

Standby supply voltage minimization for deep sub-micron SRAM

Huifang Qin*, Yu Cao, Dejan Markovic, Andrei Vladimirescu, Jan Rabaey

Department of EECS, University of California at Berkeley, Berkeley Wireless Research Center, 2108 Allston Way, Suite 200, Berkeley, CA 94704, USA

Received 17 August 2004; received in revised form 14 March 2005; accepted 20 March 2005

Abstract

Suppressing the leakage current in memories is critical in low-power design. By reducing the standby supply voltage (V_{DD}) to its limit, which is the data retention voltage (DRV), leakage power can be substantially reduced. This paper models the DRV of a standard low leakage SRAM module as a function of process and design parameters, and analyzes the SRAM cell stability when V_{DD} approaches DRV. The DRV model is verified using simulations as well as measurements from a 4 KB SRAM chip in a 0.13 μm technology. Due to a large on-chip variation, DRV of the 4 KB SRAM module ranges between 60 and 390 mV. Measurements taken at 100 mV above the worst-case DRV show that reducing the SRAM standby V_{DD} to a safe level of 490 mV saves 85% leakage power. Further savings can be achieved by applying DRV-aware SRAM optimization techniques, which are discussed at the end of this paper.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: SRAM; Leakage suppression; DRV; Data retention; State preservation; Variation

1. Introduction

Continuous technology scaling over the past four decades has been enabling higher speed and higher integration capacity in VLSI designs. While active power remains constant in the scaling, the chip leakage power increases about five times each technology generation, and becomes one of the main challenges in future CMOS design [1]. In battery-supported applications with low duty-cycles, such as the Pico-Radio wireless sensor nodes [2], cellular phones, or PDAs, under most situations active power only accounts for a small portion of the system power consumption, and leakage power ultimately determines the battery life.

On the other hand, microprocessor designs of today incorporate large memory components, which consume a significant portion of system power budget. For instance, 30% of Alpha 21264 and 60% of StrongARM power are dissipated in cache and memory structures [3]. While activity factor is usually small in memory structures, leakage causes a major part of the memory power

consumption. In 70 nm technology it has been projected that 70% of the cache power budget will be the leakage power [4].

As a result of both the technology scaling and large leakage power dissipation in memory structure, memory leakage suppression is critically important for the success of power-efficient designs, especially ultra-low power (ULP) applications. While the leakage of logic modules in a chip can be effectively controlled by gating off these paths at standby mode, the leakage suppression of memories is especially difficult due to the data retention requirement in such structures. The goal of this work is to develop an effective scheme for SRAM leakage suppression in battery-powered applications such as wireless sensor nodes. The analysis and techniques in this paper focus on the needs of such ULP designs, but are also applicable in general.

A large variety of techniques have been proposed to reduce the leakage of SRAM cells and corresponding peripheral circuits. Since leakage power of the peripheral circuits during idle period can be eliminated by turning off these leakage paths with switched source impedance (SSI) [5], this work focuses on the leakage suppression of SRAM cells only. In the proposed approach, the standby supply voltage (V_{DD}) of the whole memory is minimized with the memory states preserved. As a result of reduced voltage, all the leakage components in an SRAM cell are effectively minimized. This analysis of ultra low voltage reliable data

* Corresponding author. Tel.: +1 510 666 3153; fax: +1 510 883 0270.
E-mail address: huifangq@eecs.berkeley.edu (H. Qin).

retention and its results can also be used for the future SRAM design in ULP applications with aggressively scaled operational V_{DD} .

At the circuit level, the existing most effective low power design methods in minimizing SRAM cell leakage power are to lower supply voltage and increase transistor threshold voltage (V_{th}), both detrimental to the speed of memory read/write operations. For this reason, leakage reduction techniques at this level typically exploit dynamic control of transistor gate-source and substrate-source bias to enhance driving strength in active mode and low leakage paths during standby periods [6]. For example, the driving source-line (DSL) scheme connects source line of the cross-coupled inverters in an SRAM cell to negative voltage V_{BB} during read cycle, and leaves the source line floating during write cycle. As a result, the cell read access time is improved with boosted gate-source voltage and forward bias of the source–substrate junction of the transistors. The write cycle is also improved since the NMOS transistors in the cross-coupled inverter pair are inactive [7]. Another technique is the negative word-line driving (NWD) scheme. It uses low V_{th} access transistors with negative cut-off gate voltage and high V_{th} cross-coupled inverter pair with boosted gate voltage to achieve both improved access time and reduced standby leakage [8]. The dynamic leakage cut-off (DLC) scheme biases the substrate voltages of non-selected SRAM cells at $\sim 2V_{DD}$ for V_{NWELL} and $\sim -V_{DD}$ for V_{PWELL} [9]. A concern in some of these schemes is that the gate voltages of transistors far exceed V_{DD} , which raises reliability issues [9]. All of these techniques achieve enhanced memory operation speed and suppressed standby leakage current at sub-1V supply voltage compared to conventional cell implementation. As discussed at the end of this paper, these schemes can be applied together with the proposed ultra low voltage standby scheme, for an optimal leakage saving.

At the architectural level, leakage reduction techniques include gating-off the supply voltage (V_{DD}) of idle memory sections, or putting less frequently used sections into drowsy standby mode. To achieve optimal power-performance tradeoffs, compiler-level cache activity analysis are employed to balance the potential for saving leakage energy against the loss incurred in extra cache misses. As an example, the cache decay technique applied adaptive timing policies in cache line gating, achieving 70% leakage saving at performance penalty of less than 1% [10]. To further exploit leakage control in caches with large utilization ratio, the approach of drowsy caches allocates inactive cache lines to a low-power mode, where V_{DD} was lowered while preserving memory data [4].

With a conservatively chosen standby V_{DD} in the drowsy caches approach, leakage energy savings of over 70% in a data cache can be achieved [4]. However the question still remains on the lower bound of standby V_{DD} that still preserves the data, namely DRV. Knowledge of DRV

therefore allows a designer to exploit the maximum achievable leakage reduction for a given technology.

Furthermore, in the sub-1V low power VLSI designs of today, the reliability requirement on memories has become the bottleneck in further reducing the system V_{DD} . To enable even more aggressive memory supply voltage minimization, understanding of low voltage SRAM data preservation behavior is required to quantitatively evaluate the SRAM data retention reliability under low V_{DD} and optimize the future SRAM designs for ULV operation conditions.

This paper explores data retention voltage in SRAM cells under realistic conditions of process and design parameter variations. Section 2 develops analytical model of DRV to investigate the dependence of DRV on process and design parameters. SRAM reliability when the standby V_{DD} approaches DRV is analyzed in the same section. To verify the model and further understand the limitations of DRV under realistic conditions, a 4 KB SRAM test chip with dual-rail supply scheme was designed and fabricated in a 0.13 μm technology, as introduced in Section 3. An on-chip switch capacitor (SC) converter is used to generate the standby V_{DD} . Section 4 presents measurement results of the SRAM data preservation and leakage suppression. Using Berkeley Predictive Technology Model, scaling trend of SRAM DRV for future technologies is studied in Section 5. Sizing optimization for a DRV-aware SRAM cell design is discussed as an approach to further minimize leakage current and improve cell robustness. Section 6 concludes current work and proposes future directions.

2. ULV SRAM data retention analysis

The circuit structure of a 6T SRAM cell is shown in Fig. 1a. As in a typical SRAM design, the bitline voltages are set to V_{DD} during standby mode. To facilitate the DRV analysis, this cell can be represented by a flip-flop comprised of two inverters as shown in Fig. 1b [11].

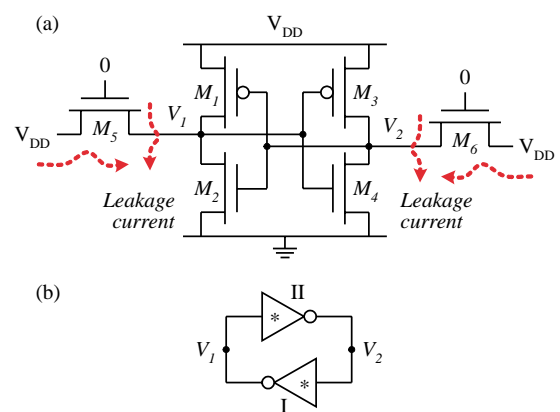


Fig. 1. Standard 6T SRAM cell structure. (a) 6T SRAM cell in standby configuration. (b) Flip-flop representation of the same SRAM cell.

The inverters include access transistors M_5 and M_6 . When V_{DD} is reduced to DRV during standby operation, all six transistors in the SRAM cell are in the sub-threshold region. Thus, the capability of SRAM data retention strongly depends on the sub-threshold current conduction behavior. In order to understand the low voltage data preservation behavior of SRAM and the potential for leakage saving through minimizing standby supply voltage, analytical models of SRAM DRV and cell leakage current are developed in this section.

2.1. Analytical DRV modeling

As the minimum V_{DD} required for data preservation, DRV of an SRAM cell is a measure of its state-retention capability under very low voltage. In order to stably preserve data in an SRAM cell, the cross-coupled inverters shown in Fig. 1(b) must have loop gain greater than one. The stability of an SRAM cell is also indicated by the static-noise margin (SNM) [11]. As shown in Fig. 2, SNM can be graphically represented as the maximum possible square between the voltage transfer characteristic (VTC) curves of the internal inverters from Fig. 1b. When V_{DD} scales down to DRV, the VTC of the internal inverters degrade to such a level that the loop gain reduces to one and SNM of the SRAM cell falls to zero, as illustrated in Fig. 2. Using the notations from Fig. 1, this condition is given by:

$$\frac{\partial V_1}{\partial V_2} \Big|_{\text{inverter I}} \times \frac{\partial V_2}{\partial V_1} \Big|_{\text{inverter II}} = 1, \text{ when } V_{DD} = \text{DRV}. \quad (1)$$

If V_{DD} is reduced below DRV, the inverter loop switches to the other biased state determined by the deteriorated inverter VTC curves, and loses the capability to hold the stored data.

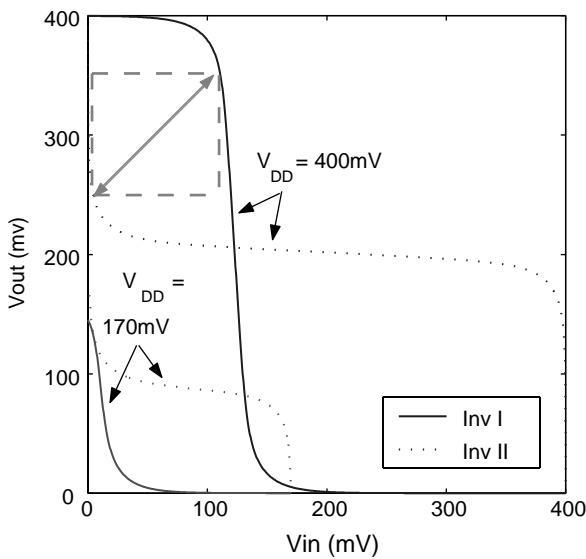


Fig. 2. Deterioration of inverter VTC under low- V_{DD} , with zero SRAM cell noise margins at DRV.

Based on Eq. (1), the DRV of an SRAM cell can be determined by solving the sub-threshold VTC equations of the two internal data-holding inverters, since all the transistors conduct in weak inversion region when V_{DD} is around DRV. The derivation is presented below.

When an SRAM cell (Fig. 1) is in standby mode, the currents in each internal inverter are balanced:

$$\text{Node } V_1 : I_1 + I_5 = I_2, \quad (2)$$

$$\text{Node } V_2 : I_3 + I_6 = I_4. \quad (3)$$

Assuming that the original state stored in SRAM cell is:

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

and that the bit-lines are connected to V_{DD} during standby, I_6 is negligible and Eq. (3) can be simplified to:

$$\text{Node } V_2 : I_3 = I_4. \quad (5)$$

In Eqs. (2) (3) and (5), I_i is the sub-threshold current of the i_{th} transistor (Fig. 1). Assuming room-temperature standby operation, I_i can be considered as dominated by the drain-source leakage in current technology (i.e., ignoring gate leakage and other leakage mechanisms which have minor contribution compared to the sub-threshold current), I_i is modeled as in [12]:

$$I_i = \beta_i I_0 \exp\left(\frac{-V_{th,i}}{n_i v}\right) \exp\left(\frac{V_{gs,i}}{n_i v}\right) (1 - e^{-V_{ds,i}/v}), \quad (6)$$

where $v = kT/q$ is the thermal voltage, equal to 26 mV when $T = 27^\circ\text{C}$; β_i is the transistor (W/L) ratio, I_0 is the leakage current of a unit sized device at $V_{gs} = 0$ and $V_{ds} \gg v$, T is the chip temperature, and n_i is the sub-threshold factor, (sub-threshold swing divided by 60 mV at room temperature). If we further define:

$$I_{off,i} = \beta_i I_0 \exp\left(\frac{-V_{th}}{n_i kT/q}\right), \quad (7)$$

I_i can be expressed as:

$$I_i = I_{off,i} \exp\left(\frac{V_{gs}}{n_i kT/q}\right) (1 - e^{-V_{ds}/(kT/q)}). \quad (8)$$

Substituting the current models in Eqs. (7) and (8), which are functions of V_1 , V_2 , V_{DD} , T , and other technology parameters, into Eqs. (2) and (5), we obtain the VTCs of the inverters in the cell. Then, together with Eq. (1), the value of the DRV (and the corresponding V_1 and V_2) can be derived.

A general solution to these equations requires numerical iterations. To avoid the iterations, we first estimate the initial value of DRV, i.e. $\text{DRV}^{(0)}$, using the approximations in Eq. (4):

$$\begin{aligned} \text{DRV}^{(0)} = & \frac{kT/q}{n_2^{-1} + n_3^{-1}} \log \left[(n_3^{-1} + n_4^{-1}) \frac{I_{off,4}}{I_{off,2} I_{off,3}} \right. \\ & \left. \times \left(\frac{I_{off,5}}{n_2} + \frac{I_{off,1}}{(n_1^{-1} + n_2^{-1})^{-1}} \right) \right] \quad (9) \end{aligned}$$

Then, using $DRV^{(0)}$, the approximations in Eq. (4) are refined as:

$$V_1 = \frac{kT}{q} \frac{I_{\text{off},1} + I_{\text{off},5}}{I_{\text{off},2}} \exp\left(\frac{-DRV^{(0)}}{n_2 kT/q}\right), \quad (10)$$

$$V_2 = DRV^{(0)} - \frac{kT}{q} \frac{I_{\text{off},4}}{I_{\text{off},3}} \exp\left(\frac{-DRV^{(0)}}{n_3 kT/q}\right). \quad (11)$$

With Eqs. (10) and (11) available, we can refine the calculation of DRV and a final expression is obtained:

$$DRV = DRV^{(0)} + \left[\frac{V_1}{2} + \frac{(DRV^{(0)} - V)^{n_2/2}}{2} \right]. \quad (12)$$

The above DRV formula only relies on the values of n_i , which can be easily extracted from transistor characterizations, either by simulation or measurement. For the industrial technology we studied, $n = 1.25$ for both PMOS and NMOS.

2.2. DRV sensitivity to variations

Process variation and temperature fluctuation are the main imperfections that cause degradations in circuit performance. For an SRAM cell, mismatch between two internal inverters has a strong impact on its DRV. As an example, Fig. 3 shows the simulated deteriorated SRAM inverter VTC under 200 mV V_{DD} , for cases both without variation and with 3σ local variations in L and V_{th} . An exaggerated SNM decrease is a clear result of the worst-case local mismatch among transistors, as indicated by the small opening between VTC curves with variations.

To further illustrate the impact of different process variation components on DRV, SPICE simulation data in

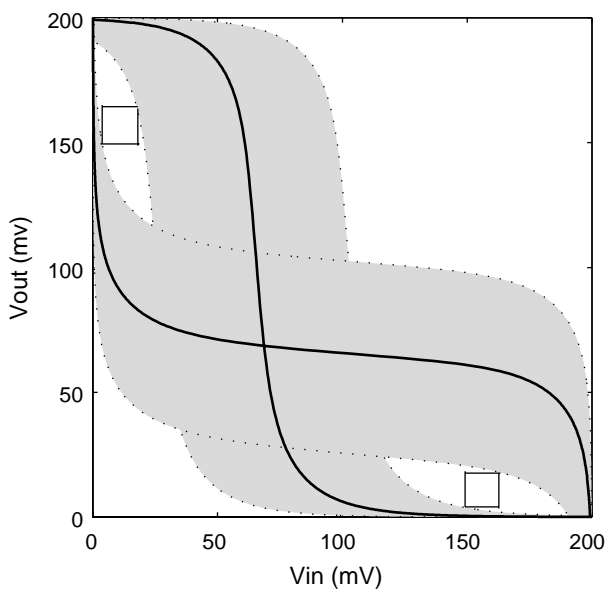


Fig. 3. VTC of SRAM cell inverters under 3σ variation in L and V_{th} . (Solid lines: ideal case with no variation.)

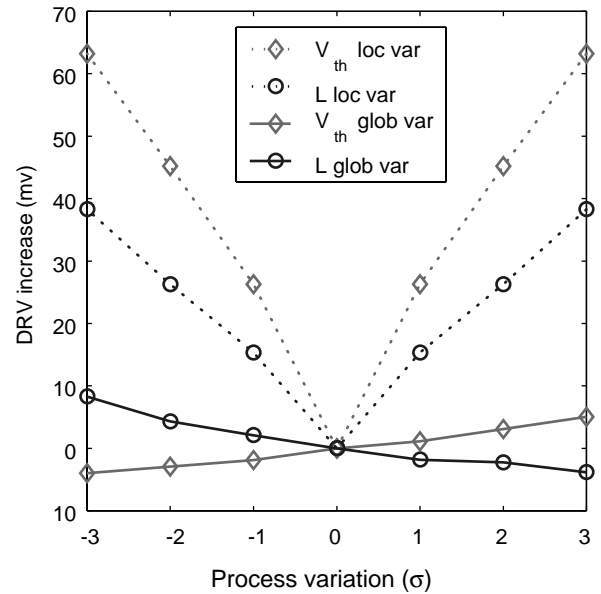


Fig. 4. DRV sensitivity to local and global parameter variation.

Fig. 4 plots the change in DRV value versus the magnitude of variations in an SRAM cell. As observed, local mismatches among transistors result in substantial DRV increase. Based on a $0.13 \mu\text{m}$ technology model, with 3σ local V_{th} mismatch DRV can be 70 mV higher than the ideal case with perfect matching. At the same time, the global shifting in both V_{th} and L , which affect both inverters (Fig. 1b) in the same direction and does not change the matching, has a much weaker impact on DRV. This is because the local mismatch changes the relative drive-strength of transistors. As indicated in Eq. (9), such drive-strength mismatch between same type of transistors (such as M_2 and M_4) causes a substantial increase in DRV value and results in a reduced SNM, as shown in Fig. 3. On the other hand, when the drive-strength of all transistors are affected in a uniform way, such as the result of a global shifting in V_{th} or L value, their impacts compensate each other and result in small change of DRV.

The chip temperature fluctuation is another global variable that has weak influence on DRV since it affects all the transistors in an SRAM cell almost uniformly. Simulation results in Fig. 5 compare the impact of process and temperature variation on DRV. As observed, the DRV increases about 100 mV in the presence of 3σ local mismatch in V_{th} and L , while the temperature impact is much smaller. When T changes from 27 to 100°C , DRV rises only about 13 mV.

In our analytical DRV-modeling Eqs. (9)–(12) are based on the individual leakage current of each SRAM cell transistor, so they capture the dependencies of DRV on process parameters (I_{off} and n), sizing β_i , and chip temperature (T). For a first-order analysis, the impact of these variation factors on DRV can be extracted from Eq. (9) as:

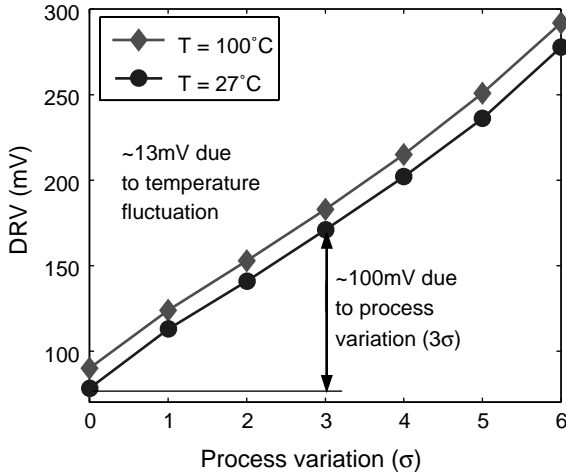


Fig. 5. SRAM cell DRV under process and T variations.

$$\begin{aligned}
 \text{DRV} &= \text{DRV}_{\text{matched}} + \Delta\text{DRV} \\
 &= \text{DRV}_{\text{matched}} + \sum_i a_i \frac{\Delta\beta_i}{\beta} + \sum_i b_i \Delta V_{\text{th}i} \\
 &\quad + c\Delta T
 \end{aligned} \tag{13}$$

where $\text{DRV}_{\text{matched}}$ is the data-retention voltage of a perfectly matched SRAM cell (i.e. no variations or with only global variation on all transistors) at room temperature; a_i , b_i , and c are fitting coefficients for each individual transistor. The $\Delta\beta$ and $\Delta V_{\text{th}i}$ terms in this model represent the local variation on individual transistors. ΔT is the overall chip temperature fluctuation. Since there is usually a small change in the $\text{DRV}_{\text{matched}}$ value caused by global variation, this model focuses on capturing the impact of individual, local, transistor variation on DRV. Considering an industrial SRAM cell design in the 0.13 μm technology used in this study, the model coefficients a_i 's are extracted from SPICE simulations as follows: $a_1 = 10 \text{ mV}$, $a_3 = -41 \text{ mV}$, $a_4 = 11 \text{ mV}$ (a_2 is negligible), assuming that the original data stored in the cell is $V_1 = 0$, $V_2 = V_{\text{DD}}$. Temperature coefficient c is extracted as $0.169 \text{ mV}/^\circ\text{C}$, which predicts an increase of 12.3 mV in DRV when T rises from 27 to 100 $^\circ\text{C}$.

The DRV predictions by Eq. (13) match well with SPICE simulations over a wide range of design parameters and their variations. This is illustrated in Table 1, which summarizes results obtained by SPICE and our analytical model from Eq. (13). The 3σ process variation condition used in this table assumes 3σ worst-case local mismatch in V_{th} and L for all six transistors in the SRAM cell. It should be noted that as a first order analysis, the model in Eq. (13) does not capture the cross-term dependency between parameters of different transistors. For example, the value of PMOS sizing ratio (β_p) affects the DRV sensitivity to NMOS size variation ($\Delta\beta_n$). For a more accurate analysis, the current model in Eq. (13) should be extended to capture these effects.

Table 1
DRV(mV) under variations and different cell sizing

DRV conditions	Spice (mV)	Model (mV)
Ideal (w/o variations)	77	78
w/ 3σ variation in V_{th} and L	170	169
200% PMOS sizing w/ 3σ V_{th} and L variation	136	138
200% NMOS sizing w/ 3σ V_{th} and L variation	182	180
T at 100 C w/ 3σ V_{th} and L variation	183	182

2.3. SRAM standby stability analysis

To reliably preserve data in an SRAM cell at ULV standby mode, certain noise margin needs to be ensured by assigning an appropriate guard-band in standby V_{DD} above the DRV. This section presents SNM analysis as a guide to understanding this guard band requirement and relationship between SNM, V_{DD} , and DRV.

SNM of an SRAM cell can be calculated in many different ways: maximum square between the normal and mirrored VTC, small-signal loop-gain, Jacobian of the Kirchoff equations, coinciding roots. These methods are well researched and it was shown that they are all equivalent [13]. Similar to [11], we take the loop-gain approach of analyzing the SNM as the maximum value of noise that can be tolerated by the flip-flop before changing states. As shown in Fig. 6, two noise sources, V_n , are inserted to assure the worst-case noise scenario when the noise is present in both gates in the same way [11].

Following the methodology of DRV derivation in Section 2.1 with inserted static noise

$$V_{\text{GS}2} + V_n = V_2 \tag{14}$$

$$V_{\text{GS}4} - V_n = V_1, \tag{15}$$

we obtain a zero-order approximation for SNM from the condition of marginal stability, that is the unity loop-gain. The maximum noise corresponding to the unity gain is given by:

$$\begin{aligned}
 \text{SNM} &= \frac{2}{3} V_{\text{DD}} - \frac{nkT/q}{3} \ln \left(\frac{2I_{\text{off},4}}{nI_{\text{off},2}I_{\text{off},3}} \right) \\
 &\quad - \frac{nkT/q}{3} \ln \left(\frac{I_{\text{off},5}}{n} + \frac{2I_{\text{off},1}}{n} e^{\text{SNM}(nkT/q)} \right)
 \end{aligned} \tag{16}$$

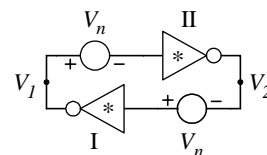


Fig. 6. Flip-flop representation of SRAM cell with inserted static noise, V_n .

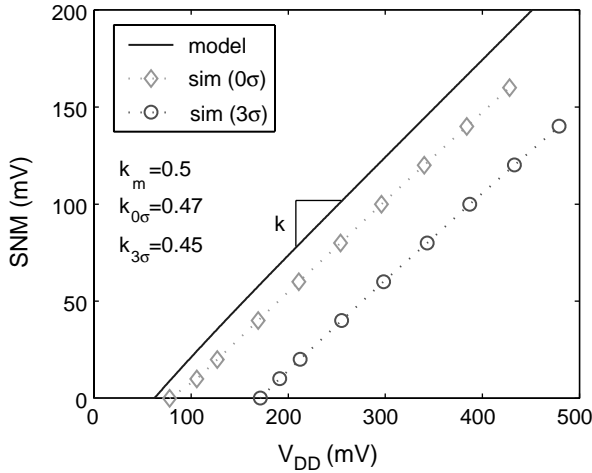


Fig. 7. Static noise margin (SNM) as a function of the SRAM supply voltage, V_{dd} . Slope of a first-order linear model agrees with simulation results.

where we assume ideal case with equal sub-threshold slope factor for NMOS and PMOS transistors, and also make approximations from Eq. (4). The above formula does not have closed-form solution, but can be solved iteratively. For each V_{DD} , the SNM value after five iterations is shown in Fig. 7. This zero-order model closely predicts the slope of the SNM- V_{DD} line, compared to simulation data for cases under 3σ local mismatch and ideal case without variation.

Furthermore, from the linear relationship indicated in Fig. 7 we can adopt simple linear macro-model given by:

$$\text{SNM} = k(V_{dd} - \text{DRV}), \quad (17)$$

Further expansion of Eq. (16) and comparison with Eq. (17) yields following approximation of the k factor:

$$k \approx \frac{2}{3+n}, \quad (18)$$

where $I_{\text{off},5}$ from Eq.(16) is neglected due to exponential nature of the other term under the logarithm. This approximation is valid for $\text{SNM} > nkT/q$. With $n=1.25$, we obtain $k=0.47$, which exactly matches the simulated data shown in Fig. 7. The result in Eq. (18) means that smaller sub-threshold factor is desirable for higher noise tolerance in standby mode.

This linear correlation of SNM and the standby V_{DD} guard-band voltage facilitates the SRAM design for reliable data retention under low voltage. For example, in order to achieve a 50 mV SNM under 3σ local process variations, the SRAM standby V_{DD} needs to be 100 mV higher than the corresponding DRV. In the event of radiation particle strikes, other special actions may be needed to combat the soft errors, such as adding additional storage capacitors, or applying error correction schemes.

2.4. SRAM standby leakage modeling

The total leakage of an SRAM cell in the subthreshold standby mode can be calculated as:

$$I_{\text{leak}} = (I_1 + I_5) + I_4 \approx (I_{\text{off},1} + I_{\text{off},5}) + I_{\text{off},4} \quad (19)$$

where $I_{\text{off},i}$ is defined in Eq. (7). After the standby SRAM V_{DD} is determined, the leakage power P_{leak} under designed standby V_{DD} is:

$$P_{\text{leak}} = V_{DD}I_{\text{leak}} = (\text{DRV} + V_{\text{gb}})I_{\text{leak}}. \quad (20)$$

where V_{gb} stands for the guard-band voltage in standby V_{DD} . Leakage power P_{leak} as provided in Eq. (20) represents the minimum leakage power required for reliable ULV data retention in standard industrial SRAM design.

3. Ultra low voltage SRAM standby: design and implementation

To obtain silicon verification of the presented DRV model and explore the potential of SRAM leakage suppression with ultra low standby V_{DD} , a 4 KB SRAM test chip with dual rail standby control was implemented in a $0.13 \mu\text{m}$ technology. Designed for ultra low-power applications, this scheme puts the entire SRAM into a deep sleep mode during the system standby period. As shown in Fig. 8, the SRAM supply rails are connected to the standard V_{DD} and the standby V_{DD} through two big power switches. The test chip consists of a standard 4 KB SRAM module and custom on-chip switch-capacitor (SC) converter that generates the standby V_{DD} with 85% conversion efficiency. The SRAM is an industrial IP module, which was embedded into the chip layout with no change in the original design. The SRAM cell transistor sizes are the same as analyzed in Section 2. Compared to the existing SRAM leakage control techniques, the simplicity of this scheme leads to minimized design effort and therefore minimum extra power necessary to support control circuitry.

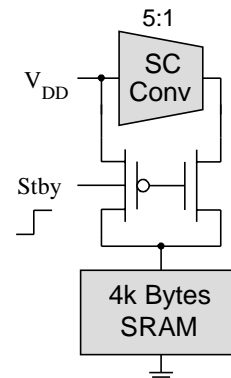


Fig. 8. Standby leakage suppression scheme.

3.1. Dual voltage scheme design considerations

When designing for an ultra low standby V_{DD} , reliability of the SRAM data retention at idle mode is the top design concern. Besides process variations, the other factors that may disturb the memory state preservation are mainly noises on the standby supply rail and radiation particles. In this scheme power supply noise is mostly caused by the output voltage ripple of the SC converter. Therefore, an appropriate noise margin needs to be provided in order to achieve the desired reliability. As analyzed in Section 2.3, assigning a guard band of 100 mV above DRV for standby V_{DD} gives about 50 mV SNM in an SRAM cell. With the 20 mV peak-to-peak ripple on the SC converter output, a guard band of 100 mV provides worst case SNM of 45 mV, which is sufficient for state preservation.

In comparison to power supply noise, the radiation particle events pose a more serious hazard. With parasitic capacitance at the data storage node of about 1fF in a 0.13 μm technology SRAM cell, the critical charge (Q_{critical}) for a 1 V V_{DD} is simulated to be approximately 3 fC. This is the minimum amount of charge injection on the storage node needed to disrupt the state preserved in this cell. For a reduced V_{DD} at 100 mV above the DRV, Q_{critical} is reduced to 0.5 fC. Considering the danger of data loss (i.e. soft error), a larger guard-band is needed. Other options to ensure reliable state preservation include additional storage capacitance [14] or implementation of error-correction schemes.

For a dual supply scheme, other design considerations include the operation delay overhead due to the power switch resistance, memory wake up delay and the power penalty during mode transition. Targeted for ultra low-power applications, the system requirements of this design are much more stringent on power than performance [2]. In this context the concern of the operation delay overhead is not crucial. A 200 μm wide PMOS power switch with 30 Ω conducting resistance is used to connect the memory module to a 1 V active-mode supply voltage. With the same switch the memory wake up time is simulated to be within 10 ns, which is typically a small fraction of the system cycle time in battery-operated applications [15].

The wake up power penalty incurred during switching from the standby mode to the full- V_{DD} mode determines the minimum standby time for the scheme if net power saving over one standby period is to be achieved. This break-even time is an important system-design parameter, as it helps the power control algorithm to decide when a power-down would be beneficial. With the parasitic capacitance information attained from process model, the minimum standby time in this design is estimated to be several tens of microseconds, which is much shorter than the typical system idle time in a battery-supported system.

3.2. Test chip implementation

Layout of the 0.13 μm SRAM test chip is shown in Fig. 9. The two main components are a 4 KB SRAM module and a SC converter. This memory is an IP module with no modifications from its original design. As shown in Fig. 10, a representative five-stage step down dc-dc SC converter topology is selected to implement the on-chip standby V_{DD} generator [16]. Compared to magnetic-based voltage regulators, SC converter provides higher efficiency, smaller output current ripple, and easier on-chip integration for small loads in the microwatt range. The design challenge here resides in handling small output load in the range of 10~20 μW . With such low power operation, power loss incurred by short-circuit currents during phase switching becomes comparable to output power and forms a significant portion of the total power loss.

To maximize power efficiency, it is desirable to minimize both the switching voltage drop and short-circuit current, which have opposite dependence on device sizes. Hence the switch devices need to be carefully designed to balance these two requirements. For example, the NMOS/PMOS switch-type selection should maximize the device gate-source overdrive voltage at conducting mode, and minimize this voltage when the switch is turned off. With these considerations in mind, Fig. 10 shows the optimized design, in which an 85% conversion efficiency is achieved with a 1 V input and an output load equal to the estimated 4 KB SRAM module leakage at standby mode.

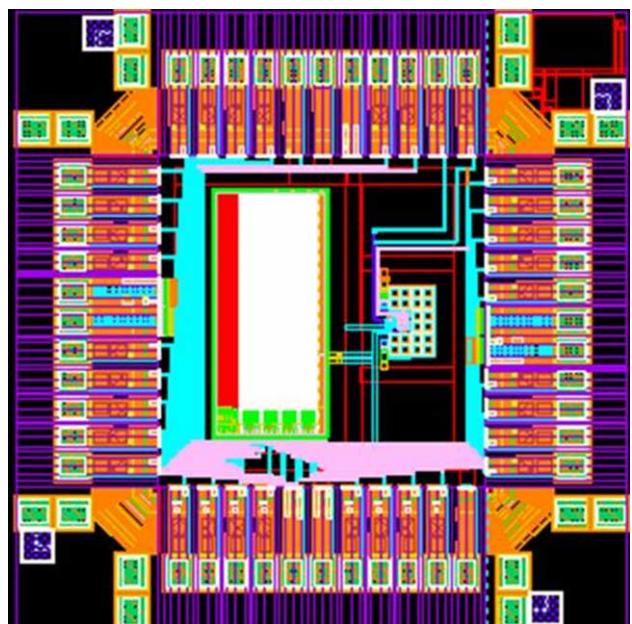


Fig. 9. A 0.13 μm SRAM leakage-control test chip.

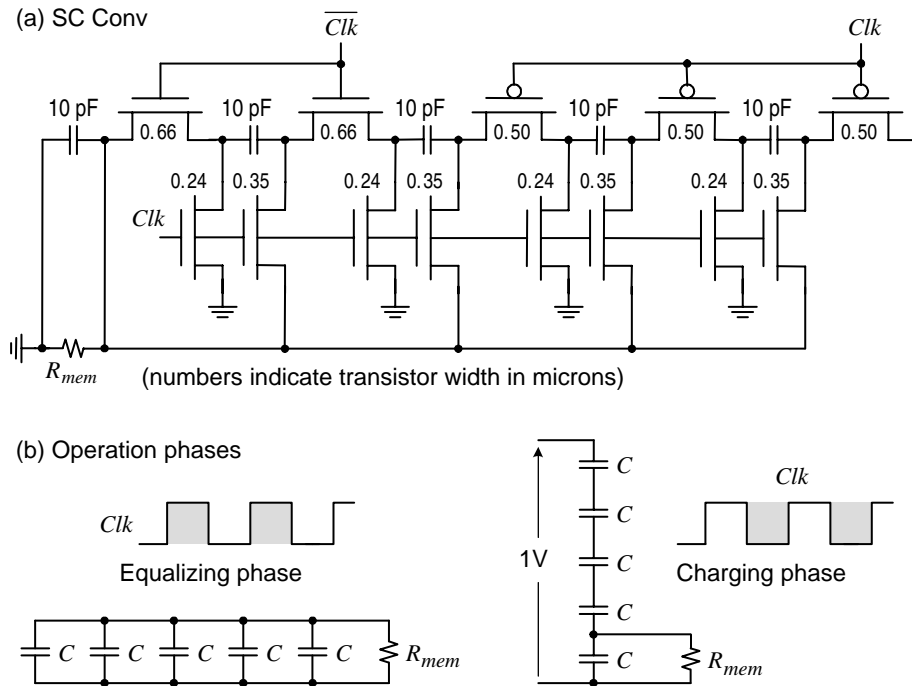


Fig. 10. (a) Schematic of switch capacitor converter, (b) Operation phases.

4. Measurement results

4.1. DRV measurement

The DRV is measured by monitoring the data retention capability of an SRAM cell with different values of standby V_{DD} , as demonstrated in Fig. 11. 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 (top

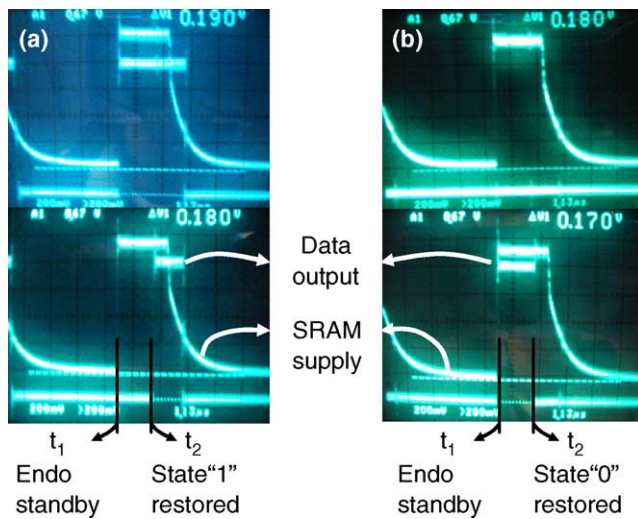


Fig. 11. Waveform of DRV measurement. (a) DRV = 190 mV in SRAM cell 1 with state '1', (b) DRV = 180 mV in SRAM cell 2 with state '0'.

traces), while the state is lost when standby V_{DD} is below DRV (bottom traces), Fig. 11.

Using automatic measurement with a logic analyzer, the DRV of all 32 K SRAM cells on one test chip was measured. Fig. 12 shows the distribution of the 32 K measurement results. The DRV values range from 60 to 390 mV with the mean value around 122 mV. Such a wide range of DRV uncertainty reflects the existence of considerable process variations during fabrication. Due to global variations, the lower end of measured DRV is slightly lower than the 78 mV ideal DRV, assuming perfect process matching. As a result of large process variation, the long DRV tail at the higher end reduces the leakage reduction achievable by minimizing the SRAM standby V_{DD} . To

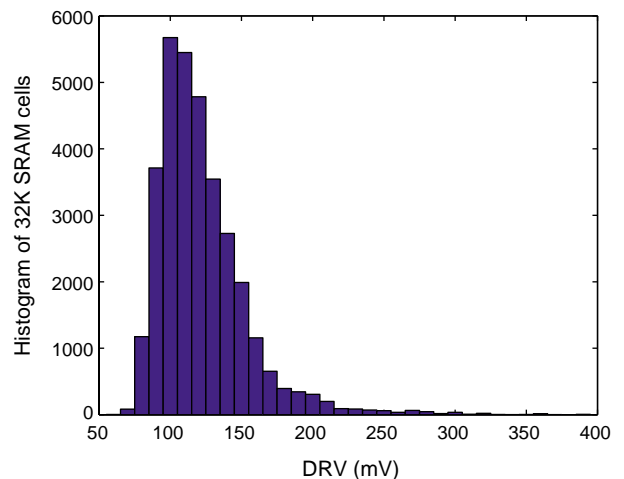


Fig. 12. Measured DRV distribution of a 4 k-byte SRAM chip.

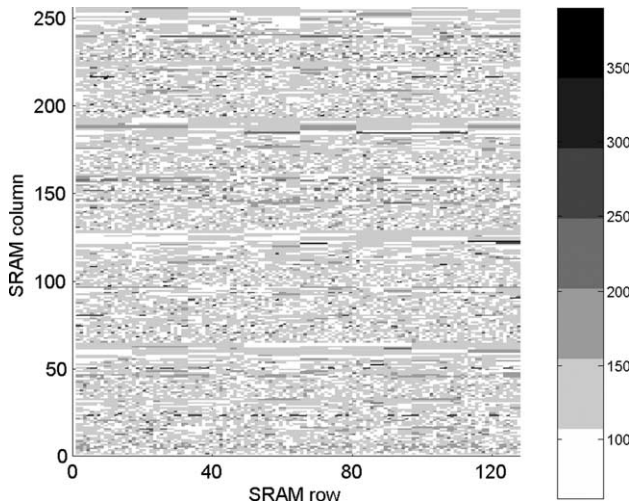


Fig. 13. DRV spatial distribution of a 4 K-byte SRAM chip.

improve the gains in leakage power, more advanced techniques, such as error tolerant schemes, are required to cope with this situation.

Temperature dependency of DRV was investigated experimentally. When the test chip was heated up to 100 °C, a 10 mV increase in DRV is observed. This result matches with the simulated temperature effect on DRV in Fig. 5. As evaluated in Section 2, the analytical DRV model proposed in this work not only predicts the ideal DRV values, but also fully captures the impact of process and temperature variations. Thus, it can serve as a convenient base for further design optimizations.

Furthermore, Fig. 13 shows the first presented spatial distribution plot of DRV on the measured SRAM chip. From the plot, it can be observed that the on-chip DRV distribution is the combination of random within-die mismatches and systematic deviations on the boundaries of SRAM sub-array blocks. The pattern of SRAM DRV spatial distribution can be exploited in the future work of designing effective error tolerant scheme for even more aggressive SRAM voltage scaling.

4.2. SRAM leakage measurement

Leakage measurement result of the 4 KB SRAM is shown in Fig. 14. The leakage current increases substantially when V_{DD} is high. This phenomenon reflects the impact of process variations on SRAM leakage, more specifically the fluctuations in channel length and V_{th} . For short channel transistors, drain-induced-barrier-lowering (DIBL) effect causes V_{th} degradation, resulting in even higher leakage in high- V_{DD} conditions. The shaded area in Fig. 13 indicates the range of measured DRV (60–390 mV). Although the memory states can be preserved at sub-400 mV V_{DD} , adding an extra guard band of 100 mV to the standby V_{DD} enhances the noise robustness of state preservation as discussed in Section 3.1. With the resulting 490 mV standby V_{DD} , SRAM leakage

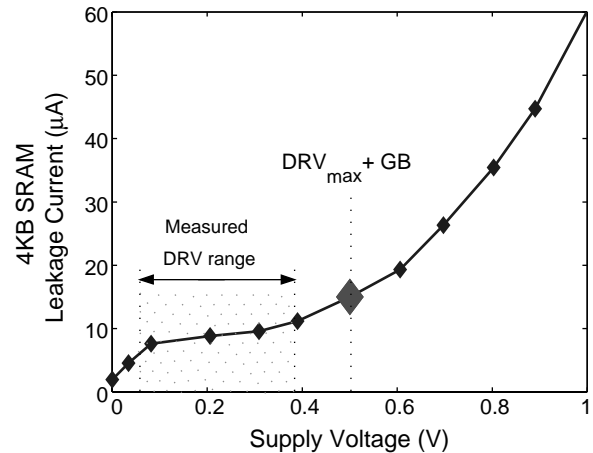


Fig. 14. Measured SRAM leakage current.

current can still be reduced by over 70%. Subsequently the leakage power, as the product of V_{DD} and leakage current, is reduced by about 85% compared to 1V operation.

4.3. Dual rail standby scheme measurement

The dual rail scheme is shown to be fully functional through the DRV measurements. With 10 MHz switch control signal, the SC converter generates the standby V_{DD} with less than 20 mV peak-to-peak ripple. Wake up time of 10 ns is observed during mode transition, while the sleep time spans around 10 μ s. The delay overhead in SRAM read operation is measured to be about 2 \times , which is reasonable for an ultra low-power application where the system clock period is typically 10 times the operation cycle of a low leakage SRAM.

5. DRV-aware SRAM design optimization

While SRAM designs have been well optimized for speed and power metrics, improving DRV for future ultra low-voltage applications poses a new challenge for low-power SRAM designers. This section presents a view of the future DRV scaling with technology, and discusses the effective methods to design for the next generation ultra low-voltage and ultra low-power SRAM.

5.1. Trend of DRV scaling

Exploring the SRAM ULV data preservation is mainly for the interest of ULP designs today, but the technology scaling will soon bring up this topic to memory designers for all-purpose applications. Based on the Berkeley Predictive Technology Model (BPTM) [17], simulation results of DRV scaling are shown in Fig. 15. In this simulation an optimistic estimation of variations is used— σ of device length variation is fixed at 10% of the mean value and σ of V_{th} variation fixed at 10 mV. The resulting SRAM DRV scales

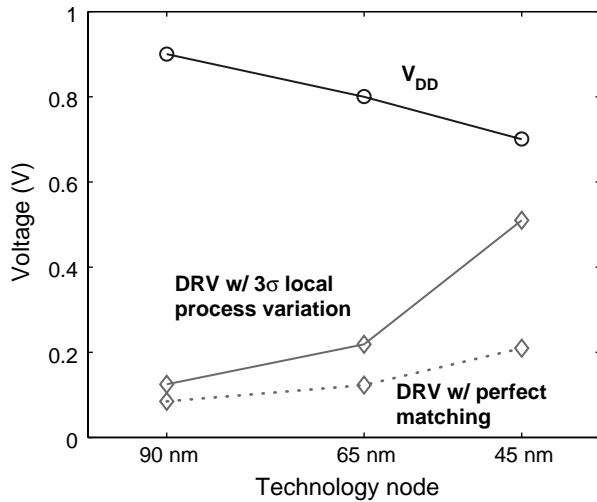


Fig. 15. Mean DRV and V_{DD} scaling trend.

against the trend of V_{DD} reduction and approaches V_{DD} at sub-45 nm nodes.

The up-scaling trend of DRV is a result of both an increasing leakage current (which leads to degradation of I_{on}/I_{off}) and a larger sensitivity of I_{off} to process variations at smaller technology dimensions. As a result of such DRV and V_{DD} scaling, severe reliability hazard of SRAM data preservation under the normal operation voltage is posed around 45 nm technology node. In order to meet the V_{DD} scaling of bulk CMOS technology and low power design requirements, the degradation of DRV must be efficiently coped with.

Measurement and simulation results in previous sections have shown that process variation is the major factor in determining the DRV value of an SRAM cell. Suppressing the process variation is therefore the most effective method to reduce SRAM DRV. As process variation control becomes more difficult in future technology nodes, it will become the limiting factor on SRAM V_{DD} scaling. Temperature variation is shown to be only secondary effect on DRV. It will not be considerable concern in the future either since most ULP applications operate at room temperature.

In SRAM design, effective techniques can help reduce the DRV value at minimum overheads of other metrics, such as area, hardware cost and performance. Following section is focused on SRAM sizing optimization as one of the solutions to DRV reduction. Several other approaches are also discussed as the future work of this study.

5.2. DRV-aware SRAM sizing optimization

DRV analytical model in Eq. (13) suggests that transistor sizing is an important factor that determines the DRV of an SRAM cell. While sizing has long been an effective technique in conventional power and speed optimization,

taking DRV into account is important for future ULP SRAM design.

In conventional performance-optimized SRAM cell, the pull-down NMOS devices are sized about $2\times$ larger than the PMOS devices. These NMOS transistors are also typically designed with a smaller L to minimize the cell area. Although providing good stability at high- V_{DD} , this imbalance in the pull-up and pull-down leakage paths leads to exacerbated VTC deterioration at low V_{DD} , and degrades DRV. The minimum L of NMOS transistors is also highly sensitive to process variations, which lead to an increase of DRV. Therefore, it is of interest to investigate impact of each of the sizing variables on DRV.

Fig. 16 shows simulated DRV over the sizing variables β_i and L_i for different transistors. For each curve all the other sizing variables are fixed at their nominal values from an industrial SRAM cell design. These simulations assume 3σ local process variations in SRAM transistor channel length and V_{th} . X-axis of the plot is the sizing ratio that each variable (β , L) is scaled by. Parameter β_i represents (W/L) ratio of the transistor: the pull-up PMOS (β_p), pull-down NMOS (β_n), and access transistors (β_a). The range of sizing ratios in Fig. 16 is constrained with $\beta_i > 1$ and $L_i > L_{min}$. From the plot in Fig. 16 it can be observed that DRV can be reduced only by increasing β_p or L_n , with impact of L_n being much stronger. This strong DRV dependence on L_n is the result of its small nominal value (chosen for minimum cell area), which is sensitive to the process variations. L_p has a much smaller impact on DRV due to its larger nominal value.

Sizing of access transistors has very small impact on DRV. SPICE simulation shows that for these two transistors, when any one or both of their β sizing ratio and L change within $3\times$ range, the resulting DRV change is less than 5 mV. This is because none of these two access transistors can significantly affect conducting path formed by the strong pull-down NMOS transistor and the weak pull-up PMOS device. Taking the SRAM cell configuration in

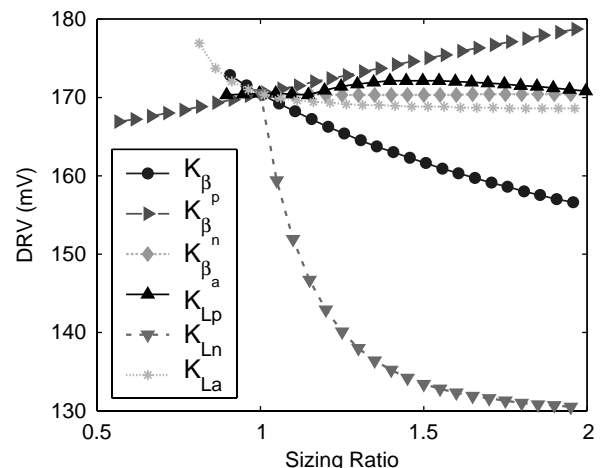


Fig. 16. DRV as a function of sizing parameters β and L .

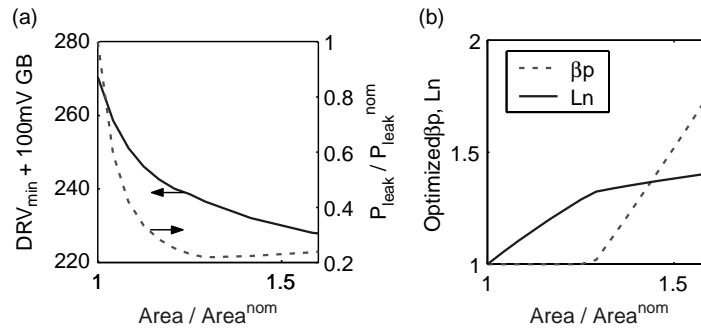


Fig. 17. DRV-aware SRAM optimization with β_p and L_n . (a) Area tradeoff with DRV and leakage power, (b) Optimized β_p and L_n .

Fig. 1 as an example, when $V_1 \approx 0$ and $V_2 \approx V_{DD}$, the inverter formed by conducting PMOS M_3 and leaky NMOS M_4 is the vulnerable path where the state toggling is initiated. Due to the same voltage level at both drain and source, the access transistor M_6 does not leak even though it is connected to the unstable path. Meanwhile the sub-threshold leakage of the other access transistor M_5 is limited by its negative gate-to-source voltage (V_1 becomes a small positive value when V_{DD} approaches DRV). Also because the M_5 leakage combats with the strong NMOS device M_2 , which is conducting at sub-threshold region, the data preservation path formed by transistors M_1 , M_2 , and M_5 is actually very stable as compared to the other path. The weak dependency of DRV on access transistor sizing can be clearly observed from Fig. 16.

In summary, following are the guidelines for DRV-aware ULV SRAM cell optimization, for applications in which the read/write performance is not a major concern:

- (1) Increasing L_n SRAM cell reduces DRV most effectively, followed by increasing β_p .
- (2) Reducing β_n and L_p both improve DRV, but the improvement space is very limited.
- (3) The sizing of access transistors has negligible effect on DRV.

As an example of power and area tradeoff, Fig. 17a plots the leakage power and SRAM cell area when tuning L_n and β_p for the minimized DRV. In this analysis the SRAM cell transistor area is simply modeled as the sum of transistor gate areas. A 30% increase in SRAM cell transistor area brings about 30 mV reduction in DRV and almost 70% additional leakage power saving. In Fig. 17b, L_n first exploits the increase of area budget due to its effectiveness in reducing DRV. Effectiveness of increasing L_n is utilized until L_n is about 25% larger than the nominal value, where its impact on DRV drops and from this point on β_p can be used to further reduce DRV under given area constraint. Although increase of β_p continuously reduces DRV, no more savings in leakage power is attained due to the positive correlation between PMOS leakage and its sizing ratio β_p .

6. Conclusions and future work

This paper explores the limit of SRAM data preservation under ultra-low standby V_{DD} . An analytical model of the SRAM DRV is developed and verified with measurement results. A commercial SRAM module with high- V_{th} process is shown to be capable of sub-400 mV standby data preservation. With additional 100 mV guard band to account for power supply ripple and cosmic particles, leakage power saving of more than 85% can be achieved with an SRAM module under 490 mV standby V_{DD} , compared to 1 V active mode. The DRV is observed to be a strong function of process variation and also SRAM cell sizing. With proper sizing optimization an additional 70% leakage power saving can be achieved with only 30% SRAM cell transistor area increase.

Besides sizing, more variables are being investigated for their impacts on ULP SRAM cell design. Such variables include the transistor V_{th} and body bias voltages. With the control of body bias voltages, the SRAM DRV and leakage current can be dynamically adjusted in different operation modes. Current efforts of this work are on evaluating these effects and attaining the silicon verification.

Besides circuit level techniques, more opportunities exist on architectural level innovations. For example, more SRAM leakage savings can be achieved with assistance from error tolerant schemes when the standby supply voltage is scaled down below DRV. In the future work such architecture-level techniques will be investigated with the goal of achieving even lower power and higher reliability in memory design.

Acknowledgements

The sponsorship of the GSRC MARCO center and fabrication support from STMicroelectronics are greatly acknowledged. The authors would also like to thank to Professor Seth Sanders and Dr. Bhusan Gupta for their enlightening technical advice. The help from Thuan Trinh in automated DRV measurement is sincerely appreciated.

References

- [1] S. Borkar, Design challenges of technology scaling, *IEEE Micro* 19 (4) (1999) 23–29.
- [2] J. Rabaey, et al., Picoradios for wireless sensor networks: the next challenge in ultra-low-power design, *Proceedings of the ISSCC* (2002) 200–201.
- [3] S. Manne, A. Klauser, D. Grunwald, Pipeline gating: speculation control for energy reduction, *International Symposium Computer Architecture* (1998) 132–141.
- [4] N. Kim, Drowsy instruction caches: leakage power reduction using dynamic voltage scaling and cache sub-bank prediction, *Proceedings of the 35th Annual Int'l Symposium Microarchitecture (MICRO-35)*, IEEE CS Press, 2002. pp. 219–230.
- [5] M. Horiguchi, T. Sakata, K. Itoh, Switched-source-impedance CMOS circuit for low standby subthreshold current giga-scale LSI's, *IEEE Journal of Solid-State Circuits* 28 (11) (1993) 1131–1135.
- [6] K. Itoh, Low voltage memories for power-aware systems, *Proceedings of the ISLPED* (2002) 1–6.
- [7] H. Mizuno, T. Nagano, Driving source-line (DSL) cell architecture for sub-1-V High-speed low-power applications, *Digest of technical papers. Symposium on VLSI circuits* (1995) 25–26.
- [8] K. Itoh, A.R. Fridi, A. Bellaouar, M.I. Elmasry, A deep sub-V, single power-supply. SRAM cell with multi-V_t, boosted storage node and dynamic load, *Digest of technical papers. Symposium on VLSI circuits* (1996) 132–133.
- [9] H. Kawaguchi, et al., Dynamic leakage cut-off scheme for low-voltage SRAM's, *Digest of technical papers. Symposium on VLSI circuits* (1998) 140–141.
- [10] S. Kaxiras, Z. Hu, M. Martonosi, Cache decay: exploiting generational behavior to reduce cache leakage power, *Proceedings of the ISCA* (2001) 240–251.
- [11] E. Seevinck, F.J. List, J. Lohstroh, Static-noise margin analysis of MOS SRAM cells, *IEEE Journal of Solid-State Circuits* SC-22 (5) (1987) 748–754.
- [12] J. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, second ed., Prentice-Hall, 2002.
- [13] J. Lohstroh, E. Seevinck, J.D. Groot, Worst-case static noise margin criteria for logic circuits and their mathematical equivalence, *IEEE Journal of Solid-State Circuits* SC-18 (6) (1983) 803–807.
- [14] C. Lage, et al., Soft error rate and stored charge requirements in advanced high-density SRAMs, *Proceedings of IEDM* (1993) 821–824.
- [15] M.J. Ammer, et al., A low-energy chip-set for wireless intercom, *Proceedings of DAC* (2003).
- [16] K.D.T. Ngo, R. Webster, Steady-state analysis and design of a switched-capacitor DC-DC converter, *Proceedings of PESC* (1992) 378–385.
- [17] Y. Cao, T. Sato, D. Sylvester, M. Orchansky, C. Hu, New paradigm of predictive MOSFET and interconnect modeling for early circuit design, *Proceedings of CICC* (2000) 201–204.