FUNDAMENTAL REDUNDANCY VERSUS POWER TRADE-OFF IN STANDBY SRAM

A. Kumar, H. Qin, P. Ishwar, J. Rabaey, and K. Ramchandran*

EECS, University of California, Berkeley, CA – 94720

*ECE, Boston University, Boston, MA – 02215

ABSTRACT

We study the problem of reducing power during data-retention in a standby static random access memory (SRAM). For successful data-retention, the supply voltage of an SRAM cell should be greater than a critical data retention voltage (DRV). Due to circuit parameter variations, the DRV for different cells on the same chip exhibits variation with a distribution having diminishing tail. For reliable data retention, the existing low-power design uses a worst-case technique in which a standby supply voltage that is larger than the highest DRV among all cells in an SRAM is used. Instead, our approach uses aggressive voltage reduction and counters the ensuing unreliability through a fault-tolerant memory architecture. The main results of this work are as follows: (i) We establish fundamental bounds on the power reduction in terms of the DRV-distribution using techniques from *information theory*. For the DRV-distribution of test-chip in [1], we show that 49% power reduction with respect to (w.r.t.) the worst-case is a fundamental lower bound while 40% power reduction w.r.t. the worst-case is achievable with a practical combinatorial scheme. (ii) We study the power reduction as a function of the block-length for low-latency codes since most applications using SRAM are latency constrained. We propose a reliable memory architecture based on the Hamming code for the next test-chip implementation with a predicted power reduction of 33% while accounting for coding overheads.

Index Terms— Memory Architecture, SRAM Chips, Error correction coding, Information Theory, Circuit Optimization.

1. INTRODUCTION

For about four-decades, digital integrated circuits have benefited from an exponential increase of transistor density on the die, known as the Moore's law [2]. Design complexity management, increasing mask cost, process-variations, soft-errors, leakage-power reduction, and power-density management are the most important challenges faced by the circuit designer today. In this paper, we will focus on leakage-power reduction and parametric process-variation aspects of static random access memories (SRAMs). Leakage power reduction is important for low-power applications, like sensor networks, to enable each device to operate within the scavenging power limit [3].

In many chips which consist of SRAM, e.g., sensor network nodes, there are two modes of operation: (i) the *active-mode* in which the SRAM is active for reading and writing, and (ii) the *standby-mode* in which the SRAM retains the data. Standby SRAM power is the dominant power consumption factor in applications which are primarily in the standby mode. The standby-mode power primarily consists of leakage-power which increases with each silicon-technology generation [4]. An effective method to reduce leakage-power is to reduce the supply voltage to the minimum operational point. For this approach, it has been shown that any SRAM cell has a critical voltage (called the data retention voltage or DRV) at which a stored bit (0 or 1) is retained reliably [5].

Leakage-power increase is aggravated by *circuit parameter variations* (process variations) [4]. In Fig. 1, we illustrate the test-chip DRV-variation in the 90nm CMOS process. For reliable data retention, the existing low-power design suggests the use of a standby supply voltage of $200mV$ which is larger than $190mV$, the highest test-chip DRV of Fig. 1.

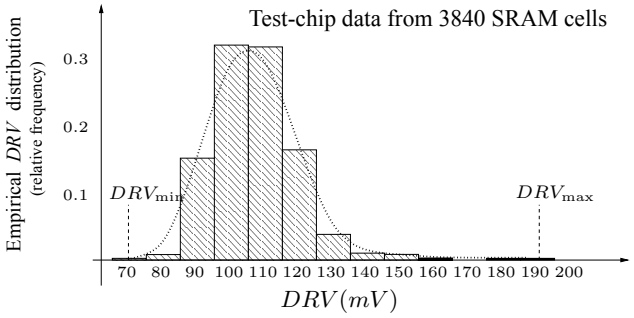

Fig. 1. **Empirical DRV-distribution:** *The intra-chip DRV varies from 70 to $190mV$ for the 90nm CMOS technology. The dotted curve is a smooth fit to the empirical DRV-distribution [1].*

In contrast to the worst-case design, we propose aggressive reduction of the standby supply voltage with error-control coding, thereby ensuring *reliable* data-storage. Using this method, we show the following main results in this paper:

This research was sponsored in part by MARCO, GSRC and supported in part by the NSF under Grant No. CCR-0330514 and Career CCF-0546598. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1. We establish fundamental bounds on the power reduction in terms of the DRV-distribution using techniques from *information theory*. For the test-chip DRV distribution in Fig. 1, we show that 49% power reduction w.r.t. the worst-case is a fundamental lower bound while 40% power reduction w.r.t. the worst-case is achievable with a bounded-distance decoding scheme.

2. We study the power reduction as a function of the block-length for low-latency codes since most applications using SRAM are latency constrained. We propose a reliable memory architecture based on the Hamming code for the next test-chip implementation with a predicted power reduction of 33% while accounting for coding and latency overheads.

The first result states fundamental bounds on the power reduction for a given DRV-distribution. The second result shows that, *while accounting for coding overheads*, a significant portion of the optimum power reduction (33% out of 40%) is achieved by a low-latency Hamming code.

Prior work: The use of error-control codes has been proposed in various SRAM-implementations in the literature (e.g., see [6, 7]). The storage capacity for a memory with stuck-at faults or random errors has been analyzed by Heegard and El Gamal [6]. Our work differs in two important aspects from the existing results: (i) We study power versus redundancy trade-off in SRAMs, unlike previous works which studied reliability versus redundancy trade-offs, and (ii) We account for coding and latency overheads in our analysis and results.

Modeling assumptions: We will model the parametric variation of the DRV by the observed (discrete) probability distribution $\mu(x), x \in \{70, 80, \ldots, 190\}$ (see Fig. 1). The cumulative distribution function is $F(x) = \sum_{z \le x} \mu(z)$. Since the available DRV data is quantized at a resolution of $10mV$, we will sweep the supply voltage in multiples of $10mV$. A cell will retain the stored data successfully if the supply voltage is strictly greater than the cell's DRV voltage. We model the DRV as a fixed voltage after realization. [1]

Notation: In the rest of the paper, the *standby power* will be called as *power* for brevity. The distribution in Fig. 1 will be referred as $F(x)$. The supply voltage will be represented by v_S. The symbol $\mathbb{P}$ will be used for the probability of a set with respect to the distribution $F(x)$. Vectors like $(x_1, x_2, \ldots, x_n)$ will be represented as x_1^n. Finally, $h(t) = -t \log_2 t - (1-t)\log_2(1-t)$ stands for the binary entropy function [9].

Organization: We present the proposed standby SRAM architecture in Sec. 2. We discuss fundamental bounds on power reduction in Sec. 3. We also discuss the power reduction for a few known family of codes in the same section. Finally, we conclude in Sec. 4.

[1] Recent work suggests that the critical voltage at which an SRAM cell works in *read*, *write*, and *store* modes may vary temporally [8]. However, DRV in 90nm CMOS process does not depends significantly on gate-leakage, which is the cause behind temporal variation.

2. PROPOSED STANDBY SRAM ARCHITECTURE

We will present the SRAM cell retention model followed by our proposed standby SRAM architecture. The description of the retention model is important for understanding the architecture.

SRAM cell retention model: For each SRAM cell, there is a data-retention-voltage (DRV), above which the data bit 0 or 1 is stored reliably [5]. However, if the supply voltage is lowered below the DRV, then the stored bit degenerates to a preferred digital state $S \in \{0, 1\}$ [5].

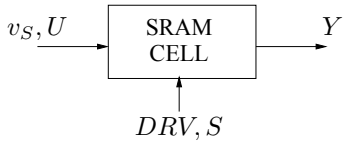

Fig. 2. **Standby SRAM cell:** *The SRAM cell has two statistically independent parameters: (i) a time-invariant positive continuous-valued threshold-voltage DRV, and (ii) a binary bias-state $S \in \{0, 1\}$. The inputs are the supply voltage v_S and a bit $U \in \{0, 1\}$ to be stored. The output is $Y = U$ if $v_S > DRV$ and S otherwise.*

We capture these features of an SRAM cell in the following mathematical model (see Fig. 2). The cell has two statistically independent parameters: (i) a time-invariant, positive and continuous-valued threshold-voltage DRV, and (ii) an equally likely binary bias-state $S \in \{0, 1\}$. The inputs to the cell are the supply voltage v_S and a bit $U \in \{0, 1\}$ to be stored. The retention model for the SRAM cell is as follows:

$$\begin{aligned} Y = U \quad &\text{if} \quad DRV < v_S, \\ = S \quad &\text{if} \quad DRV \ge v_S, \end{aligned} \tag{1}$$

where $Y \in \{0, 1\}$ is the output bit. If $v_S \le DRV$, then there is DRV *failure*. This digital abstraction is sufficient for this paper. The proposed standby SRAM architecture is discussed next.

Proposed standby SRAM architecture: Let the standby supply voltage be $v_S \in \{0, 10, \ldots, 200\}$ in mV at $10mV$ resolution. The worst-case solution is to use $v_S = 200mV$ in which every cell retains the data reliably.

In contrast, we propose an error-protected SRAM as follows. Let $B_1^k = (B_1, B_2, \ldots, B_k)$ be the data vector to be stored. Using a suitable error-control code, B_1^k is encoded into U_1^n and stored in n memory cells ($n \ge k$). Cells have i.i.d. pairs of independent DRV, S realization. [2] The i^{th} stored bit is stuck-at S_i if $DRV_i \ge v_S$, otherwise U_i is successfully retained. At the end of standby, Y_1^n is decoded to $\widehat{B}_1^k$. Let $0 \le i \le 2^k - 1$ be the binary representation of B_1^k. The voltage v_S is chosen such that the outage probability,

$$\mathbb{P}(\text{outage}) = \mathbb{P}\big(\exists i, \text{ such that } \widehat{B}_1^k \ne i | B_1^k = i\big), \tag{2}$$

[2] The assumption that DRV across cells are independent is a worst-case assumption as discussed at the end of Sec. 3.

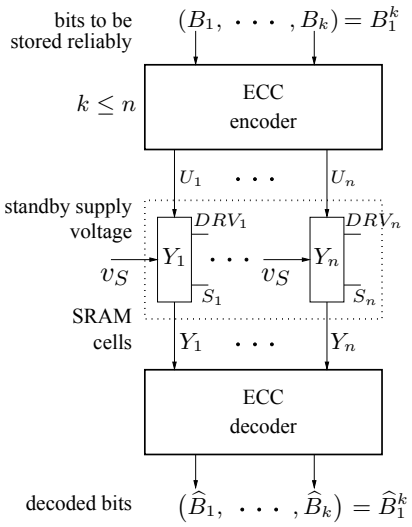

Fig. 3. **Standby-SRAM architecture:** *Let B_1^k be the data vector to be stored. Then B_1^k is encoded into U_1^n and stored in n memory cells. The i^{th} stored bit is stuck-at S_i if $DRV_i \ge v_S$, otherwise U_i is read-out. The decoder reads Y_1^n and outputs $\widehat{B}_1^k$. The voltage v_S is selected such that $\mathbb{P}(\text{outage})$ negligible (see (2)).*

is negligible. This condition ensures that an n-bit row of memory works with high probability for *all* input-words i. The outage failures will be corrected by row-redundancy [10].

Since v_S is a free variable, power per useful-bit (or other performance metrics) can be *optimized* over its range. For an outage of ϵ, we define the power per bit as,

$$\mathcal{P}_\epsilon(v_S) := \frac{1}{k} \cdot \big(\text{Total standby power}\big). \tag{3}$$

If ϵ can be made arbitrarily small by choosing $n \longrightarrow \infty$, then the power per bit function will be called as $\mathcal{P}(v_S)$. The dependence of power per bit on v_S will be established next.

3. POWER PER BIT BOUNDS

In this section, we will derive DRV-distribution dependent fundamental bounds on the power per bit $\mathcal{P}(v_S)$. We first discuss the standby power dependence on the supply voltage.

3.1. Power dependence on the supply voltage

Let T_s be the standby duration. Let $E_{\mathcal{C}}$ be the average encoder-decoder computational energy (over codewords B_1^k) of the error-control code $\mathcal{C}$. The total-power (including coding) is,

$$P_T = P_L + \frac{E_{\mathcal{C}}}{T_s}, \tag{4}$$

where P_L is the total leakage-power. The leakage-current in the $60 - 200mV$ range is approximately linear in the supply voltage, i.e., $I_L = G v_S$, where G is a constant. This is confirmed by our test-chip leakage-current measurements. Thus, the power per bit of the SRAM cell is,

$$\mathcal{P}_\epsilon(v_S) = \frac{n}{k} \cdot G v_S^2 + \frac{E_{\mathcal{C}}}{k T_s}, \tag{5}$$

where the code $\mathcal{C}$ has an outage ϵ [3].

3.2. Fundamental bounds on the power reduction

For deriving bounds, we note the following important points: (i) For $T_s \longrightarrow \infty$, the encoder decoder energy gets normalized to zero. Under this condition, the standby power is minimum and will be explored first, (ii) We will account for the coding and latency overheads after establishing fundamental benchmark asymptotic bounds (see Sec. 3.3 and Sec. 3.4), and (iii) The outage $\epsilon > 0$ can be made arbitrarily small in an asymptotic setting, i.e., when $n \longrightarrow \infty$. The DRV-failure probability is given by,

$$p(v_S) = \sum_{z \ge v_S} \mu(z). \tag{6}$$

Then, we have the following theorem:

Theorem 3.1 *Let v_S be the standby supply voltage and $p(v_S)$ be as in (6). For each voltage $v_S : p(v_S) < 0.25$, the minimum power per bit satisfies,*

$$\frac{G v_S^2}{1 - h\big(p(v_S)/2\big)} \le \mathcal{P}(v_S) \le \frac{G v_S^2}{1 - h\big(2p(v_S)\big)}, \tag{7}$$

where G is a constant. The optimum reduction in $\min_{v_S} \mathcal{P}(v_S)$ w.r.t. the worst-case lies between 40% and 49%. ■

The bounds on $\mathcal{P}(v_S)$ are derived using ideas from Information theory [9, Ch. 8] and error-control code theory [11], respectively. We omit the details for brevity.

Fig. 4 illustrates the power per bit bounds as a function of $p(v_S)$. The minimum value of the upper bound and the lower bound are 40% and 49% less than the worst-case, at $v_S = 130mV$ and $v_S = 150mV$, respectively.

3.3. Power reduction with low-latency codes

Practical SRAM design requires low latency of a few clock cycles. We explore power per bit reduction as a function of n for Hamming and Reed Muller codes. We will study the power reduction at an outage of $\epsilon = 0.01$. Rows in outage will be corrected by row-redundancy [10]. The outage condition simplifies to

$$\epsilon = \mathbb{P}\big[DRV_{(n-u)} \ge v_S\big], \tag{8}$$

where the code can correct up to u errors and $DRV_{(t)}$ is the t^{th} largest random DRV. The power per bit function is

$$\mathcal{P}_{0.01}(v_S) = G \cdot \frac{n}{k} \cdot \big(v_S\big)^2. \tag{9}$$

[3] The constant G is random but fixed for a given chip. In a given chip, the randomly realized value of G does not affects the percentage-reduction in $\mathcal{P}_\epsilon(v_S)$ (for large T_s) and hence its variation is ignored.

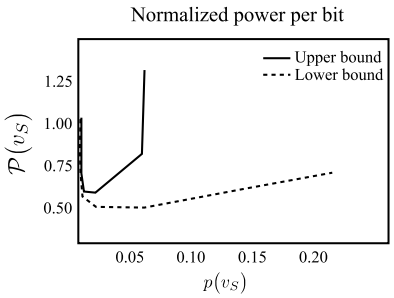

Fig. 4. **Bounds on $\mathcal{P}(v_S)$:** *The minimum values of upper and lower bounds are 40% and 49% lower than the worst-case.*

The trade-off curves are shown in Fig. 5. For Hamming codes, the minimum $\mathcal{P}_{0.01}(v_S)$ is 33% less than the worst-case and is achieved at $n = 31$. The corresponding numbers for Reed Muller code are 33% and 256, respectively. A significant fraction, 33% out of the optimum 40% (see Thm. 3.1), power per bit reduction is achieved with a single clock-cycle latency Hamming code. The gap can be reduced with higher-complexity coding. The returns are marginal, e.g., 2% extra power per bit can be saved by a Reed Muller code with 8-times larger block length.

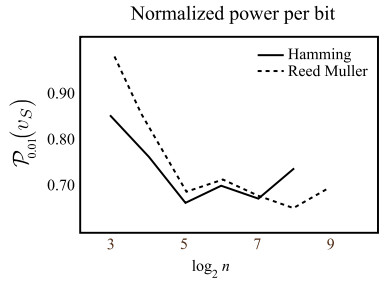

Fig. 5. **$\mathcal{P}(v_S)$ for finite n:** *For an outage $\epsilon = 0.01$, the optimum power reduction for Hamming and Reed Muller codes are plotted.*

3.4. Accounting for coding and latency overheads

We selected the Hamming code with a block length $n = 31$ for implementation. We synthesized the encoder-decoder using CAD tools (90nm CMOS technology). The estimated average encoding and decoding energy for 26-bit word were $0.93pJ$ and $2.32pJ$, respectively. The measured leakage-current at $200mV$ for 256 cells was $55.76nA$. Based on this data, *we estimated that $T_s \geq 100$ms is sufficient to achieve power per bit reduction of 33%.* The latency of the Hamming code is 1-clock cycle (2ns).

Independence of DRV: Correlations in the DRV can be exploited with better coding strategies. However, from the test-chip measurements, we observed a small spatial correlation factor (< 0.1) in the DRV data. Since the measured correlation is small, the resulting gains will not be significant. Therefore, we work with the pessimistic i.i.d. assumption.

4. CONCLUSIONS

We studied the problem of reducing power during data-retention in a standby SRAM. For successful data-retention, the supply voltage of an SRAM cell should be greater than a critical threshold voltage called DRV. For reliable data retention, the existing low-power design technique uses a standby supply voltage that is higher than the worst-case DRV voltage. Instead, we have advocated aggressive voltage reduction with a fault-tolerant memory architecture to optimize standby power. We established fundamental bound on the reduction of standby power. We also studied the dependence of power-reduction on block-length for low-latency codes and showed that most of power-reduction can be achieved by a Hamming code with a block-length 31. We proposed a practical reliable memory architecture based on the Hamming code for the next test-chip implementation with a predicted power reduction of 33% while accounting for the coding and latency overheads.

5. REFERENCES

[1] H. Qin, R. Vattikonda, T. Trinh, Y. Cao, and J. Rabaey, "SRAM cell optimization for ultra-low power standby operation," *Journal of Low Power Electronics*, vol. 2, no. 3, pp. 401–411, Dec 2006.

[2] J. M. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits – A design perspective*, Prentice Hall, NJ, USA, 2nd edition, 2003.

[3] S. Roundy, B. Otis, Y. H. Chee, J. Rabaey, and P. Wright, "A 1.9 GHz RF transmit beacon using environmentally scavenged energy," in *Proc. of Digital IEEE Intl. Symposium, Low Power Electronic Devices*, 2003.

[4] System Drivers, "International Technology Roadmap for Semiconductors," *http://www.itrs.net*, pp. 1–25, 2005.

[5] H. Qin, Y. Cao, D. Markovic, A. Vladimirescue, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *ISQED'04: Proc. of Fifth Intl. Symposium on Quality Electronic Design*, 2004, pp. 55–60.

[6] C. Heegard and A. E. Gamal, "On the capacity of computer memory with defects," *IEEE Trans. on Information Theory*, vol. 29, no. 5, pp. 731–739, Sept 1983.

[7] C. W. Slayman, "Cache and memory error detection, correction, and reduction techniques for terrestrial servers and workstations," *IEEE Trans. on Reliability*, vol. 5, no. 3, pp. 397–404, Sept 2005.

[8] M. Agostinelli et al., "Erratic fluctuations of SRAM cache vmin at the 90nm process technology node," in *IEEE International Electron Devices Meeting, 2005*. IEDM Technical Digest, Dec 2005, pp. 655–658.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, New York, NY, USA, 1991.

[10] W. K. Huang, Y. Shen, and F. Lombardi, "New approaches for the repairs of memories with redundancy by row/column deletion for yield enhancement," *IEEE Trans. on CAD of Integrated Circuits and Systems*, pp. 323–328, Mar. 1990.

[11] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice Hall, NJ, USA, 1st edition, 1995.