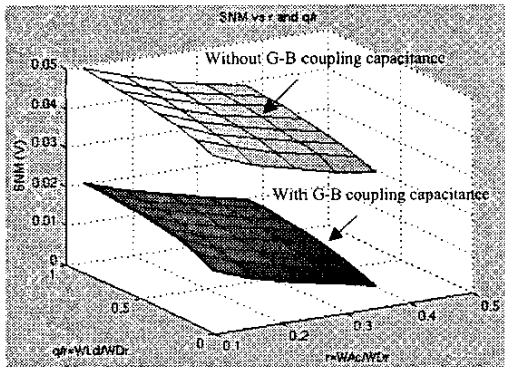


Fig. 6 SRAM Cell SNM vs.  $W_p/W_n$  and  $W_A/W_D$

Simulation verifies the above estimates; assigning a guard band of 100mV above DRV for standby  $V_{DD}$  typically gives 60mV in SRAM cell Static Noise Margin (SNM), and the SNM degrades linearly with the  $V_{DD}$  guard band (becoming zero at the DRV). Soft errors are another element to be considered in the noise budget when setting the guard band above DRV [8].

### C. Measurements

The two main components of the test chip are a 4KB SRAM module and a switch capacitor dc-dc converter.

DRV is measured by monitoring the data retention capability of an SRAM cell with different values of standby  $V_{DD}$ , as shown in Fig. 7. With  $V_{DD}$  switching between active and standby modes, a specific state is written into the SRAM cell under test at the end of each active period ( $t_2$ ), and then read out at the beginning of the next active period ( $t_1$ ). Preservation of the assigned logic state is observed when standby  $V_{DD}$  is higher than DRV, while the state is lost when standby  $V_{DD}$  is below DRV.

The DRV was measured for 50 SRAM cells. The DRV values range from 80mV to 250mV with the mean value of the approximately normal distribution at about 170mV. Such a wide range of DRV uncertainty reflects considerable process variations in fabrication. The 78 mV ideal DRV, assuming perfect process matching as obtained by simulation and analytical modeling, shows up as the lower bound of the measured DRV distribution. With mismatches included, the measured mean value of 170mV matches the analytical model with  $3\sigma$  variations in  $V_{T0}$  and  $L$ . Moreover, temperature dependency of DRV was investigated with measurement. When the test chip was heated up to 100°C, DRV was obtained as 183mV, which verifies the calculation of 181mV from Eq. 7.

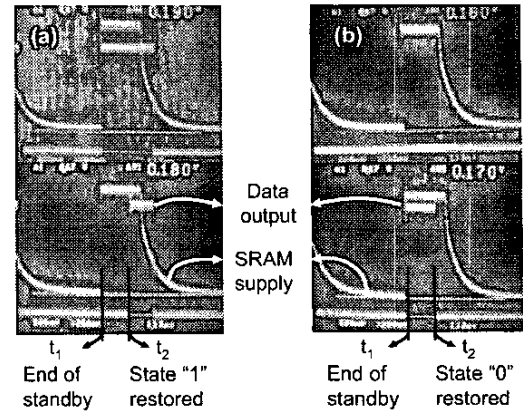


Fig. 7 Waveform of DRV measurement.

(a) DRV = 190mV in SRAM cell 1 with state "1",  
(b) DRV = 180mV in SRAM cell 2 with state "0".

## V. CONCLUSION

This paper has presented a performance evaluation model for digital circuits operating at ULV. First a new subthreshold MOSFET estimation model has been introduced which is applied to accurately predict delay and minimum  $V_{DD}$  for proper operation of SRAM cells.

Monte-Carlo simulations have shown that very favorable tradeoff among delay, energy and yield is possible; for logic circuits in 0.13 $\mu$ m technology lowering the voltage from a nominal 1.2V to 0.4V can achieve an 80-90% reduction in energy for a 30% lower yield. The most important factor is the  $V_{DD}/V_{T0}$  ratio with the device width  $W$  playing only a secondary role.

Reliable operation of SRAM cells at ULV has been analyzed and measures for stability, DRV and SNM, have been analytically derived. Measurements on an SRAM test chip have validated our estimation model and demonstrated that an SRAM cell state can be preserved at sub-300mV standby  $V_{DD}$ , with more than 90% leakage power savings.

## REFERENCES

- [1] J. Rabaey et al, "PicoRadios for Wireless Sensor Networks: The Next Challenge in Ultra-Low-Power Design," *Proc. of ISSCC*, pp. 200-201, San Francisco, Feb 2002.
- [2] O. Thomas, A. Valentian, A. Vladimirescu, and A. Amara, "An Accurate Estimation Model for Subthreshold CMOS SOI Logic", *Proc. ESSCIRC*, pp. 275-279, Florence, Italy, Sep 2002.
- [3] S.R. Nassif, "Design for variability in DSM technologies," *ISQED 2000*, pp. 451-454, 2000.
- [4] Y. Cao, H. Qin, R. Wang, P. Friedberg, A. Vladimirescu, and J. Rabaey, "Yield Optimization with Energy-Delay Constraints in Low-Power Digital Circuits," *Proc. EDSSC*, pp. 285-288, Hong Kong, Dec. 2003.
- [5] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," *ISQED 2004*, pp. 451-454, Mar. 2004.
- [6] O. Thomas, A. Amara, and A. Vladimirescu, "Stability Analysis of a 400 mV 4-Transistor CMOS-SOI SRAM Cell Operated in Subthreshold." *Proc. EDSSC*, pp. 247-250, Hong Kong, Dec. 2003.
- [7] O. Thomas, and A. Amara, "An SOI 4 Transistors Self-Refresh Ultra-Low-Voltage memory cell", *ISCAS*, May 2003, Thailand, Bangkok.
- [8] Y. Nakagome, M. Horiguchi, T. Kawahara, and, K. Itoh, "Review and future prospects of low-voltage RAM circuits," *IBM J. Res & Dev.*, Vol. 47, No. 5/6, pp. 525-552, Sep/Oct 2003.