Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS

Saibal Mukhopadhyay, *Student Member, IEEE*, Hamid Mahmoodi, *Student Member, IEEE*, and Kaushik Roy, *Fellow, IEEE*

***Abstract*—In this paper, we have analyzed and modeled failure probabilities (access-time failure, read/write failure, and hold failure) of synchronous random-access memory (SRAM) cells due to process-parameter variations. A method to predict the yield of a memory chip based on the cell-failure probability is proposed. A methodology to statistically design the SRAM cell and the memory organization is proposed using the failure-probability and the yield-prediction models. The developed design strategy statistically sizes different transistors of the SRAM cell and optimizes the number of redundant columns to be used in the SRAM array, to minimize the failure probability of a memory chip under area and leakage constraints. The developed method can be used in an early stage of a design cycle to enhance memory yield in nanometer regime.**

***Index Terms*—Leakage, performance, random dopant fluctuation (RDF), robustness, synchronous random-access memory (SRAM), yield.**

I. Introduction

THE random variations in process parameters have emerged as a major design challenge in circuit design in the nanometer regime [1]–[3]. The sources of the inter-die and the intra-die variations in process parameters includes variations in channel length, channel width, oxide thickness, threshold voltage, line-edge roughness, and random dopant fluctuations [the random variations in the number and location of dopant atoms in the channel region of the device resulting in the random variations in transistor threshold voltage (RDF)] [1]–[5]. These different sources of variations result in significant variation in the delay and the leakage of digital circuits [1]–[5]. The inter-die variation in a parameter [say threshold voltage (V_{t})] modifies the value of that parameter of all transistors in a die in the same direction (i.e., threshold voltage of all the transistors either increase or reduce). This principally results in a spread in the delay and the leakage, but does not cause a mismatch between different transistors in a die. On the other hand, the intra-die variations shift the process parameters of different transistors in a die in different directions (e.g., V_{t} of some transistors increase whereas that of some others reduce). The intra-die (or on-die) variations can be systematic (i.e., shift in a parameter of one transistor depends on the shift of that parameter of a neighboring transistor) or random (i.e., shifts in a parameter of two neighboring transistors are completely independent). An example of the systematic intra-die variation can be the change in the channel length of different transistors of a die that are spatially correlated. The RDF induced V_{t} variation is a classic example of the random intra-die variation. The systematic variation does not result in large differences between the two transistors that are in close spatial proximity. The random component of the intra-die variation can result in a significant mismatch between the neighboring transistors in a die [1]–[5].

In a static random-access memory (SRAM) cell, a mismatch in the strength between the neighboring transistors, caused by intra-die variations, can result in the failure of the cell [7]–[9]. For example, a cell failure can occur due to: 1) an increase in the cell access time (access time failure); 2) unstable read (flipping of the cell data while reading) and/or write (inability to successfully write to a cell) operations (read/write failure); or 3) failure in the data holding capability of the cell (flipping of the cell data with the application of a supply voltage lower than the nominal one) at the standby mode (hold failure in the standby mode). Since these failures are caused by the variations in the device parameters, these are known as the parametric failures [8], [9]. There can also be hard failures (caused by open or short) or soft failures due to soft error. In this paper, we will concentrate only on the parametric failures, and hereafter, by the word "failure," we will refer to the parametric failures. A failure in any of the cells in a column of the memory will make that column faulty. In a memory, the redundant columns are used to improve the fault tolerance of the memory and when a column is detected as a faulty one, it gets replaced by an available redundant column. Thus, if the number of faulty columns in a memory chip is larger than the number of available redundant columns, then the chip is considered to be faulty (a similar argument holds for the memory designed with the row redundancy). Hence, the probability of failure of a cell is directly related to the yield of a memory chip. Thus, the intra-die-variation-induced device mismatch can significantly reduce the yield of a memory. As the effect of the intra-die variations increases with the technology

Manuscript received September 14, 2003; revised December 2, 2004. This work was supported in part by the Semiconductor Research Corporation, the Defence Advance Research Project Agency Power Aware Computing and Communication (DARPA PACC) Program, Intel, and IBM Corporation. This paper was recommended by Associate Editor S. Sapatnekar.

The authors are with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: sm@ecn.purdue.edu; mahmoodi@ecn.purdue.edu; kaushik@ecn.purdue.edu).

Digital Object Identifier 10.1109/TCAD.2005.852295

0278-0070/$20.00 © 2005 IEEE

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

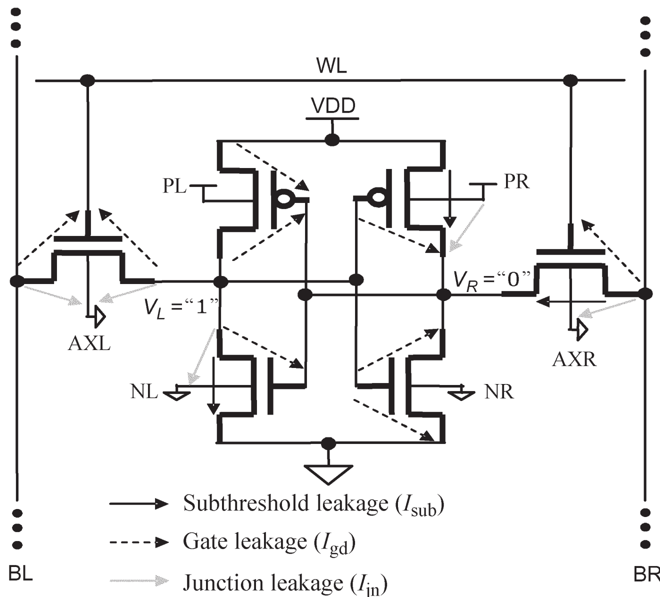

Fig. 1. SRAM cell.

scaling [1]–[5], analysis and reduction of the mismatch-induced parametric failures in an SRAM cell is extremely necessary to enhance the yield of a memory designed in nanoscaled complementary metal–oxide–semiconductor (CMOS) [8], [9]. In this paper, we have analyzed and analytically modeled the different types of parametric failures (mentioned above) that can occur in an SRAM cell.

Among the different sources of random intra-die variations, the most significant one is the threshold-voltage (V_t) variation due to RDF. The impacts of random dopant effect are most pronounced in minimum-geometry transistors commonly used in area-constrained circuits such as SRAM cells [7]. Hence, in this work we have principally concentrated on the on-die random variation in the threshold voltage of the different transistors in a cell due to RDF. However, the analysis is equally applicable to all the other sources of the intra-die variations such as channel length, channel width, etc. Our analysis is primarily focused on the random component of the intra-die variation. However, we also have considered the effect of the correlation among the threshold voltage of the different transistors in a cell to understand the impact of the systematic variations.

The parametric variations, and in particular the V_t fluctuation due to RDF, is a strong function of the size of different transistors in the cell [channel length (L), width (W)]. Hence, the failure probability of a memory can be reduced by optimally designing the size of different transistors. However, any such optimization has to consider its impact on the overall area and the leakage of the SRAM array. Moreover, the memory organization [i.e., number of rows (N_{ROW}) and number of columns (N_{COL}) and the number of redundant columns (N_{RC})] also have a strong impact on the memory-failure probability. Hence, a statistical design of the SRAM cell and memory organization is very important to reduce the memory-failure probability and to improve the yield in nanoscaled SRAM.

In this paper, we have developed a statistical methodology to design the size of the different transistors of an SRAM cell and the memory organization, in order to reduce the memory-failure probability considering on-die V_t variations, which are constrained by the overall memory area and leakage power. In particular, we have:

1) presented semianalytical models for the access time and the read, write, and hold failures of a cell due to the on-die random variation of transistor threshold voltages;
2) analyzed the effect of the correlation of the threshold voltages of different transistors on the failure probabilities;
3) developed a method to estimate the failure probability of a memory (considering the memory architecture and the number of redundant columns) and to predict the yield of a memory chip;
4) presented a statistical analysis of the impact of circuit (transistors sizing) and architecture (N_{ROW}, N_{COL}, and N_{RC}) on the cell- and memory-failure probability;
5) proposed a statistical-design strategy to reduce the memory-failure probability and improve the yield in nanoscaled SRAM.

The statistical-design strategy is developed considering the on-die threshold voltage (V_t) variation, but can be extended to consider on-die L and W variations.

The remainder of this paper is organized as follows. Section II briefly describes the different failure mechanisms in an SRAM cell. The modeling of the different failure probabilities in an SRAM cell are explained in Section III. The models developed in Section III are based on the complex short-channel transistors models that require numerical solutions to estimate the probability values. However, such a numerical solution (although accurate) is computationally very expensive. Hence, in Section IV, we have presented the analytical estimation of the failure probabilities using simple long-channel transistor models. The models using long-channel current models predict the probabilities fast, but they are less accurate than the complete numerical solutions. In Section V, we have analyzed the sensitivity of the failure probabilities to environmental conditions and transistor sizes. Section VI illustrates the statistical-design approach for SRAM. Finally, Section VII concludes the paper.

II. Mechanisms of Failure in an SRAM Cell

On-die variations in the process parameters (e.g., threshold voltage, channel length, channel width, etc., of transistors) result in the mismatch in the strength of the different transistors in an SRAM cell. This device mismatch can result in the failure of the SRAM cell. The parametric failures in an SRAM cell (Fig. 1) are principally due to:

1) destructive read (i.e., flipping of the stored data in a cell while reading—known as read failure);
2) unsuccessful write (inability to write to a cell—defined as write failure);
3) an increase in the access time of the cell resulting in a violation of the delay requirement—defined as access-time failure;
4) the destruction of the cell content in standby mode with the application of a lower supply voltage (primarily to reduce leakage in standby mode)—known as hold failure.

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

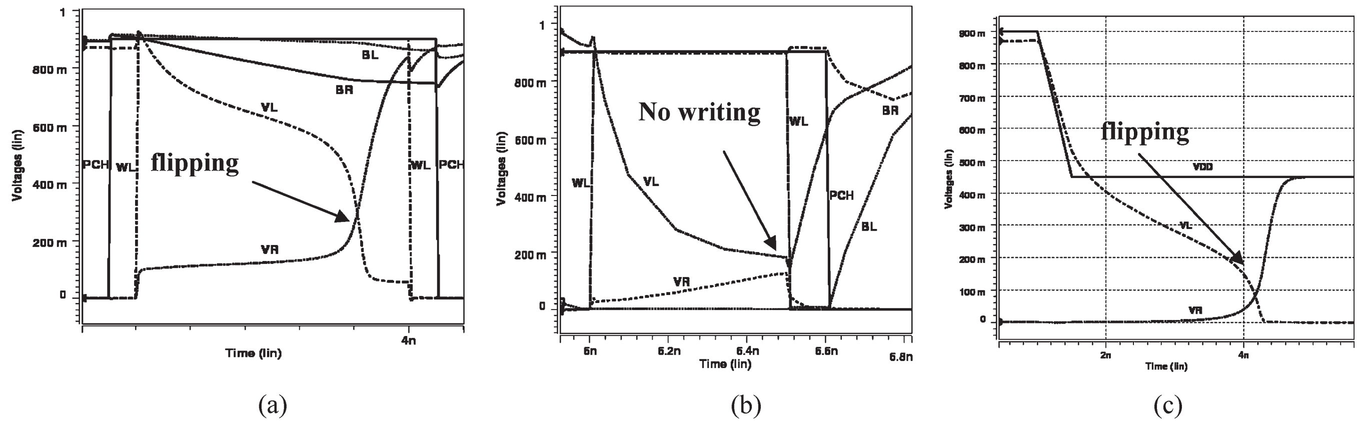

Fig. 2. Unstable read, write, and hold operations: (a) read, (b) write, and (c) hold failure.

In this section, we briefly discuss the mechanisms of each of these failures caused by the mismatch in the threshold voltage (due to random intra-die variations) of the different transistors in an SRAM cell.

A. Read Failure

While reading the cell shown in Fig. 1 ($V_L =$ "1" and $V_R =$ "0"), due to the voltage divider action between AX_R and N_R, the voltage at node R (V_R) increases to a positive value $V_{\rm READ}$. If $V_{\rm READ}$ is higher than the trip point of the inverter $P_L - N_L$ ($V_{\rm TRIPRD}$), then the cell flips while reading the cell [Fig. 2(a)] [11]. This represents a read-failure event. If the strength of the access transistor (AX_R) is higher than that of the pull-down N-type metal oxide semiconductor (NMOS) transistor (N_R), the voltage division action between the two transistors increases the voltage $V_{\rm READ}$. A measure of the relative strength of the AX_R and N_R is the ratio of the "ON" current [known as the beta ratio ($BR_{\rm npd-nax}$)] of these two transistors and is given by

$$BR_{\rm npd-nax} = \frac{\beta_{\rm npd}}{\beta_{\rm nax}} = \frac{\frac{\mu_{\rm eff} C_{\rm ox} W_{\rm npd}}{L_{\rm npd}}}{\frac{\mu_{\rm eff} C_{\rm ox} W_{\rm nax}}{L_{\rm nax}}} \tag{1}$$

where $\mu_{\rm eff}$ is the effective mobility, $C_{\rm ox}$ is the oxide capacitance (assumed to be same as the oxide thickness of both the transistors are same), $W_{\rm nax}$ and $W_{\rm npd}$ are the widths of the access transistor and the pull-down NMOS, respectively, and $L_{\rm nax}$ and $L_{\rm npd}$ are the lengths of the access transistor and the pull-down NMOS, respectively. A decrease in $BR_{\rm npd-nax}$ increases $V_{\rm READ}$, thereby facilitating read failure. Hence, while designing an SRAM cell, the size of the access transistor is usually reduced from that of the pull-down NMOS to increase $BR_{\rm npd-nax}$. However, such a design strategy does not consider the effect of the parameter variation resulting in the random variation in the strengths of different transistors. For example, due to the random variation in the threshold voltage, a reduction in the $V_{\rm t}$ of the access transistor (increase in strength) and an increase in the $V_{\rm t}$ (reduction in strength) of the pull-down NMOS results in an increase in the $V_{\rm READ}$ from its nominal value (i.e., value designed by optimizing the beta ratio), thereby resulting in a read failure. Similarly, the trip point of the inverter $P_L - N_L$ depends on the strengths of the pull-up P-type metal oxide semiconductor (PMOS) and pull-down NMOS. Under nominal condition, the cell is designed to have a weaker PMOS (to facilitate writing, as we will explain in the next section) that results in a lower value of $V_{\rm TRIP}$. Although the nominal value of $V_{\rm TRIP}$ is not less than the nominal value of $V_{\rm READ}$, parameter variation can result in an increase in the $V_{\rm t}$ of P_L and/or reduction in the $V_{\rm t}$ of N_L. This can lower $V_{\rm TRIP}$ below $V_{\rm READ}$, thereby resulting in read failure. It should be noted that the read failure is caused by the mismatch in the strength of the different transistors (e.g., if strength of AX_R increases, while that of N_R reduces). This mismatch can only be caused by the effect of random intra-die variation and not by the inter-die variation (inter-die variation will shift the threshold voltage of all the transistors in the same direction). Hence, an increase in the random intra-die variation can significantly increase the read failure.

B. Write Failure

While writing a "0" to a cell storing "1," the node V_L gets discharged through BL to a low value ($V_{\rm WR}$) determined by the voltage division between the PMOS P_L and the access transistor AX_L [11]. If V_L cannot be reduced below the trip point of the inverter $P_R - N_R$ ($V_{\rm TRIPWR}$) within the time when word-line is high ($T_{\rm WL}$), then a write failure occurs [Fig. 2(b)]. The discharging current (I_L) at node L is the difference in the ON currents of the access transistor AX_L ($I_{\rm AXL}$) and the PMOS P_L ($I_{\rm PL}$) (i.e., $I_L = I_{\rm AXL} - I_{\rm PL}$). Hence, a stronger PMOS and a weaker access transistor can significantly slow down the discharging process, thereby causing a write failure. Thus, while designing the cell, the beta ratio between the access transistor and the PMOS ($BR_{\rm nax-pup} = \beta_{\rm nax}/\beta_{\rm pup}$) needs to be designed (by upsizing the access and downsizing the pull-up transistors) in such a way ($BR_{\rm nax-pup} > 1$) that under nominal conditions, the write time is less than the word-line turn-on time. However, the variation in the device strengths due to random variations in process parameters can increase the write time. For example, if $V_{\rm t}$ of P_L reduces and that of AX_L increases, which can result in an increase in the write-time thereby causing write failure. Hence, a proper static

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

beta-ratio is not sufficient to reduce the write failure. Moreover, upsizing the access transistor and/or downsizing the PMOS transistors increases the read failure. Thus, an optimum design of the size of the different transistors (considering the parameter variation) is necessary to reduce the read and the write failures. It should be noted that write failure is also primarily caused by the mismatch in the strength in the transistors in a cell.

C. Access Time Failure

The cell access time (T_{ACCESS}) is defined as the time required to produce a prespecified voltage difference ($\Delta_{\mathrm{MIN}} \approx 0.1V_{\mathrm{DD}}$) between two bit-lines (bit-differential). If due to V_{t} variation, the access time of the cell is longer than the maximum tolerable limit (T_{MAX}), an access time failure is said to have occurred. Access failure is caused by the reduction in the strength of the access and the pull-down transistors. Thus, access failure limits the reduction in the size of the access transistor (required to increase $BR_{\mathrm{npd-nax}}$ to reduce V_{READ}). An increase in the V_{t} of the access transistor and the pull-down NMOS (caused by the process variation) can significantly increase the access time from its nominal value thereby resulting in an access failure. It should be noted that the access failure is caused by increase in the V_{t} of AX_R and/or V_{t} of N_R. Thus, both intra-die and inter-die variation in process parameters increase the access failure.

D. Hold Failure

In the stand-by mode, the V_{DD} of the cell is reduced to reduce the leakage power consumption. However, if lowering of V_{DD} causes the data stored in the cell to be destroyed, then the cell is said to have failed in the hold mode [15] [Fig. 2(c)]. As the supply voltage of the cell is lowered, the voltage at the node storing "1" (node L in Fig. 1) also gets reduced. Moreover, for a low supply voltage (when P_L is not strongly "ON") leakage of the pull-down NMOS N_L reduces the voltage at node L, even below the supply voltage applied to the cell. If the voltage at the node L is reduced below the trip-point of the inverter $P_R - N_R$, then flipping occurs and the data are lost in the hold mode. The supply voltage to be applied in the hold mode is chosen to ensure the holding of the data under nominal condition. However, variation in the process parameter can result in the device mismatch causing hold failures. For example, if the V_{t} of N_L reduces while that of P_L increases (which facilitates the reduction of the voltage at node L from the applied supply voltage) and/or V_{t} of N_R increases, while that of P_R reduces (increase in the trip-point of $P_R - N_R$) the possibility of data flipping in the hold mode increases. Consequently, an increase in the random intra-die variation can significantly increase the hold-failure.

III. Modeling of Failure Probabilities

In an SRAM cell, mismatches in the device parameters (L, W, V_{t}) of different transistors (cause by intra-die variations) result in different types of failures as explained in Section II. Because of the small geometry of the SRAM cell, the principal source of the device mismatch is the intrinsic fluctuation of the V_{t} of different transistors due to RDF [4]. As the transistors in a cell are in very close spatial proximity, the effect of mismatch in the channel length or width is small. Hence, in this work we have considered the V_{t} variation due to RDF as the major source of intra-die variation, while estimating the probabilities of different failure events. However, the proposed method can also be extended to include other sources of variations (such as L and W variations or the impact of L and W variations can also be represented as an additional contribution to the V_{t} variation). In this section, we will explain the basic methodology used to model the different failure probabilities.

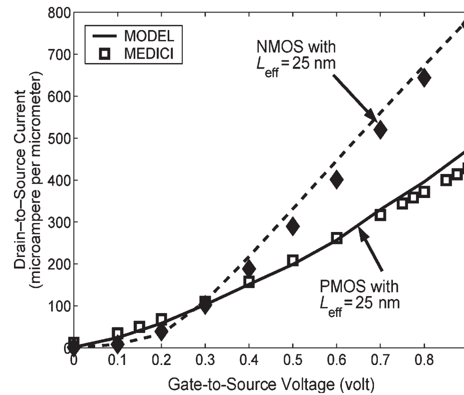

Fig. 3. Device characteristics. Discrete points represent the result obtained in MEDICI simulation and the lines represent the results from the models.

The failure probabilities are estimated using an SRAM cell designed with bulk CMOS transistors of 50-nm gate length ($L_{\mathrm{eff}} = 25$ nm). The transistors are designed using the two-dimensional Gaussian doping profiles [11] and simulated using the device simulator MEDICI [12]). In our analysis, we have used the short-channel metal oxide semiconductor field effect transistor (MOSFET) theory to model the current and the threshold voltage considering the device geometry and doping profile [3], [5]. Fig. 3 shows the $I_{\mathrm{d}} - V_{\mathrm{g}}$ characteristics of the designed transistors. The leakage current models presented in [10] are used to represent different leakage components.

While modeling the failure probabilities considering the threshold voltage variations due to RDF, V_{t} fluctuations (δV_{t}) in the six transistors in an SRAM cell are considered as six independent Gaussian random variables (mean $= 0$) [4]. The assumption of the independent random variable is justified as we have considered primarily the effect of RDF. The placement and the number of dopants in the channel of one transistor depend only on the geometry of that transistor and are independent of the placements and the number of dopants in the channel of a neighboring transistor [4]. Thus, the V_{t} fluctuation due to the RDF of one transistor does not depend on the V_{t} fluctuation of any neighboring transistor. Hence, the δV_{t} of the cell transistors can be assumed as independent random variables [4]. However, we have also investigated the effect of the correlation of δV_{t}s of the cell transistors on the failure probabilities. The standard deviation of the V_{t} fluctuation ($\sigma_{V_{\mathrm{t}}}$) due to RDF depends on the manufacturing process, doping profile, and the transistor sizing [6]. In the proposed method, $\sigma_{V_{\mathrm{t}}}$ for a

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

minimum-sized transistor (σ_{V_t0}) is an input parameter and the dependence of σ_{V_t} on the transistor size is given by [6]

$$\sigma_{V_t} = \sigma_{V_t0}\sqrt{\left(\frac{L_{\min}}{L}\right)\left(\frac{W_{\min}}{W}\right)}. \quad (2)$$

A. Modeling Methodology

In this section, we will summarize the key mathematical bases used to estimate the failure probabilities. Let us consider $y = f(x_1, \ldots, x_n)$ as a function, where $x_1, \ldots, x_n$ are independent Gaussian random variables with mean $\eta_1, \ldots, \eta_2$ and standard deviation (STD) $\sigma_1, \ldots, \sigma_2$. The mean (μ_y) and the STD (σ_y) of the random variable y can be estimated as (using multivariable Taylor-series expansion) [13]

$$\mu_y = f(\eta_1, \ldots, \eta_n) + \frac{1}{2}\sum_{i=1}^{n} \left.\frac{\partial^2 f(x_1, \ldots, x_n)}{\partial (x_i)^2}\right|_{\eta_i} \sigma_i^2$$

$$\sigma_y^2 = \sum_{i=1}^{n}\left(\left.\frac{\partial f(x_1, \ldots, x_n)}{\partial (x_i)}\right|_{\eta_i}\right)^2 \sigma_i^2. \quad (3)$$

Assuming the probability distribution function (PDF) of y to be also Gaussian $[N_y(y : \mu_y, \sigma_y)]$, the probability of $(y > Y_0)$ is given by

$$P[y > Y_0] = \int_{y=Y_0}^{\infty} N_y(y : \mu_y, \sigma_y)\mathrm{d}y = 1 - \int_{y=-\infty}^{Y_0} N_y(y)\mathrm{d}y = 1 - \Phi_y(Y_0) \quad (4)$$

where Φ_y is the cumulative distribution function (CDF) of y.

Let us assume $y = f(x_1, \ldots, x_n)$ and $z = g(x_1, \ldots, x_n)$ are two Gaussian random variables $N_y(y : \mu_y, \sigma_y)$ and $N_z(y : \mu_z, \sigma_z)$, respectively. The probability of $(y > Y_0$ and $z > Z_0)$ is given by

$$\begin{aligned} &P[(y > Y_0) \text{ and } (z > Z_0)] \\ &\quad = 1 - P[(y \le Y_0) + (z \le Z_0)] \\ &\quad = 1 - \{P[y \le Y_0] + P[z \le Z_0] \\ &\qquad\quad - P[(y \le Y_0) \,\&\, (z \le Z_0)]\} \\ &\quad = \{P[y > Y_0] + P[z > Z_0] - 1\} + \Phi_{y,z}(Y_0, Z_0) \end{aligned} \quad (5)$$

where $\Phi_{y,z}(y, z)$ is the joint CDF of y and z. $\Phi_{y,z}(Y_0, Z_0)$ is given by

$$\Phi(Y_0, Z_0) = \int_{y=-\infty}^{Y_0}\int_{z=-\infty}^{Z_0} N_{y,z}(y : \mu_y, \sigma_y; z : \mu_z, \sigma_z)\,\mathrm{d}y\mathrm{d}z. \quad (6)$$

The joint PDF $N_{y,z}(y : \mu_y, \sigma_y; z : \mu_z, \sigma_z)$ is given by

$$N_{y,z}(y, z) = \frac{1}{2\pi\sigma_y\sigma_z\sqrt{1-\rho^2}} \times \exp\left[-\frac{\left(\frac{y-\mu_z}{\sigma_z}\right)^2 - 2\rho\left(\frac{y-\mu_z}{\sigma_z}\right)\left(\frac{z-\mu_z}{\sigma_z}\right) - \left(\frac{z-\mu_z}{\sigma_z}\right)^2}{2(1-\rho^2)}\right]. \quad (7)$$

The correlation coefficient ρ can be computed as follows:

$$\rho = \frac{E(yz) - E(y)E(z)}{\sigma(y)\sigma(z)} = \frac{E(yz) - \mu_y\mu_z}{\sigma_y\sigma_z}$$

$$\begin{aligned} E(yz) &= f(\eta_1, \ldots, \eta_n)g(\eta_1, \ldots, \eta_n) + \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2(fg)}{\partial(x_i)^2}\sigma_i^2 \\ &= f_0g_0 + \frac{1}{2}\sum_{i=1}^{n}\left(g_0\frac{\partial^2 f}{\partial x_i^2} + 2\frac{\partial f}{\partial x_i}\frac{\partial g}{\partial x_i} + f_0\frac{\partial^2 g}{\partial x_i^2}\right)\sigma_i^2. \end{aligned} \quad (8)$$

The above results will be used in this paper to estimate the failure probabilities of different events.

B. Read Failure (R_{F})

As explained in Section II-A, read failure occurs when during reading voltage at node R (V_{READ}) increases to a value higher than the trip-point of the inverter $P_L - N_L$ (V_{TRIPRD}) [Fig. 3(a)] [14]. Hence, the read-failure probability (P_{RF}) is given by

$$P_{\mathrm{RF}} = P[V_{\mathrm{READ}} > V_{\mathrm{TRIPRD}}]. \quad (9)$$

V_{READ} can be obtained by simultaneously solving Kirchhoff's Current Law (KCL) at node R and L, as given by

$$\begin{aligned} \text{At } R \equiv\, & I_{\mathrm{dsatAXR}} + I_{\mathrm{gsAXR}} + I_{\mathrm{subPR}} + I_{\mathrm{gdPR}} \\ & + I_{\mathrm{jnPR}} + I_{\mathrm{gdNR}} + I_{\mathrm{gdNL}} + I_{\mathrm{gdPL}} + I_{\mathrm{gsPL}} \\ =\, & I_{\mathrm{dlinNR}} + I_{\mathrm{jnNR}} + I_{\mathrm{jnAXR}}, \\ \text{At } L \equiv\, & I_{\mathrm{dsNL}} + I_{\mathrm{jnNL}} + I_{\mathrm{gdNL}} + I_{\mathrm{gdPL}} \\ =\, & I_{\mathrm{dsPL}} + I_{\mathrm{jnPL}} + I_{\mathrm{dlinAXL}} \end{aligned} \quad (10)$$

where I_{dsatXX} is the saturation current (from drain to source), I_{dlinXX} is the drain current at the linear region of operation, I_{gsXX} represents the gate-to-source component of the gate leakage, I_{gdXX} represents the gate-to-drain component of the gate leakage, I_{subXX} represents the subthreshold leakage, and I_{jnXX} represents the junction leakage of the transistor XX [where XX represents different cell transistors used in (10)]. Similarly, V_{TRIPRD} can be obtained by solving [14]

$$\begin{aligned} & I_{\mathrm{dsatNL}}(V_{\mathrm{gate}} = V_{\mathrm{TRIPRD}}, V_{\mathrm{drain}} = V_{\mathrm{TRIPRD}}, V_{\mathrm{source}} = \mathrm{gnd}) \\ & \approx I_{\mathrm{dsatPL}}(V_{\mathrm{gate}} = V_{\mathrm{TRIPRD}}, V_{\mathrm{drain}} = V_{\mathrm{TRIPRD}}, V_{\mathrm{source}} = V_{\mathrm{DD}}) \\ & \quad + I_{\mathrm{dsatAXL}}(V_{\mathrm{gate}} = V_{\mathrm{DD}}, V_{\mathrm{drain}} = V_{\mathrm{DD}}, V_{\mathrm{source}} = V_{\mathrm{TRIPRD}}). \end{aligned} \quad (11)$$

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

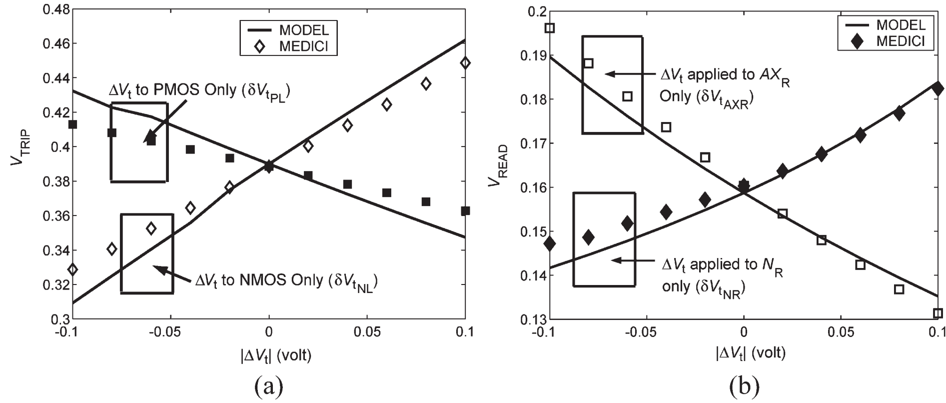

Fig. 4. Variation of (a) V_{TRIP} of $P_L - N_L$ and (b) V_{READ}, with δV_{t} applied to different transistors.

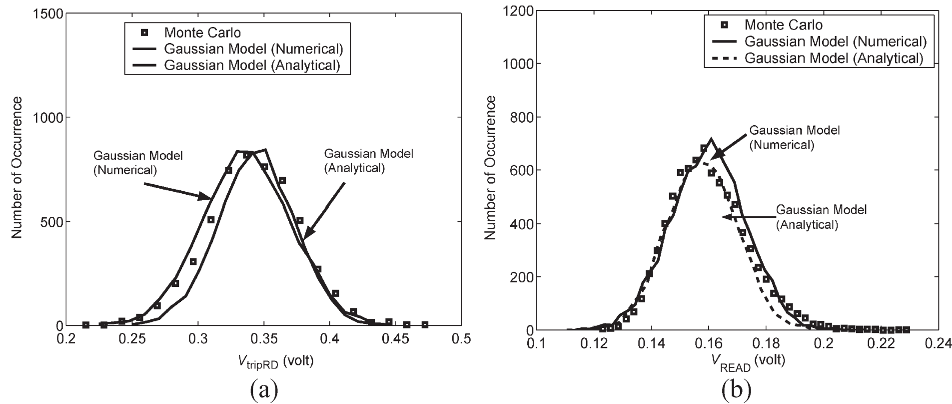

Fig. 5. Distributions of (a) V_{TRIP} of $P_L - N_L$ and (b) V_{READ}. The curves entitled Gaussian Model (Analytical) represent solutions explained in Section IV.

TABLE I
FAILURE-PROBABILITY ESTIMATIONS FOR DIFFERENT CELLS (MONTE CARLO ESTIMATION)

σ_{VT0}=80 mV (A Large Value of σ_{VT0} is used to verify the results with Monte Carlo simulations of 10000 Vector)
V_{DD}=0.9 V, V_{HOLD}=0.45 V, T_{MAX}=75 ps, T_{WL}=90 ps)
Cells: C1≡($L_{\mathrm{p}}=L_{\mathrm{npd}}=L_{\mathrm{nax}}$=50 nm, W_{p}=100 nm, W_{npd}=200 nm, W_{nax}=150 nm);
C2≡($L_{\mathrm{p}}=L_{\mathrm{npd}}=L_{\mathrm{nax}}$=50 nm, W_{p}=150 nm, W_{npd}=200 nm, W_{nax}=150 nm),

	P_{RF}	P_{AF}	P_{WF}	P_{HF}	$P[A_{\mathrm{F}}R_{\mathrm{F}}]$	$P[R_{\mathrm{F}}W_{\mathrm{F}}]$	$P[R_{\mathrm{F}}H_{\mathrm{F}}]$	$P[A_{\mathrm{F}}W_{\mathrm{F}}]$	$P[A_{\mathrm{F}}H_{\mathrm{F}}]$	$P[W_{\mathrm{F}}H_{\mathrm{F}}]$	$P[A_{\mathrm{F}}R_{\mathrm{F}}H_{\mathrm{F}}]$	$P[A_{\mathrm{F}}R_{\mathrm{F}}H_{\mathrm{F}}]$, $P[R_{\mathrm{F}}W_{\mathrm{F}}H_{\mathrm{F}}]$, $P[W_{\mathrm{F}}H_{\mathrm{F}}A_{\mathrm{F}}]$	P[All]	P_{F}
C1	0.041/ 0.038	0.159/ 0.152	0.005/ 0.005	0.067/ 0.062	0.008/ 0.009	0/0	0.029/ 0.027	0/0.001	0.014/ 0.0155	0/0	0.002/0	0/0	0/0	0.223/ 0.21
C2	0.009/ 0.008	0.150/ 0.152	0.028/ 0.022	0.029/ 0.031	0/0.002	0/0	0.008/ 0.006	0.007/ 0.005	0.006/ 0.008	0/0	0/0	0/0	0/0	0.221/ 0.192

Fig. 4 shows that V_{TRIPRD} [obtained using (11) and the MEDICI simulation] is a linear function of independent random variables: $\delta V_{\mathrm{t_{NL}}}$ and $\delta V_{\mathrm{t_{PL}}}$. Similarly, V_{READ} [obtained using (10) and the MEDICI simulation] is a linear function of independent random variables: $\delta V_{\mathrm{t_{AXR}}}$ and $\delta V_{\mathrm{t_{NR}}}$. As explained in Section II-A, V_{READ} increases if the strength of AX_R increases ($V_{\mathrm{t_{AXR}}} \downarrow => I_{\mathrm{dsatAXR}} \uparrow$) and/or that of N_R reduces ($V_{\mathrm{t_{NR}}} \uparrow => I_{\mathrm{dlinNR}} \downarrow$). On the other hand, V_{TRIPRD} reduces when the strength of P_L reduces ($V_{\mathrm{t_{PL}}} \uparrow$) and/or that of N_L increases ($V_{\mathrm{t_{NL}}} \downarrow$). The PDF of V_{READ} [$= N_{\mathrm{RD}}(v_{\mathrm{READ}})$] and V_{TRIP} [$= N_{\mathrm{TRIP}}(v_{\mathrm{TRIP}})$] can be approximated as Gaussian distributions with the means and the variances obtained using (3) [Fig. 5(a) and (b)]. P_{RF} is given by

$$P_{\mathrm{RF}} = P\left[Z_R \equiv (V_{\mathrm{READ}} - V_{\mathrm{TRIPRD}}) > 0\right] = 1 - \Phi_{\mathrm{ZR}}(0) \tag{12}$$

where $\eta_{\mathrm{ZR}} = \eta_{V_{\mathrm{READ}}} - \eta_{V_{\mathrm{TRIP}}}$ and $\sigma^2_{\mathrm{ZR}} = \sigma^2_{V_{\mathrm{READ}}} - \sigma^2_{V_{\mathrm{TRIPD}}}$.

The estimated value of P_{RF} closely follows the values obtained from Monte Carlo simulations (Table I).

C. Write Failure (W_{F})

Following the discussion in Section II-B, the write failure occurs when, while writing a "0" to the node storing "1" (node L in Fig. 1), the voltage at node L (V_L) is not reduced below the trip-point of the inverter $P_R - N_R$ (V_{TRIPWR}) within the time when word-line is high (T_{WL}). The write-failure probability (P_{WF}) is given by

$$P_{\mathrm{WF}} = P\left[(T_{\mathrm{WRITE}} > T_{\mathrm{WL}})\right] \tag{13}$$

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

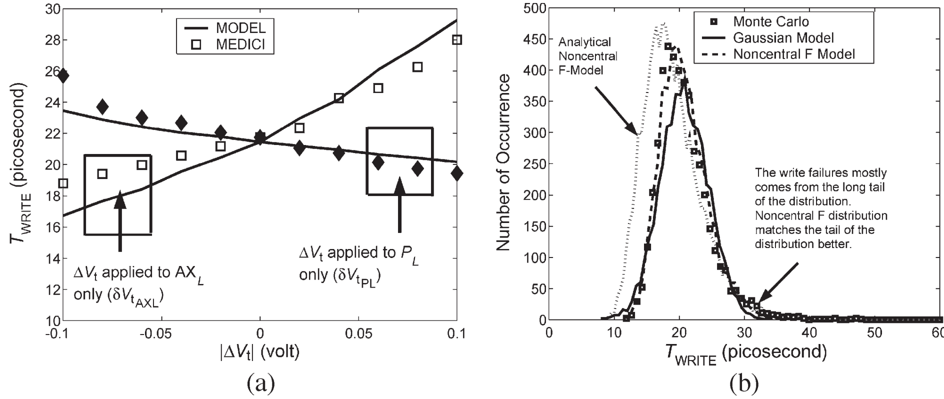

Fig. 6. Variation and distribution of T_{WRITE} with variation in δV_{t}. (a) T_{WRITE} variation with δV_{t}, and (b) distribution of T_{WRITE}. The curves entitled Gaussian Model (Analytical) represent solutions explained in Section IV.

where T_{WRITE} is the time required to pull down V_L from V_{DD} to V_{TRIPWR}. T_{WRITE} is obtained by solving

$$T_{\mathrm{WRITE}} = \begin{cases} \left| \int_{V_{\mathrm{DD}}}^{V_{\mathrm{TRIP}}} \frac{C_L(V_L)\mathrm{d}V_L}{I_{\mathrm{in}(L)}(V_L) - I_{\mathrm{out}(L)}(V_L)} \right|, & \text{if } (V_{\mathrm{WR}} < V_{\mathrm{TRIPWR}}) \\ \infty, & \text{if } (V_{\mathrm{WR}} \geq V_{\mathrm{TRIPWR}}) \end{cases}$$
$$I_{\mathrm{in}(L)} = \text{current into L} \approx I_{\mathrm{dsPL}}, I_{\mathrm{out}(L)} = \text{current out of L} \approx I_{\mathrm{dsAXL}} \quad (14)$$

where C_L is the net capacitance at the node L. V_{WR} can be obtained by simultaneously solving KCL at node L and R. V_{TRIPWR} can be obtained by solving for the trip-point of the inverter $P_R - N_R$ using (11). T_{WRITE} obtained using (14) closely matches the MEDICI simulation result with V_{t} variation of different transistors [Fig. 6(a)]. It can be observed that T_{WRITE} increases if the strength of AX_L reduces ($V_{\mathrm{t_{AXL}}} \uparrow => I_{\mathrm{dsAXL}} \downarrow$) and/or that of P_L increases ($V_{\mathrm{t_{PL}}} \downarrow => I_{\mathrm{dsPL}} \uparrow$). Moreover, V_{TRIP} (P_R, N_R) reduces (thereby increasing T_{WRITE}) when the strength of P_R reduces ($V_{\mathrm{t_{PL}}} \uparrow$) and/or that of N_R increases ($V_{\mathrm{t_{NR}}} \downarrow$). Hence, T_{WRITE} is a strong function of the random variables: $\delta V_{\mathrm{t_{PL}}}$, $\delta V_{\mathrm{t_{NR}}}$, $\delta V_{\mathrm{t_{PR}}}$, and $\delta V_{\mathrm{t_{AXR}}}$. Using (3), we can estimate the mean (η_{TWR}) and the standard deviation (σ_{TWR}) and approximate its PDF as a Gaussian one $(f_{\mathrm{WR}}(t_{\mathrm{WR}}))$ [Fig. 6(b)]. However, most of the write-failures originate from the "tail" of the distribution function. Hence, to improve the accuracy of the model at the tail region, we can use a noncentral F distribution [13]. Using the PDF (Gaussian/noncentral F) of T_{WRITE} $[N_{\mathrm{WR}}(t_{\mathrm{WR}})]$, P_{WF} is given by

$$P_{\mathrm{WF}} = \int_{t_{\mathrm{WR}}=T_{\mathrm{WL}}}^{\infty} N_{\mathrm{WR}}(t_{\mathrm{WR}}) d(t_{\mathrm{WR}}) = 1 - \Phi_{\mathrm{WR}}(T_{\mathrm{WL}}). \quad (15)$$

$\Phi_{\mathrm{WR}}(t_{\mathrm{WR}})$ represents the CDF of the probability distribution (Gaussian/noncentral F) [13]. P_{WF} obtained using (15) closely matches the result using Monte Carlo simulations (Table I).

D. *Access-Time Failure* (A_{F})

Access-time failure occurs if the access time of the cell (T_{ACCESS}) is longer than the maximum tolerable limit (T_{MAX}) (Section II-C). The probability of access-time failure (P_{AF}) of a cell is given by

$$P_{\mathrm{AF}} = P(T_{\mathrm{ACCESS}} > T_{\mathrm{MAX}}). \quad (16)$$

While reading the cell storing V_L = "1" and V_R = "0" (Figs. 1, 3), bit-line BR will discharge through AX_R and N_R (by the current I_{BR}). Simultaneously, BL will discharge by the gate leakage, the subthreshold leakage, and the junction leakage of AX_L of all the cells connected to BL (I_{BL}). The discharging currents I_{BR} and I_{BL} are given by

$$I_{\mathrm{BR}} = I_{\mathrm{dsatAXR}} + \sum_{i=1,\ldots,N} \left[I_{\mathrm{gdAXR}(i)} + I_{\mathrm{jnAXR}(i)} \right] \quad (17a)$$

$$I_{\mathrm{BL}} = \sum_{i=1,\ldots,N} \left[I_{\mathrm{gdAXL}(i)} + I_{\mathrm{jnAXL}(i)} + I_{\mathrm{subAXL}(i)} \right] \quad (17b)$$

where N is the number of cells attached to a bit-line (or column). Hence, T_{ACCESS} can be obtained by solving

$$T_{\mathrm{ACCESS}} = \int_{V_{\mathrm{DD}}}^{V_{\mathrm{DD}} - \Delta V_{\mathrm{BL}} - \Delta_{\mathrm{MIN}}} \frac{C_{\mathrm{BR}} \mathrm{d}V_{\mathrm{BR}}}{I_{\mathrm{BR}}} = \int_{V_{\mathrm{DD}}}^{V_{\mathrm{DD}} - \Delta V_{\mathrm{BL}}} \frac{C_{\mathrm{BL}} \mathrm{d}V_{\mathrm{BL}}}{I_{\mathrm{BL}}} \quad (18)$$

where $C_{\mathrm{BR/BL}}$ is the bit-line capacitance that includes the diffusion capacitance (C_{jn}) of the access transistors and the interconnect capacitances (C_{IC}). To simplify the above

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

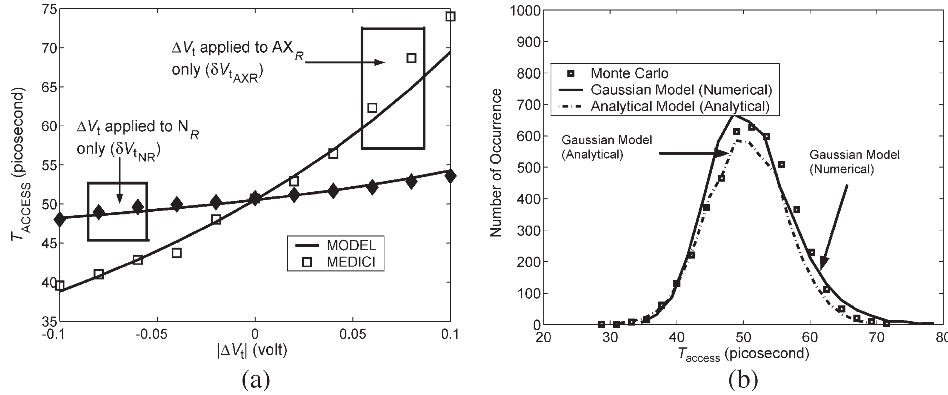

Fig. 7. Variation and distribution of T_{ACCESS} with variation in δV_{t}: (a) T_{ACCESS} variation with δV_{t}, and (b) distribution of T_{ACCESS}. The curves entitled Gaussian Model (Analytical) represent solutions explained in Section IV.

calculation, we assume that I_{dsatAXR} is constant (valid for small ΔV_{BR} since AX_R is in saturation) and I_{gd}, I_{jn}, and I_{sub} are constant at their values for $V_{\text{BL}} = V_{\text{DD}}$ (valid for small ΔV_{BR} and ΔV_{BL}). Hence, ΔV_{BR} and ΔV_{BL} are linear functions of time [Fig. 3(a)]. We further assume that $C_{\text{BL}} = C_{\text{BR}} = C_B$, $I_{\text{gdAXR}(i)} = I_{\text{gdAXL}(i)}$ and $I_{\text{jnAXR}(i)} = I_{\text{jnAXL}(i)}$ (since they are not a strong function of V_{t}). Using these assumptions, T_{ACCESS} is given by

$$T_{\text{ACCESS}} = \frac{C_{\text{BR}}C_{\text{BL}}\Delta_{\text{MIN}}}{C_{\text{BL}}I_{\text{BR}} - C_{\text{BR}}I_{\text{BL}}} = \frac{C_B\Delta_{\text{MIN}}}{I_{\text{dsatAXR}} - \sum_{i=1,\ldots,N} I_{\text{subAXL}(i)}}. \quad (19)$$

Fig. 3(a) shows that during a nondestructive read operation, the voltage at node R quickly rises to V_{READ} and stays stable at that value. Hence, to simplify (19), we first solve (10) for V_{READ} and use that V_{READ} to evaluate T_{ACCESS}. The access time given by (19) closely follows the MEDICI simulation result [Fig. 7(a)]. From Fig. 7, it can be observed that T_{ACCESS} increases with an increase in $V_{\text{t}_{\text{AXR}}}$ and/or $V_{\text{t}_{\text{NR}}}$. Hence, T_{ACCESS} principally depends on V_{t} of AX_R and N_R that determines I_{dsatAXR}. The total subthreshold leakage of the cells associated with BL (i.e., $\Sigma I_{\text{subAXL}(i)}$) is approximated as $N \times E[I_{\text{subAXL}}]$, where $E[I_{\text{subAXL}}]$ is the expected value of I_{subAXL} considering random variation in $\delta V_{\text{t}_{\text{AXL}}}$. The PDF of T_{ACCESS} can be approximated as a Gaussian one with the mean (η_{TAC}) and the standard deviation (σ_{TAC}) obtained from (3). Using the derived PDF $[N_{\text{TACCESS}}(t_{\text{ACCESS}})]$, P_{AF} can be estimated as

$$P_{\text{AF}} = \int_{t_{\text{ACCESS}}=T_{\text{MAX}}}^{\infty} N_{\text{TACCESS}}(t_{\text{ACCESS}})\mathrm{d}(t_{\text{ACCESS}}) = 1 - \Phi_{\text{TACCESS}}(T_{\text{vMAX}}) \quad (20)$$

where $\Phi_{\text{TACCESS}}(t_{\text{ACCESS}})$ is the CDF of T_{ACCESS}. P_{AF} of a cell using the model closely matches the one obtained using Monte Carlo simulation (Table I).

E. Hold Failure (H_{F})

A hold failure occurs if the minimum supply voltage that can be applied to the cell in the hold mode (V_{DDHmin}), without destroying the data, is higher than the designed stand-by mode supply voltage (V_{HOLD}) (Section II-D). Thus, the probability of hold failure (P_{HF}) is given by

$$P_{\text{HF}} = P[V_{\text{DDHmin}} > V_{\text{HOLD}}]. \quad (21)$$

Lowering the V_{DD} of the cell (say V_{DDH} represents the cell V_{DD} at the hold mode) reduces the voltage at the node storing "1" (V_L in Fig. 1). Due to leakage of N_L, V_L will be less than V_{DDH} for low V_{DDH}. The hold failure occurs if $V_L < V_{\text{TRIP}}$ of $P_R - N_R$. Hence, V_{DDHmin} can be obtained by numerically solving

$$V_L(V_{\text{DDHmin}}, \delta V_{\text{t}_{\text{PL}}}, \delta V_{\text{t}_{\text{NL}}}) = V_{\text{TRIP}}(V_{\text{DDHmin}}, \delta V_{\text{t}_{\text{PR}}}, \delta V_{\text{t}_{\text{NR}}}). \quad (22)$$

The estimated value of V_{DDHmin}, using (22), closely follows the values obtained from MEDICI simulation [Fig. 8(a)]. From Fig. 8(a), it can be observed that V_L reduces as the strength of N_L increases ($V_{\text{t}_{\text{NL}}} \downarrow$) or that of P_L reduces ($V_{\text{t}_{\text{PL}}} \uparrow$). On the other hand, V_{TRIP} (P_R, N_R) increases as $V_{\text{t}_{\text{NR}}} \uparrow$ or $V_{\text{t}_{\text{PR}}} \downarrow$. From (17), it is evident that V_{DDHmin} is a function of random variables: $\delta V_{\text{t}_{\text{PL}}}$, $\delta V_{\text{t}_{\text{NL}}}$, $\delta V_{\text{t}_{\text{PR}}}$ and $\delta V_{\text{t}_{\text{NR}}}$. The distribution of V_{DDHmin} $[N_{\text{VDDHmin}}(v_{\text{DDHmin}})]$ can be approximated as a Gaussian one with mean and variance obtained using (7) (a noncentral χ^2 distribution improves the accuracy for V_{DDHmin} values close to 0) [Fig. 8(b)]. Hence, we can estimate P_{HF} as

$$P_{\text{HF}} = \int_{V_{\text{HOLD}}}^{\infty} f_{\text{VDDHmin}}(v_{\text{DDHmin}})\mathrm{d}(v_{\text{DDHmin}}) = 1 - F_{\text{VDDHmin}}(V_{\text{HOLD}}). \quad (23)$$

The P_{HF} obtained using (23) closely matches the result using Monte Carlo simulations (Table I).

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

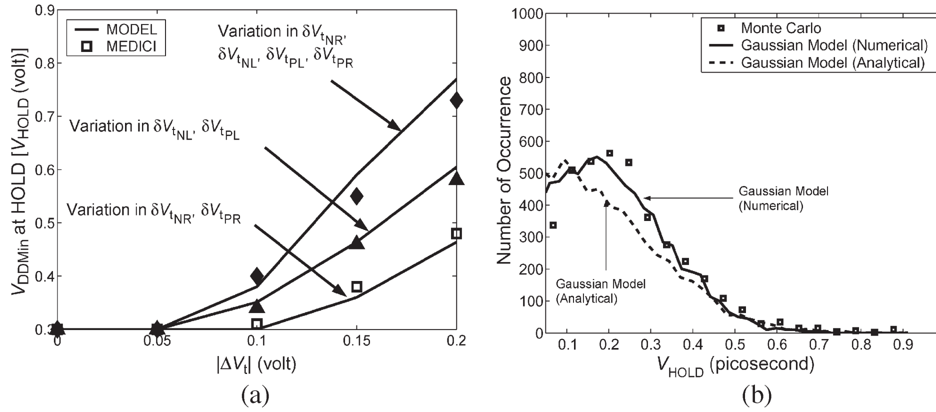

Fig. 8. Variation and distribution of V_{HOLD}. (a) V_{DDHmin} variation with δV_{t}, and (b) distribution of V_{DDHmin}. In (a) δV_{t} is applied in the directions: $\delta V_{\mathrm{tNR}} > 0$, $\delta V_{\mathrm{tPR}} < 0$, $\delta V_{\mathrm{tNL}} < 0$, and $\delta V_{\mathrm{tPL}} > 0$. The curves entitled Gaussian Model (Analytical) represent solutions explained in Section IV.

TABLE II
ESTIMATION OF PROBABILITIES OF JOINT EVENTS

	δV_{tAXR}	δV_{tAXL}	δV_{tNR}	δV_{tNL}	δV_{tPR}	δV_{tPL}
R_{F}	$\delta V_{\mathrm{t}} < 0$	NME*	$\delta V_{\mathrm{t}} > 0$	$\delta V_{\mathrm{t}} < 0$	NME	$\delta V_{\mathrm{t}} > 0$
A_{F}	$\delta V_{\mathrm{t}} > 0$	NME	$\delta V_{\mathrm{t}} > 0$	NME	NME	NME
W_{F}	NME	$\delta V_{\mathrm{t}} > 0$	$\delta V_{\mathrm{t}} < 0$	NME	$\delta V_{\mathrm{t}} > 0$	$\delta V_{\mathrm{t}} < 0$
H_{F}	NME	NME	$\delta V_{\mathrm{t}} > 0$	$\delta V_{\mathrm{t}} < 0$	$\delta V_{\mathrm{t}} < 0$	$\delta V_{\mathrm{t}} > 0$

*NME ≡ does not have a major effect

F. Estimation of Overall Cell Failure Probability (P_{F})

The overall failure probability is given by

$$\begin{aligned} P_{\mathrm{F}} &= P[\mathrm{Fail}] = P[A_{\mathrm{F}} + R_{\mathrm{F}} + W_{\mathrm{F}} + H_{\mathrm{F}}] \\ &= P_{\mathrm{AF}} + P_{\mathrm{RF}} + P_{\mathrm{WF}} + P_{\mathrm{HF}} - P[A_{\mathrm{F}}R_{\mathrm{F}}] - P[A_{\mathrm{F}}W_{\mathrm{F}}] \\ &\quad - P[A_{\mathrm{F}}H_{\mathrm{F}}] - P[R_{\mathrm{F}}W_{\mathrm{F}}] - P[R_{\mathrm{F}}H_{\mathrm{F}}] - P[W_{\mathrm{F}}H_{\mathrm{F}}] \\ &\quad + P[A_{\mathrm{F}}R_{\mathrm{F}}W_{\mathrm{F}}] + P[A_{\mathrm{F}}R_{\mathrm{F}}H_{\mathrm{F}}] + P[R_{\mathrm{F}}W_{\mathrm{F}}H_{\mathrm{F}}] \\ &\quad + P[W_{\mathrm{F}}H_{\mathrm{F}}A_{\mathrm{F}}] - P[\mathrm{All}]. \end{aligned} \tag{24}$$

Table II shows the direction in which δV_{t} of different transistor has to move ($\delta V_{\mathrm{t}} > 0$ or $\delta V_{\mathrm{t}} < 0$) to cause each type of failure. For example, let us consider the joint event $(A_{\mathrm{F}}R_{\mathrm{F}})$. It can be observed that among the four different ways of causing read failure only $\delta V_{\mathrm{t_{NR}}} > 0$ also causes the access-time failure. An accurate estimate of the probability of joint events is possible by constructing the joint PDF representing the two events using the procedure given in (5). We have also assumed that probabilities of simultaneous occurrence of more than two events are negligible (≈ 0). The estimated probabilities match the Monte Carlo results very closely (Table I). All of the different failure probabilities increase significantly with an increase in the sigma of V_{t} variation, as shown in Fig. 9 (for cell C1 in Table I).

G. Estimation of Column and Memory-Failure Probability (P_{COL} *and* P_{MEM})

The failure probability of a column (P_{COL}) or row (P_{ROW}) is defined as the probability that any of the cells in that column (out of N_{ROW} cells) or row (out of N_{COL} cells) fails. Assuming column redundancy, the probability of failure of a memory chip (P_{MEM}) designed with N_{COL} number of columns and N_{RC} number of redundant columns, is defined as the probability that more than N_{RC} (i.e., at least $N_{\mathrm{RC}}+1$) columns fail (similar definition is applicable for row redundancy). Hence, P_{COL} and P_{MEM} can be given by

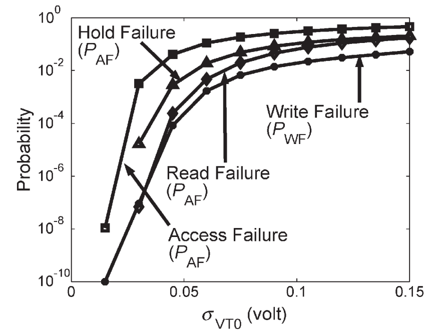

Fig. 9. Variation of failure probability with σ_{VT0}.

$$\begin{aligned} P_{\mathrm{COL}} &= 1 - (1 - P_{\mathrm{F}})^{N_{\mathrm{ROW}}} \\ P_{\mathrm{MEM}} &= \sum_{i=N_{\mathrm{RC}}+1}^{N_{\mathrm{COL}}+N_{\mathrm{RC}}} \binom{N_{\mathrm{COL}}+N_{\mathrm{RC}}}{i} \\ &\quad \times P_{\mathrm{COL}}^{i}(1 - P_{\mathrm{COL}})^{N_{\mathrm{COL}}+N_{\mathrm{RC}}-i}. \end{aligned} \tag{25}$$

H. Effect of Correlation of Threshold Voltage of Different Transistors

In the previous discussions, we have assumed the V_{t} of different transistors in an SRAM cell are independent random variables. This assumption is valid if we are considering the V_{t} variation due to RDF [1]–[3]. However, in general, due to the presence of systematic intra-die variations (e.g., V_{t} variation due to channel-length variation), the V_{t}s of different transistors can be correlated. In this section, we will investigate the impact of such correlation on the failure probabilities. The proposed

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

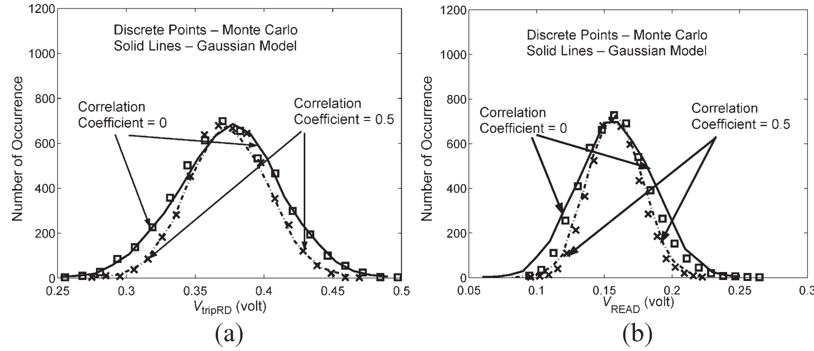

Fig. 10. Effect of correlation of V_{t} on the distribution of (a) V_{TRIPRD} and (b) V_{READ}.

models in Section III-A can be easily extended to account for the effect of correlation as shown below [13]

$$\mu_y = f(\eta_1,\ldots,\eta_n) + \frac{1}{2}\sum_{i=1}^{n} \left.\frac{\partial^2 f}{\partial (x_i)^2}\right|_{\eta_i} \sigma_i^2 + \sum_{k=1}^{N}\sum_{i=1;i\neq k}^{N} \left.\frac{\partial^2 f}{\partial x_i \partial x_k}\right|_{\eta_i,\eta_k} r_{(i,k)}\sigma_i\sigma_k$$
$$\sigma_y^2 = \sum_{i=1}^{n}\left(\left.\frac{\partial f}{\partial (x_i)}\right|_{\eta_i}\right)^2 \sigma_i^2 + 2\sum_{k=1}^{N}\sum_{i=1;i\neq k}^{N}\left(\left.\frac{\partial f}{\partial x_i}\right|_{\eta_i}\right)\left(\left.\frac{\partial f}{\partial x_k}\right|_{\eta_k}\right) r_{(i,k)}\sigma_i\sigma_k \tag{26}$$

where $r_{(i,k)}$ is the correlation coefficient between x_i and x_j. Since all the transistors in a cell are in very close spatial proximity, we can assume the correlation coefficients among different transistors are same, i.e., $r_{(i,k)} = r$ for all i and k. Although the proposed model can handle different values of correlation coefficients for different transistors, we have used the above assumption to simplify the calculation. Fig. 10 shows that the extended analytical model shown in (26) closely follows the Monte Carlo distributions for V_{READ} and V_{TRIPRD} in the presence of correlation. The Gaussian model for T_{ACCESS}, T_{WRITE}, and V_{HOLD} also closely follow the corresponding Monte Carlo distributions. It has been also observed that the effect of the correlation on the mean value of V_{READ}, V_{TRIPRD}, T_{ACCESS}, T_{WRITE}, and V_{HOLD} is not very significant, whereas it has a stronger impact on the standard deviations of the above-mentioned parameters.

It can be observed from the discussions in the previous sections and from Table II that for read, write, and hold failures to occur the threshold voltages of different transistors need to shift in opposite directions. In other words, these three failures are enhanced by the increase in the mismatch between the V_{t}s of the transistors. For example, the read failure increases if δV_{t} of AX_R becomes negative and δV_{t} of N_R becomes positive, which increases V_{READ}. If δV_{t} of AX_R and δV_{t} of N_R are completely uncorrelated (independent random variables with mean = 0) and $\delta V_{\text{t}_{\text{AXR}}} < 0$, δV_{t} of N_R has equal probability of being positive or negative. On the other hand, if they are positively correlated and $\delta V_{\text{t}_{\text{AXR}}} < 0$, δV_{t} of N_R has a lower probability of being negative. This reduces the probability of occurrence of a high value of V_{READ}. Similarly, the probability of occurrence ($\delta V_{\text{t}_{\text{AXR}}} > 0$ and $\delta V_{\text{t}_{\text{NR}}} < 0$) (i.e., low value of V_{READ}) also reduces. Hence, the spread of the distribution of V_{READ} reduces with an increase in the correlation [Fig. 10(a)]. Similar reduction in the spread of V_{TRIPRD} is also observed with an increase in the correlation between $\delta V_{\text{t}_{\text{NL}}}$ and $\delta V_{\text{t}_{\text{PL}}}$ [Fig. 10(b)]. Consequently the standard deviation of the variable $Z = (V_{\text{READ}} - V_{\text{TRIPRD}})$ reduces with an increase in the correlation coefficient [Fig. 11(a)], which results in a reduction of the read-failure probability [Fig. 11(b)]. Since the write failure and the hold failure are also enhanced by the mismatch between the transistor threshold voltages, an increase in the correlation reduces both the write and the hold failure probabilities [Fig. 11(b)] by decreasing the standard deviation of T_{WRITE} and V_{HOLD}, respectively [Fig. 11(a)]. However, increase in the correlation increases the standard deviation of T_{ACCESS} [Fig. 11(a)], and hence, the access-time failure probability [Fig. 11(b)]. This is because of the fact that access-time failure is caused if $\delta V_{\text{t}_{\text{AXR}}} > 0$ and $\delta V_{\text{t}_{\text{NR}}} > 0$. The correlation between $\delta V_{\text{t}_{\text{AXR}}}$ and $\delta V_{\text{t}_{\text{NR}}}$ increases the probability of occurrence of this event (i.e., if $\delta V_{\text{t}_{\text{AXR}}} > 0$ due to positive correlation $\delta V_{\text{t}_{\text{NR}}}$ has a higher probability of being positive and vice-versa). Hence, an increase in the correlation coefficient increases the access time failure probability.

In this section, we have analyzed the impact of the correlation of the threshold voltage on the failure probabilities. It has been observed that the read, write, and hold failures reduce with an increase in the correlation whereas the access-time failure probability increases. However, due to extremely small geometry of the SRAM cell, the effect of RDF on the threshold voltage is very high in the cell transistors [4]. Hence, the correlation between V_{t}s is expected to be small particularly for the SRAM designed with the nanoscaled CMOS devices. In the rest of the paper, we have neglected the effect of the correlation while estimating the failure probabilities and proposing a statistical approach for designing a robust memory. It should be noted that neglecting the correlation results in an overestimation of the read, write, and hold failures (i.e., pessimistic estimation) and an underestimation of access time failure (i.e., optimistic estimation).

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

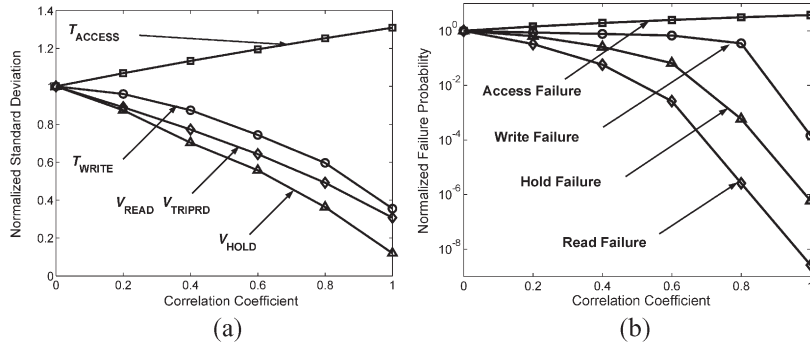

Fig. 11. Effect of correlation of V_t on (a) standard deviation of T_{ACCESS}, T_{WRITE}, $V_{\mathrm{READ}} - V_{\mathrm{TRIPRD}}$, and V_{HOLD}; and (b) failure probabilities. All the values are normalized to their corresponding values for the completely uncorrelated case (i.e., correlation coefficient $= 0$). (a) Standard deviation versus correlation. (b) Failure probability versus correlation.

IV. Failure-Probability Estimation Using Simple Long-Channel Transistor Model

In the estimation method described in the previous sections, the use of sophisticated short-channel transistor model increases the accuracy. However, due to their complex nature, these models require numerical solutions of (10), (11), (14), (19), and (22) to obtain the failure probabilities (i.e., KCL at the intermediate nodes need to be solved numerically). Numerical solution increases the estimation time. Hence, to reduce the computation cost, we have derived a set of analytical models to estimate the failure probabilities using long-channel transistor equations (with short-channel threshold-voltage model). These analytical models (although less accurate) allow a fast estimation of the failure probabilities. These models are particularly useful in the generation of a good initial guess for the statistical-optimization problem discussed in Section VI. The numerical models are used in the final optimization stage to ensure the accuracy.

A. Long-Channel Transistor Equations

The long-channel transistor characteristics that are used in deriving the analytical models are summarized below

$$\begin{aligned} I_{\mathrm{dsub}} &= \beta(m-1)\left(\frac{kT}{q}\right)^2 \\ &\quad\times \exp\left(\frac{q(V_{\mathrm{gs}}-V_{\mathrm{th}})}{mkT}\right)\left(1-\exp\left(\frac{-qV_{\mathrm{ds}}}{kT}\right)\right) \\ I_{\mathrm{dlin}} &= \frac{\beta\left[(V_{\mathrm{gs}}-V_{\mathrm{th}})V_{\mathrm{ds}}-\left(\frac{m}{2}\right)V_{\mathrm{ds}}^2\right]}{1+\left(\frac{\mu_{\mathrm{eff}}V_{\mathrm{ds}}}{v_{\mathrm{sat}}L}\right)} \text{ and} \\ I_{\mathrm{dsat}} &= \beta\frac{(V_{\mathrm{gs}}-V_{\mathrm{th}})^2}{2m} \\ \beta &= \frac{\mu_{\mathrm{eff}}C_{\mathrm{ox}}W}{L},\ I_{\mathrm{sub0}} = \beta(m-1)\left(\frac{kT}{q}\right)^2 \end{aligned} \tag{27}$$

where μ_{eff} is the effective mobility, and m is the body effect coefficient $(m = 1 + 3T_{\mathrm{ox}}/(\text{Width of depletion region}))$. In this work, we have used m and μ_{eff} to match the MEDICI simulation result as closely as possible.

B. Estimation of Failure Probabilities

Using a simple square-law model to estimate the ON current through a transistor and neglecting the contribution of the leakage currents, V_{READ} and V_{TRIPRD} are obtained from (assuming a short-channel V_{th} model)

$$V_{\mathrm{TRIPRD}} = \frac{\left[V_{\mathrm{DD}} - V_{\mathrm{tPL}} + V_{\mathrm{tNL}}\sqrt{\left(\frac{\beta_{\mathrm{NL}}}{\beta_{\mathrm{PL}}}\right)}\right]}{\left(1+\sqrt{\left(\frac{\beta_{\mathrm{NL}}}{\beta_{\mathrm{PL}}}\right)}\right)} \tag{28a}$$

$$\begin{aligned} I_{\mathrm{dsatAXR}} &\equiv 0.5\beta_{\mathrm{AXR}}\left(V_{\mathrm{WL}} - V_{\mathrm{READ}} - V_{\mathrm{tAXR}}\right)^2 \\ &= \beta_{\mathrm{NR}}\left(V_{\mathrm{DD}} - V_{\mathrm{tNR}} - 0.5V_{\mathrm{READ}}\right)V_{\mathrm{READ}} \\ &\equiv I_{\mathrm{dlinNR}}. \end{aligned} \tag{28b}$$

Assuming simplified square-law transistors models, the integration in (14) can be performed analytically as shown below

$$\begin{aligned} T_{\mathrm{WRITE}} &= C_L \int_{V_{\mathrm{DD}}}^{V_{\mathrm{WL}}-V_{\mathrm{tAXL}}} \frac{dV_L}{I_{\mathrm{dsatAXL}} - I_{\mathrm{dlinPL}}} \\ &\quad + C_L \int_{V_{\mathrm{WL}}-V_{\mathrm{tAXL}}}^{V_{\mathrm{tNR}}+V_{\mathrm{tPL}}} \frac{dV_L}{I_{\mathrm{dlinAXL}} - I_{\mathrm{dlinPL}}} \\ &\quad + C_L \int_{V_{\mathrm{tNR}}+V_{\mathrm{tPL}}}^{V_{\mathrm{TRIP}}} \frac{dV_L}{I_{\mathrm{dlinAXL}} - I_{\mathrm{dsatPL}}}. \end{aligned} \tag{29}$$

The analytical model for access-time failure can be obtained by using the solution of the V_{READ} from (28b). The estimated value of V_{READ} can be used to determine the saturation current of a transistor with $V_{\mathrm{GS}} = V_{\mathrm{DD}} - V_{\mathrm{READ}}$, $V_{\mathrm{DS}} = V_{\mathrm{DD}} - V_{\mathrm{READ}}$, and $V_{\mathrm{BS}} = -V_{\mathrm{READ}}$ (in this step, we can use short-channel saturation-current model also as we need to estimate the current for one bias point only). This allows an analytical estimation of T_{ACCESS} in (19).

In order to obtain an estimate of hold voltage (V_{DDHmin}), let us assume at $V_{\mathrm{DD}} = V_{\mathrm{DDH}}$, that all transistors are in subthreshold region. The current of A_{XL} can be neglected due to its small V_{ds} drop. Using these assumption and neglecting

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

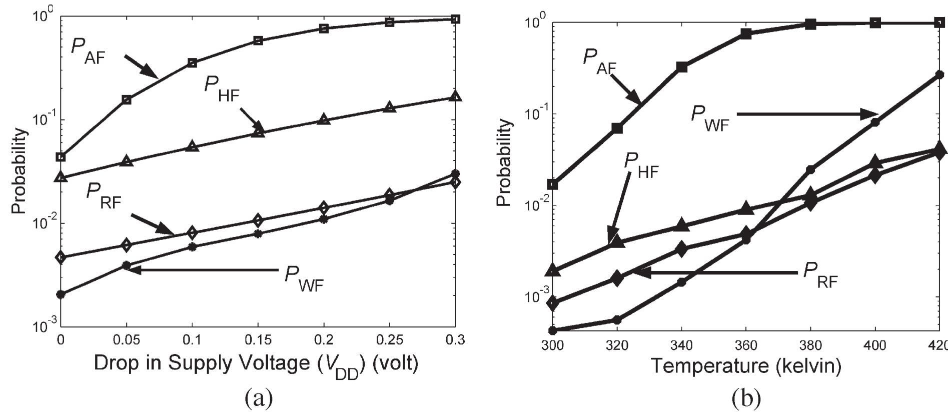

Fig. 12. Impact of (a) supply-voltage drop and (b) temperature increase on the failure probability.

the contribution of the gate leakage and the junction leakage currents (as $V_{\rm DD}$ is low), the KCL at node L becomes

$$\begin{aligned} Id_{\rm NL} &= Id_{\rm PL} \\ &\Rightarrow \beta_{\rm NL}(m_n - 1)\exp\left(\frac{V_R - V_{\rm t_{NL}}}{m_n v_T}\right) \\ &\quad\times\left(1 - \exp\left(\frac{-V_L}{v_T}\right)\right) \\ &= \beta_{\rm PL}(m_p - 1)\exp\left(\frac{V_{\rm DDH} - V_R - V_{\rm t_{PL}}}{m_p v_T}\right) \\ &\quad\times\left(1 - \exp\left(\frac{V_L - V_{\rm DDH}}{v_T}\right)\right) \\ &\Rightarrow \text{Analytical solution for } V_L. \end{aligned} \tag{30}$$

Since node R is storing "0," due to large $V_{\rm ds}$ for P_R and $A_{\rm XR}$ [neglecting the $\exp(-V_{\rm ds}/v_T)$ term in their current equations], the KCL at node R can be approximated as

$$\begin{aligned} Id_{\rm NR} &= Id_{\rm PR} + Id_{\rm AXR} \\ &\Rightarrow \beta_{\rm NR}(m_n - 1)\exp\left(\frac{V_L - V_{\rm t_{NR}}}{m_n v_T}\right) \\ &\quad\times\left(1 - \exp\left(\frac{-V_R}{v_T}\right)\right) \\ &= \beta_{\rm PR}(m_p - 1)\exp\left(\frac{V_{\rm DDH} - V_L - V_{\rm t_{PR}}}{m_p v_T}\right) \\ &\quad+ \beta_{\rm AXR}(m_n - 1)\exp\left(\frac{-V_R - V_{\rm t_{AXR}}}{v_T}\right) \\ &\Rightarrow \text{Analytical solution for } V_R. \end{aligned} \tag{31}$$

In the above equation, the m_n of $A_{\rm XR}$ is approximated to 1, in order to be able to derive an analytical solution for V_R. Assuming initial values of $V_L = V_{\rm DDH}$ and $V_R = 0$, the above equations can be iteratively solved to find the solution for V_L and V_R and therefore to decide whether the cell fails (flips) at the assigned supply voltage. The minimum hold voltage ($V_{\rm DDHmin}$) can be found by a binary search.

The simplified analytical models of $V_{\rm READ}$, $V_{\rm TRIPRD}$, $T_{\rm WRITE}$, $T_{\rm ACCESS}$, and $V_{\rm HOLD}$ match the Monte Carlo simulation result and the numerical models reasonably closely (Figs. 5–8). It should be remembered that, although the currents are estimated using the long-channel equations, the threshold voltage is still evaluated using the short-channel model, which increases the accuracy. Moreover, the values of $V_{\rm READ}$, $V_{\rm TRIPRD}$, $T_{\rm WRITE}$, and $V_{\rm HOLD}$ depend on the relative magnitude of different current components and not on the absolute values of the currents. For example, $V_{\rm READ}$ depends on the relative strength of $I_{\rm dsatAXR}$ and $I_{\rm dlinNR}$. Thus, the error introduced by using the long-channel model in the estimation of $V_{\rm READ}$ is less than the error introduced in the magnitudes of the currents themselves. Due to these reasons, use of the simplified long-channel current models does not introduce high errors in the estimation of the failure probabilities. Thus, these simplified models can be used to generate preliminary estimation of the failure probabilities. However, the analytical models developed here neglect the contributions of the leakage currents while determining the node voltages [e.g., see (10), (11), (14)]. As the magnitude of the leakage currents increases, the error in these simplified models also becomes higher.

V. Sensitivity Analysis of Failure Probability

A. Impact of Supply Voltage ($V_{\rm DD}$) and Temperature

A drop in the supply voltage increases the cell failure probabilities. [Fig. 12(a)]. The impact of supply voltage drop is most significant for the access-time failure. The derived model can also be extended to include $V_{\rm DD}$ of a cell as an independent Gaussian random variable. Fig. 12(b) shows that the failure probabilities increase with an increase in the temperature. The impacts of temperature increase are more severe on the access-time failure and the write failure because of: 1) reduction in the ON current of the access transistors; and 2) increase in the junction capacitances.

B. Transistor Size and Cell-Failure Probability

The length and width of different transistors of the cell (i.e., $L_{\rm nax}$, $W_{\rm nax}$, $L_{\rm npd}$, $W_{\rm npd}$, $L_{\rm pup}$, and $W_{\rm pup}$,) impact the cell-failure probability principally by modifying: 1) the nominal values of $T_{\rm ACCESS}$, $V_{\rm TRIP}$, and $V_{\rm READ}$, $T_{\rm WRITE}$, and $V_{\rm DDHmin}$; 2) the sensitivity of these parameters to $V_{\rm t}$ variation, thereby changing the mean and the variance of these parameters; and

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

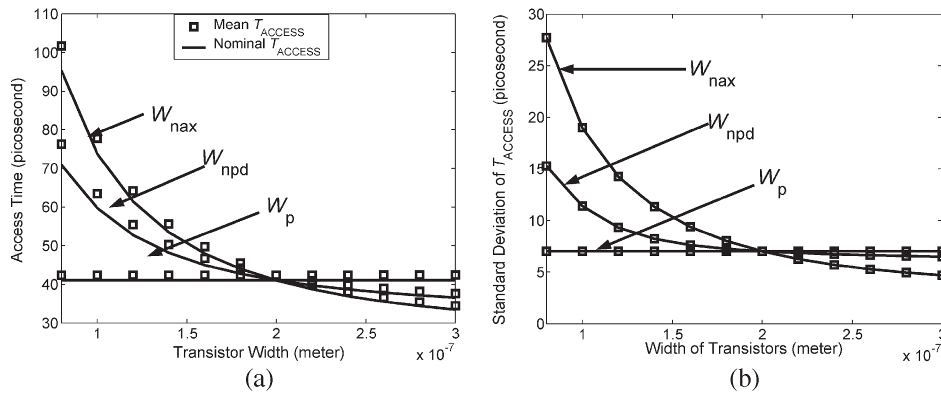

Fig. 13. Impact of transistor size on distributions of T_{ACCESS}. (a) Mean of T_{ACCESS}. (b) Standard deviation of T_{ACCESS}.

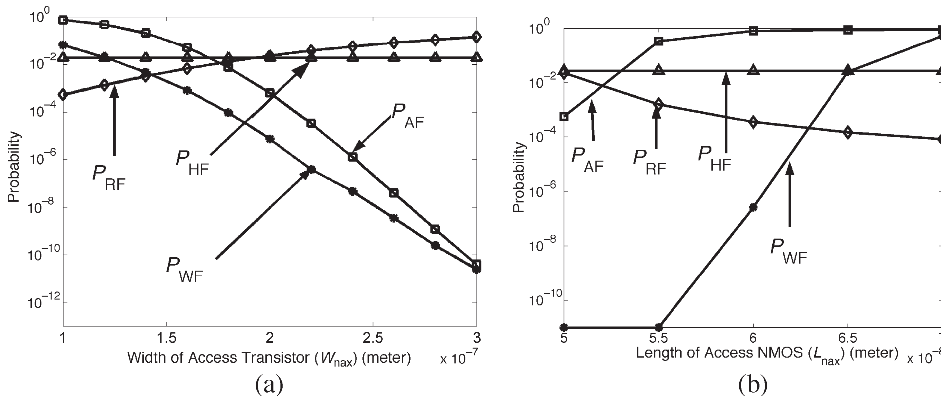

Fig. 14. The impact of (a) width (W_{nax}) and (b) length (L_{nax}) of access NMOS transistor on failure.

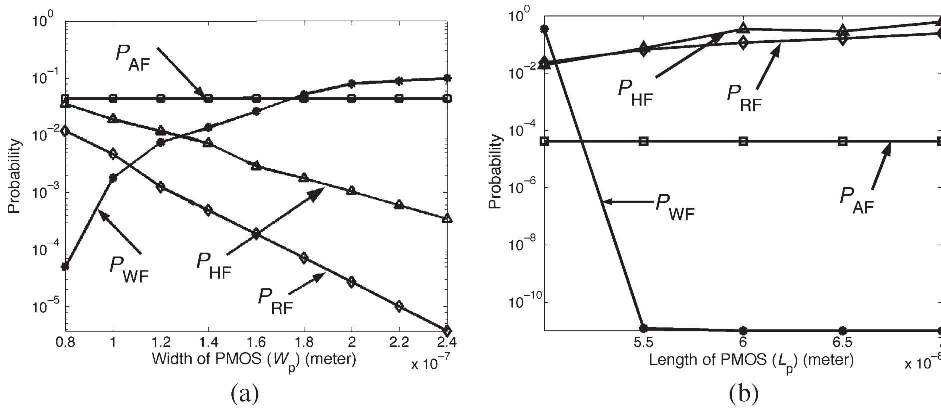

Fig. 15. The impact of (a) width (W_{pup}) and (b) length (L_{pup}) of pull-up PMOS transistor on the failure.

3) the standard deviation of the V_{t} variation [see (2)]. For example, Fig. 13 shows that along with the nominal value, the mean and the standard deviation of T_{ACCESS} significantly vary with W_{npd} and W_{nax}. In this section, we study the impact of variation of strength of different transistors on the cell-failure probability. Fig. 14 shows that a weak access transistor (small W_{nax} and/or large L_{nax}) reduces P_{RF} (V_{READ} decreases); however, it increases P_{AF} and P_{WF} (Fig. 14) and has very small impact on P_{HF}. Reducing the strength of the PMOS pull-up transistors (by decreasing W_p or increasing L_p) reduces P_{WF} (reducing I_{dsPL}), but increases P_{RF} (lowers V_{TRIPRD}). P_{AF} does not depend strongly on PMOS strength (Fig. 15). P_{HF} improves with an increase in W_p or a reduction in L_p as node L is more strongly coupled to the supply voltage ($V_L \rightarrow V_{\mathrm{DDH}}$) (Fig. 15). Increasing W_{npd} (and/or reducing L_{npd}) increases the strength of pull-down NMOS transistors (N_L and N_R). This reduces P_{RF} ($V_{\mathrm{READ}} \downarrow$) and P_{AF} by increasing the strength of N_R (Fig. 16). Increase in width of N_R has little impact on P_{WF}. Although it slightly increases the nominal value of T_{WRITE}, the reduction of σ_{VT} of N_R [see (2)] tends to reduce σ_{TWRITE} and hence P_{WF} remains almost constant (Fig. 16). However, increasing L_{npd} reduces both T_{WRITE} (by increasing the trip-point of $P_R - N_R$) and σ_{VT} of N_R [see (2)], which results in a significant reduction in P_{WF} (Fig. 16). An increase in the V_{TRIP} of $P_R - N_R$ initially reduces P_{HF} with the increase in the strength of N_R (i.e., higher W_{npd} or lower L_{npd}).

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

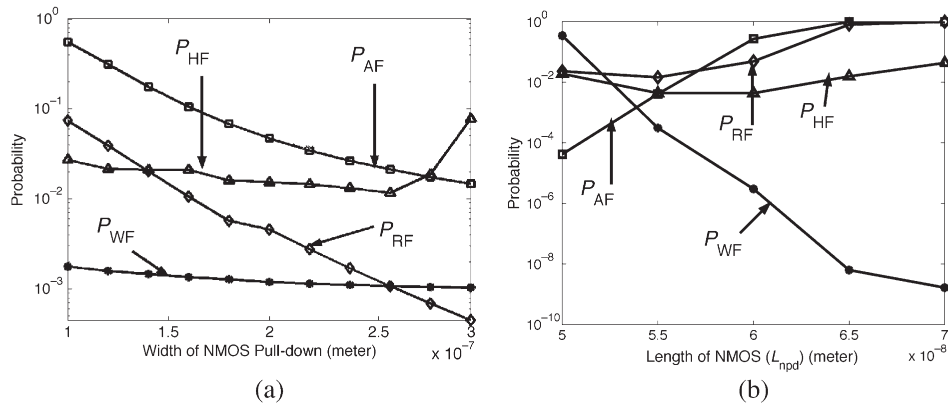

Fig. 16. The impact of strength of pull-down NMOS transistor on the failure probabilities: (a) width (W_{npd}); (b) Length (L_{npd}).

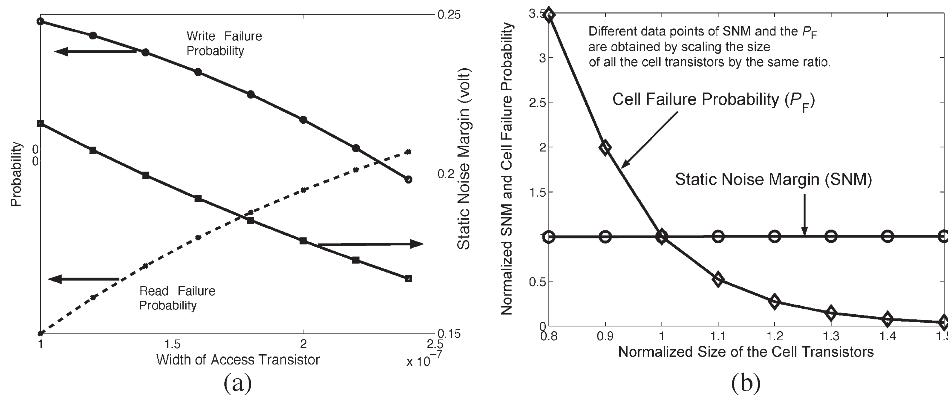

Fig. 17. Variation of SNM and failure probability with (a) width of the access transistors; and (b) normalized cell area.

However, a higher width of N_L (or lower L_{npd}) reduces V_L (from the applied V_{DDH}) due to an increase in the leakage of N_L. Consequently, a very high W_{npd} (or low L_{npd}) increases the P_{HF} (Fig. 16).

Due to the variation in the failure probability, the choice of the transistor sizes has a strong impact on the yield. Hence, it can be concluded that a statistical approach to the design of transistor sizes is necessary to maximize the yield. The derived failure-probability models can be effectively used for such statistical optimizations.

C. Static-Noise Margin and Cell-Failure Probability

The static-noise margin (SNM) of a cell is often used as a measure of the robustness of an SRAM cell against flipping [4]. However, an increase in SNM makes the cell difficult to write by increasing its data-holding capability, which increases write failures. For example, reducing the size of the access transistor improves the SNM [4]. On the other hand, reducing W_{nax} decreases read-failure probability, but increases the write-failure probability [Fig. 17(a)]. Hence, the size of W_{nax} that results in a maximum SNM does not correspond to minimum-failure probability [Fig. 17(a)]. Moreover, increasing the size of all the transistors in a cell by the same factor does not modify the SNM. However, an increase in the size of all the transistors in a cell considerably reduces its failure probability by reducing the standard deviation of the V_{t} variation [Fig. 17(b)]. Using the proposed models, it is observed that SNM does not have a strong relationship with the parametric failure of the memory. Consequently, an increase in the SNM does not necessarily reduce the overall failure probability and an SNM-based analysis of the cell does not directly correspond to the memory-failure probability and the yield. Hence, a statistical analysis and design of the cells and memory structure is necessary to ensure acceptable yield in nanometer regimes.

VI. Statistical Design of the SRAM Array

A. SRAM Yield-Estimation Model

The hierarchy of the failure probabilities in an SRAM array is shown in Fig. 18. We first estimate the failure probability of a cell (P_{F}) [see (24)]. The cell failure probability (P_{F}) is used to determine the probability of failure of a column (P_{COL}) using the total number of cells in that column (i.e., column length) [see (25)]. The estimated value of P_{COL} is then used to calculate the failure probability of a memory array (P_{MEM}). The memory-failure probability depends on the number of actual columns (N_{COL}) and the redundant columns (N_R) [see (25)]. The memory-failure probability is directly related to the yield of the memory chip. To estimate the yield, we have used Monte Carlo simulations for inter-die distributions of L, W, and V_{t} (assumed to be Gaussian). For each inter-die value of the parameters (say L_{INTER}, W_{INTER}, and $V_{\mathrm{t_{INTER}}}$)

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

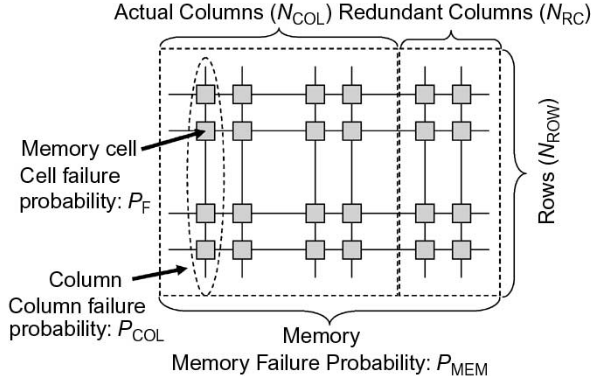

Fig. 18. Memory hierarchy and failure probability.

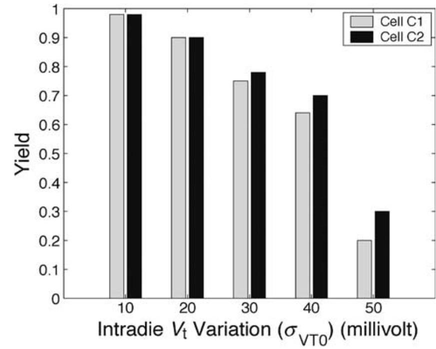

Fig. 19. Variation of yield with σ_{VT0}. The results are obtained using cells C1 and C2 shown in Table I.

we estimate P_{F}, P_{COL}, and P_{MEM} considering the intra-die distribution of δV_{t}. Finally, the yield is defined as

$$\mathrm{Yield} = 1 - \left(\frac{\sum_{\mathrm{INTER}} P_{\mathrm{MEM}} \left(L_{\mathrm{INTER}}, W_{\mathrm{INTER}}, V_{\mathrm{t_{INTER}}} \right)}{N_{\mathrm{INTER}}} \right) \tag{32}$$

where N_{INTER} is the total number of inter-die Monte Carlo simulations (i.e., total number of chips). An increase in the intra-die variation (i.e., σ_{VT0}) increases the memory-failure probability, thereby reducing the yield (Fig. 19). In this estimation, we have assumed a standard deviation of 7% for inter-die distribution of L, W, and V_{t}, $N = 32$ cells per column, $N_{\mathrm{COL}} = 512$, $N_{\mathrm{RC}} = 24$.

As observed in Section V, the cell-failure probabilities are dependent on memory-cell configurations including transistor sizing. Hence, proper choice of the size of the cell transistors has a strong impact on the memory yield [Fig. 20(a)]. Yield also depends strongly on the design of the memory architecture. Increasing the number of cells in a column (column length or number of rows) increases the cell-failure probability (particularly P_{AF} as C_{BL} and I_{BL} increases in (19), resulting in higher T_{ACCESS}). Moreover, P_{COL}, and hence P_{MEM}, increases significantly with the column length [see (25)]. However, for a constant memory size, increasing column length reduces the number of columns, which tends to reduce P_{MEM} (assuming a constant redundancy). Fig. 20(b) shows the variation of column-failure probability, the memory-failure probability, and yield of a 2-kB cache with the column length (number of redundant columns kept constant). It can be observed that yield strongly depends on the column length. Therefore, the memory yield is also impacted by memory architecture. Hence, the cell configurations and the memory architecture can be optimized for maximizing memory yield.

In order to improve the yield of the memory, the memory-failure probability (P_{MEM}) needs to be minimized. From (25), it can be observed that P_{MEM} depends on: 1) length (L_{nax}, L_{npd}, L_{pup}) and width (W_{nax}, W_{npd}, W_{pup}) of the transistors in the cell that determine P_{F}; 2) number of cells in a row (i.e., number of rows N_{ROW}), which determines P_{COL}; and 3) number of actual columns (N_{COL}) and the number of redundant columns (N_{RC}). It should be noted that N_{COL} and N_{ROW} are not independent since $N_{\mathrm{COL}} \times N_{\mathrm{ROW}} =$ size of the memory. However, N_{ROW} and N_{COL} are principally determined by the memory architecture (e.g., memory pitch, complexity of column and row decoder, etc.). Any modifications of N_{COL} and N_{ROW} have to consider their impact on architectural-level parameters, such as memory pitch, complexity of column and row decoder, etc. Thus, to simplify the present design problem, we have not considered N_{ROW} and N_{COL} as design parameters. Minimization of P_{MEM} has to consider the impact on the total leakage and the total area (A_{MEM}). In the following section, we present the models we have used to estimate the leakage and the area of an SRAM array.

B. Statistical Estimation of Leakage in SRAM

The leakage current in SRAM cell is the major contributor to the power in the SRAM cell array. Hence, the optimization of the cell structure has to consider its impact on the cell leakage. The total leakage in a cell principally consists of the subthreshold leakage, the gate leakage, and the junction band-to-band tunneling leakage through different transistors in the cell (Fig. 1) [10]. Considering all of the different components, the total leakage of the cell can be computed as

$$\begin{aligned} I_{\mathrm{sub}} &= I_{\mathrm{subAXR}} + I_{\mathrm{subNL}} + I_{\mathrm{subPR}} \\ I_{\mathrm{jn}} &= 2I_{\mathrm{jnAXL}} + I_{\mathrm{jnAXR}} + I_{\mathrm{jnNL}} + I_{\mathrm{jnPR}} \\ I_{\mathrm{gate}} &= I_{\mathrm{gdAXL}} + I_{\mathrm{gsAXL}} + I_{\mathrm{gdAXR}} + I_{\mathrm{gdPR}} \\ &\quad + I_{\mathrm{gdNR}} + I_{\mathrm{gsNR}} + I_{\mathrm{gdPL}} + I_{\mathrm{gsPL}} + I_{\mathrm{gdNL}} \\ I_{\mathrm{leak}} &= I_{\mathrm{sub}} + I_{\mathrm{jn}} + I_{\mathrm{gate}}. \end{aligned} \tag{33}$$

We have used the leakage current expressions presented in [10] to evaluate different leakage components and the total cell leakage. However, the V_{t} variation in the transistors of a cell results in significant variation in the leakage (particularly, the subthreshold leakage) of the cell. The mean (μ_{LCELL}) and the standard deviation (σ_{LCELL}) of the leakage of a cell considering RDF-induced V_{t} fluctuation can be obtained using the process described in (3). Since the subthreshold leakage is an exponential function of the threshold voltage, we have assumed a lognormal PDF to describe the distribution of the cell leakage [13]. Fig. 21 shows that the lognormal distribution

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

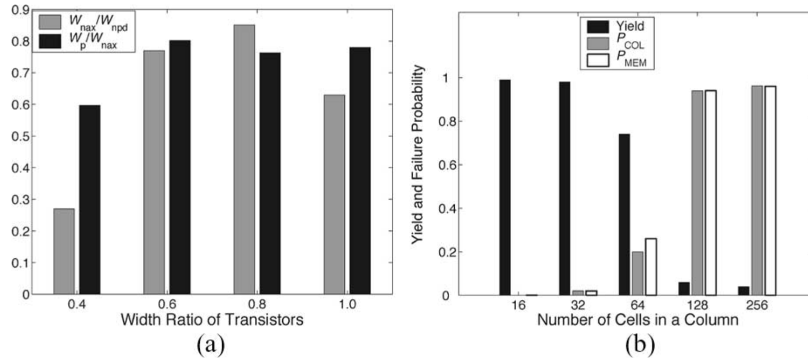

Fig. 20. Impact of (a) circuit (transistor size) and (b) architecture [number of row (column length) and number of columns] on yield. In (b), transistor sizes were chosen to maximize yield following (a). ($\sigma_{\mathrm{VT0}} = 20$ mV).

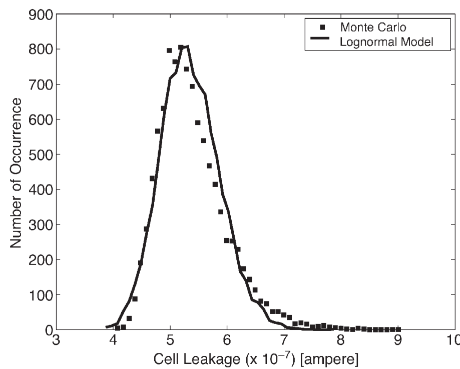

Fig. 21. Distributions leakage of SRAM cell (I_{leak}).

model with the mean and the standard deviation estimated using (3) closely follows the Monte Carlo simulation results. Considering the different cells to be independent identical random variables, (i.e., the mean and the standard deviation of the leakage of all the cells are same) the total SRAM array leakage is given by

$$I_{\mathrm{LeakMem}} = \sum_{i=1}^{N_{\mathrm{CELLS}}} I_{\mathrm{leak}} = \sum_{i=1}^{N_{\mathrm{ROW}}(N_{\mathrm{COL}}+N_{\mathrm{RC}})} I_{\mathrm{leak}} \quad (34)$$

where I_{leak} is the random variable representing the leakage of a cell. Applying the Central Limit Theorem [13], the distribution of the total leakage can be approximated as a Gaussian one with the mean (μ_{L}) and the standard deviation (σ_{L}) given by

$$\mu_L = N_{\mathrm{CELLS}}\mu_{\mathrm{LCELL}} \text{ and } \sigma_L^2 = N_{\mathrm{CELLS}}\sigma_{\mathrm{LCELL}}^2. \quad (35)$$

To consider the effect of leakage distribution in the statistical design of the SRAM array, we have defined the probability (P_L) that the total memory leakage will meet a given leakage bound as

$$P_{\mathrm{LeakMem}} = P(I_{\mathrm{LeakMem}} \leq I_{\mathrm{LMAX}}) = \Phi\left(\frac{I_{\mathrm{LMAX}} - \mu_{\mathrm{LMEM}}}{\sigma_{\mathrm{LMEM}}}\right). \quad (36)$$

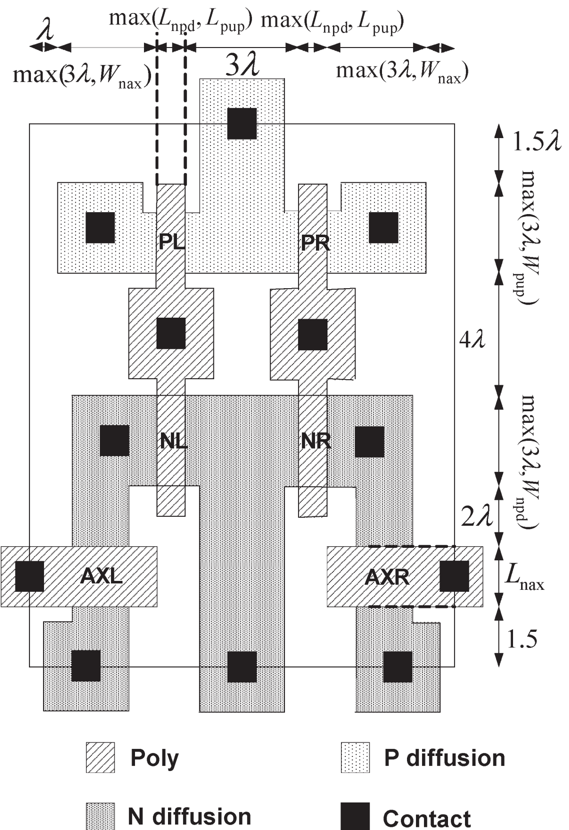

Fig. 22. SRAM cell layout.

C. Area Estimation of SRAM

Using the layout shown in Fig. 22 [16], the total cell area can be computed as

$$\begin{aligned} X_{\mathrm{cell}} &= 5\lambda + 2\max(3\lambda, W_{\mathrm{nax}}) + 2\max(L_{\mathrm{npd}}, L_{\mathrm{pup}}) \\ Y_{\mathrm{cell}} &= 9\lambda + \max(3\lambda, W_{\mathrm{pup}}) + \max(3\lambda, W_{\mathrm{npd}}) + L_{\mathrm{nax}} \\ A_{\mathrm{cell}} &= X_{\mathrm{cell}} \times Y_{\mathrm{cell}} \end{aligned} \quad (37)$$

where λ is the minimum feature size of a technology. In our model, $\lambda \approx L_{\min} = 50$ nm. Although, there are different types of layout possible for the SRAM cell, we have only considered the one shown in Fig. 22 for the sake of simplicity. The total

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

memory area, including the area of the redundant columns, is given by

$$\begin{aligned} A_{\mathrm{actual}} &= N_{\mathrm{ROW}} N_{\mathrm{COL}} A_{\mathrm{cell}} \\ A_{\mathrm{redundant}} &= N_{\mathrm{ROW}} N_{\mathrm{RC}} A_{\mathrm{cell}} \\ A_{\mathrm{MEM}} &= A_{\mathrm{actual}} + A_{\mathrm{redundant}} \\ &= N_{\mathrm{ROW}} (N_{\mathrm{COL}} + N_{\mathrm{RC}}) A_{\mathrm{cell}} \end{aligned} \tag{38}$$

where A_{actual} is the required memory area (given by the memory size) and $A_{\mathrm{redundant}}$ is the area overhead of the redundant columns.

D. Statistical-Design Procedure

In order to improve the yield of an SRAM array under parameter variation, we have developed a statistical-design method to reduce the memory-failure probability under the area, the performance, and the leakage constraints. In the proposed method, the size of the different transistors in a cell and the number of redundant columns in an array is properly chosen to minimize the memory-failure probability. The proposed design method can be formulated as a minimization problem as shown below

$$\begin{aligned} &\text{Minimize } P_{\mathrm{MEM}} = f(\boldsymbol{X}) \\ &\text{where } \boldsymbol{X} \equiv [\, L_{\mathrm{nax}} \;\; W_{\mathrm{nax}} \;\; L_{\mathrm{npd}} \;\; W_{\mathrm{npd}} \;\; L_{\mathrm{pup}} \;\; W_{\mathrm{pup}} \;\; N_{\mathrm{RC}} \,] \\ &\text{Subject to :} \\ &P_{\mathrm{LeakMem}} \geq P_{\mathrm{LeakMin}} \\ &A_{\mathrm{MEM}} \leq \text{Maximum Area } (A_{\mathrm{MAX}}) \\ &E[T_{\mathrm{AC}}] = \mu_{\mathrm{TACCESS}} \\ &\qquad \leq \text{Maximum access time mean } (\mu_{\mathrm{TAC-MAX}}) \\ &\text{For all the parameters : } \{X_{\mathrm{MIN}}\} \leq \{X\} \leq \{X_{\mathrm{MAX}}\}. \end{aligned} \tag{39}$$

This is essentially a nonlinear optimization problem with nonlinear constraints. The upper bound on the mean-access time is given to ensure that robustness of the memory has not been achieved by significantly sacrificing the performance. It should be noted that the total memory area (i.e., actual + redundant cell area) is used as a constraint instead of only the cell area. This allows the tradeoff between the area of the individual cells and the amount of memory redundancy. Also, in order to consider the effect of leakage distribution, we have used a lower bound on the probability that the leakage will be less than the maximum allowable limit, as a constraint (instead of using a deterministic leakage bound).

Fig. 23 shows the basic steps of the proposed design process. It should be noted that N_{RC} (number of redundant columns) can have only discrete integer values. As we have considered redundancy only to correct the parametric failures, we allow the minimum value of N_{RC} to be zero [it should be noted that there can be other kinds of failures (e.g., hard short or open soft error) that will require $N_{\mathrm{RCmin}} > 0$]. The upper bound of N_{RC} is determined using (37) and (38) as shown below:

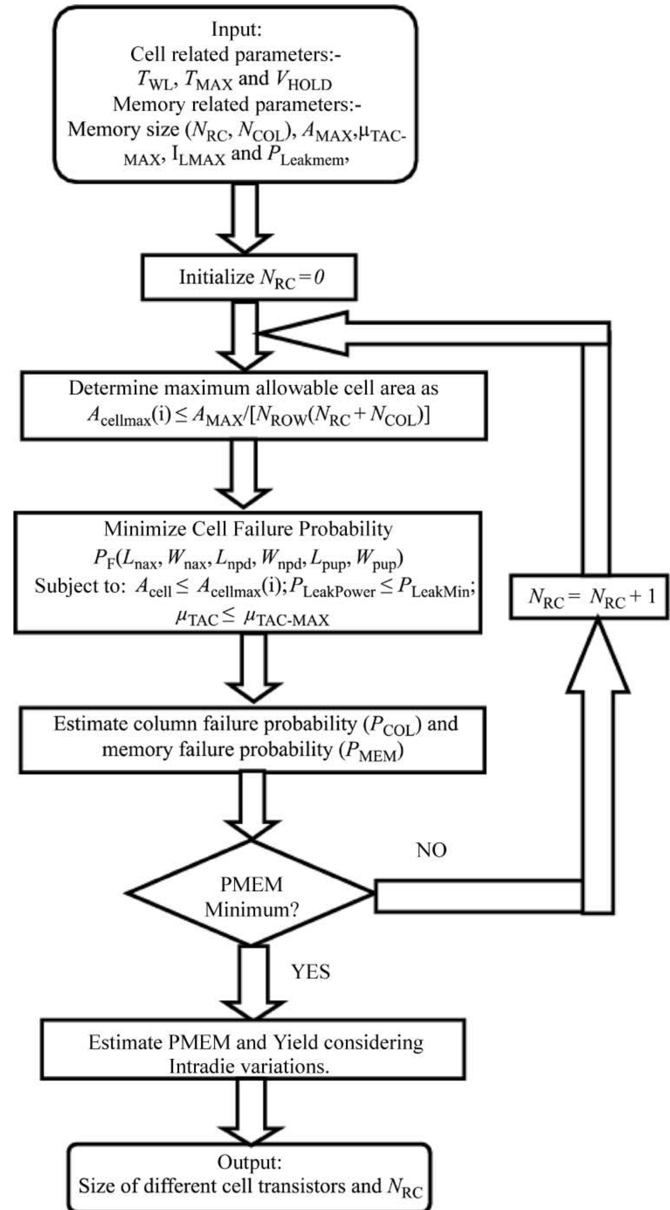

Fig. 23. Statistical-design procedure of SRAM.

$$\begin{aligned} X_{\mathrm{cellmin}} &= 5\lambda + 6\lambda + 2\lambda = 13\lambda \\ Y_{\mathrm{cellmin}} &= 9\lambda + 3\lambda + 3\lambda + \lambda = 16\lambda \\ A_{\mathrm{cellmin}} &= (X_{\mathrm{cellmin}})(Y_{\mathrm{cellmin}}) \\ N_{\mathrm{RCmax}} &= \frac{A_{\mathrm{MEM}} - N_{\mathrm{ROW}} N_{\mathrm{COL}} A_{\mathrm{cellmin}}}{A_{\mathrm{cellmin}} N_{\mathrm{ROW}}} \\ &= \frac{A_{\mathrm{MEM}}}{A_{\mathrm{cellmin}} N_{\mathrm{ROW}}} - N_{\mathrm{COL}}. \end{aligned} \tag{40}$$

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

1. **Initialize $x(0)$ and $\lambda(0)$.**
2. ***repeat***
3. **Evaluate**

$$L_d(x^{(x)},\lambda^{(k)}) = f(x^{(k)}) + \lambda^{(k)T} H(h(x^{(k)})) + (1/2)\left\|h(x^{(k)})\right\|^2$$

4. **Evaluate the direction of maximum descent:**

$$\Delta_x L_d(x^{(k)},\lambda^{(k)}) = y - x \text{ ; where } y = \min_{x' \in N(x^{(k)}) \cup \{x^{(k)}\}} L_d(x',\lambda)$$

5. **Update $x^{(k+1)}$ and $\lambda^{(k+1)}$ using:**

$$x^{(k+1)} = x^{(k)} + \Delta_x L_d(x^{(k)},\lambda^{(k)}) \text{ and } \lambda^{(k+1)} = \lambda^{(k)} + \varsigma H(h(x^{(k)}))$$

6. ***until*** $\Delta_x L_d(x^{(k)},\lambda^{(k)}) = 0$ ***and*** $h(x) = 0$
7. **Report optimum x, and $f(x)$.**

Fig. 24. Optimization procedure using DLM.

The minimization of $P_{\rm F}$ requires the estimation of the joint probabilities given in (5), which are computationally expensive. However, it should be noted that

$$\begin{aligned} P_{\rm F} &= P[A_{\rm F} + R_{\rm F} + W_{\rm F} + H_{\rm F}] \\ &\le P_{\rm AF} + P_{\rm RF} + P_{\rm WF} + P_{\rm HF} = P_{\rm FMOD}. \end{aligned} \tag{41}$$

This allows us to minimize $P_{\rm FMOD}$ instead of minimizing $P_{\rm F}$. Thus, the minimization problem in step 4 in Fig. 23 can be formulated as

$$\begin{aligned} &\text{Minimize } \boldsymbol{f}(\boldsymbol{X}) = P_{\rm FMOD} \\ &\text{where } \boldsymbol{X} = [\,L_{\rm nax} \quad W_{\rm nax} \quad L_{\rm npd} \quad W_{\rm npd} \quad L_{\rm pup} \quad W_{\rm pup}\,] \\ &\text{Subject to: } h_1(X) = \left(\frac{A_{\rm cell}}{A_{\rm cellmax}(i)}\right) - 1 \le 0 \\ &h_2(X) = \left(\frac{P_{\rm LeakMin}}{P_{\rm LeakPower}}\right) - 1 \le 0 \\ &h_3(X) = \left(\frac{\mu_{\rm TAC}}{\mu_{\rm TAC-MAXr}}\right) - 1 \le 0. \end{aligned} \tag{42}$$

The above problem can be solved using Lagrange multiplier-based algorithm [17], [18]. It should be noted that the parameters $L_{\rm nax}$, $W_{\rm nax}$, $L_{\rm npd}$, $W_{\rm npd}$, $L_{\rm pup}$, and $W_{\rm pup}$ can be considered both as continuous (i.e., any width and length larger than the minimum dimension λ is allowed) and discrete (i.e., only a finite set of discrete values of L and W are allowed). In this work, we have considered the discrete-variable space (to account for the minimum limit on the lithographic controllability of L and W). To solve the discrete space Lagrangian problem, we have used the Discrete Lagrangian method (DLM) described in [17]. The basic steps of this procedure are summarized in Fig. 24. In the described DLM process, inequality constraints are converted into the equality constraints using the function: $h_i(X) = \max(h_i(X), 0)$. $L_{\rm d}(x,\lambda)$ is the generalized augmented Lagrangian function [17], defined as

$$L_{\rm d}(x,\lambda) = f(x) + \lambda^{\rm T} H\left(h(x)\right) + \left(\frac{1}{2}\right)\|h(x)\|^2 \tag{43}$$

where H is a continuous transformation function satisfying $H(y) = 0$, if $y = 0$ [realized as $H(y) = y^2$] and $\lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ are the Lagrange multipliers.

It can be observed from Fig. 23 that the complexity of the proposed design flow has a polynomial dependence on the memory size. To understand this, let us analyze the dependence of the complexity of different steps of the proposed flow on the memory size (say, $M_{\rm Size} = N_{\rm COL} \times N_{\rm RC}$). First, it should be noted that in the proposed design flow, the number of loops required to find the optimum number of redundant columns ($\rm N_{RC_opt}$) increases linearly with an increase in the number of memory columns. Let us now analyze the dependence of the complexity of different steps in the main loop on the memory size. The complexity of the estimation of $P_{\rm MEM}$ depends linearly on $N_{\rm COL}$ [see (25)]. Moreover, $A_{\rm cellmax}$ reduces in successive iterations of the loop (due to increase in $N_{\rm RC}$). A reduction in $A_{\rm cellmax}$ reduces the feasible solution space for $\{L_{\rm nax}, W_{\rm nax}, L_{\rm npd}, W_{\rm npd}, L_{\rm pup}, W_{\rm pup}\}$ (i.e., all possible values of the length and widths of the transistors that satisfy the area constraint). This suggests that the complexity of the minimization problem of $P_{\rm FMOD}$ reduces in successive iterations (due to a reduction in $A_{\rm cellmax}$). To simplify the analysis, we assume that the complexity of $P_{\rm FMOD}$ minimization depends linearly on $A_{\rm cellmax}$. Hence, the complexity of the design ($C_{\rm Design}$) flow with respect to the memory size is given by

$$\begin{aligned} &C_{\rm Design} \\ &= \sum_{N_{\rm RC}(i)=1}^{N_{\rm RC_min}} \Bigg[\underbrace{O(N_{\rm COL})}_{P_{\rm MEM}\text{ Computation}} + \underbrace{O(A_{\rm cellmax})}_{P_{\rm FMOD}\text{ Minimization}} \Bigg] \\ &= \sum_{N_{\rm RC}(i)=1}^{N_{\rm RC_min}} \left[O(N_{\rm COL}) + O\left(\frac{kM_{\rm Size}}{\sqrt{M_{\rm Size}}N_{\rm RC}(i) + M_{\rm Size})}\right)\right] \\ &\qquad \text{where } A_{\rm cellmax} = \frac{A_{\rm MAX}}{N_{\rm ROW}\,(N_{\rm RC}(i) + N_{\rm COL})} \end{aligned} \tag{44}$$

where for simplicity, we assume $A_{\rm MAX} = kM_{\rm Size}$ and $N_{\rm ROW} = N_{\rm COL} = M_{\rm Size}$. Simplifying the above equation, we get [18]

$$\begin{aligned} &C_{\rm Design} \le O(N_{\rm COL}N_{\rm RC_min}) \\ &\qquad + O\left(N_{\rm RC_min} k \sqrt{M_{\rm Size}} \ln\left(\frac{N_{\rm RC_min} + \sqrt{M_{\rm Size}}}{\sqrt{M_{\rm Size}}}\right)\right) \\ &\quad \Rightarrow C_{\rm Design} \le O(N_{\rm COL}N_{\rm COL}) \le O(M_{\rm Size}) \\ &\qquad \left[\because N_{\rm RC_min} = O(N_{\rm COL}) = O(\sqrt{M_{\rm Size}})\right]. \end{aligned} \tag{45}$$

The above analysis shows that the complexity of the proposed design flow has a linear dependence on the memory size.

E. Results and Discussions

The statistical-design methodology described in the previous section is used to optimize the cell structure and the use of redundancy to minimize the memory-failure probability. We

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

TABLE III
RESULTS OF STATISTICAL DESIGN

	β_{nax}/β_p	β_{npd}/β_{na} x	I_{Leak}	T_{AC}	Yield
Initial Design (scaled from 250 nm) [19]	1.5	1.36	851 nA	55 ps	47%
Statistically Designed Cell	1.2	1.48	950 nA	46 ps	95%

have applied the developed design methodology to an SRAM cell given in [19]. The cell was originally designed in 250 nm and was scaled down to the 50-nm node. It is observed that application of the cell optimization successfully reduces the failure probability and improves yield (Table III). To understand how the optimization reduces the cell-failure probability, let us consider Fig. 25, which shows the read, write, access, and hold failure probabilities of the initial cell (i.e., scaled-down version of the cell from [19]) and the statistically designed cell. It can be observed from the figure that the initial cell had a large read and access failure, whereas write failure was low. This is because of the fact that, to improve the beta ratio between the pull-up PMOS and access transistor (to facilitate writing), the initial cell was designed with a longer PMOS. However, as explained in Section V, a weaker PMOS tends to increase the read failure and hold failure (Fig. 15). Hence, the optimization reduces the length of the PMOS resulting in a lower read and hold failure. The extra area obtained from reducing the length of the PMOS is used in increasing the width of the pull-down NMOS, which simultaneously improves the access failure (also reduces the access time) and read failure. However, the width of the NMOS cannot be increased arbitrarily as that would increase the hold failure by increasing the leakage through N_L (as explained in Section V). It should be noted that reduction in the channel length of the PMOS transistors results in an increase in the mean value of leakage. Hence, it can be observed that a statistical design of the SRAM cell can significantly improve the design yield.

The proposed design strategy allows to trade off between the redundancy area and the active cell area. Reducing the number of redundant columns allows more area for each of the actual cells. This reduces the failure probability of the cells, thereby reducing P_{MEM}. On the other hand, from (25) it can be observed that reducing N_{RC} will tend to increase P_{MEM}. Fig. 26 shows the variation of P_{MEM} with the variation of N_{RC}, considering constant A_{MEM}. It can be observed that increasing the redundancy beyond a certain point increases the memory-failure probability. Thus, increasing the redundancy at the cost of the silicon area does not necessarily reduce the memory-failure probability. It should be further noted that with the application of a higher value of the σ_{V_t0}, the optimized value of the redundancy (that minimizes failure probability) reduces. This indicates that with larger amount of variations, design of robust cell (with larger area) is more effective in reducing the failure probability (improving yield) as compared to increasing the number of redundant columns (at the cost of reducing the cell area). It should be noted that, in this analysis, we have

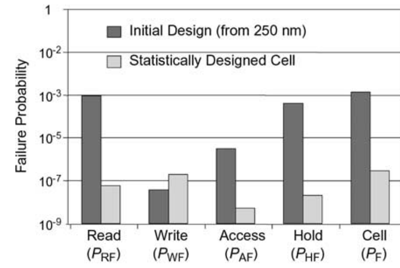

Fig. 25. Modification of different failure probabilities by statistical-design strategy.

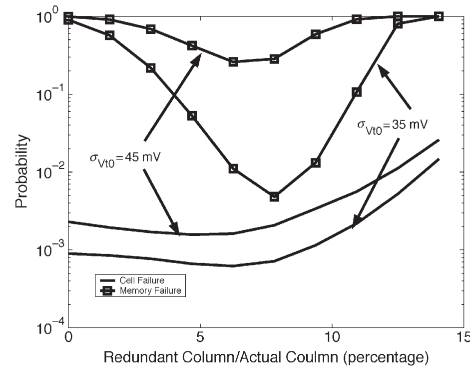

Fig. 26. Impact for number of redundant columns (N_{RC}) on memory yield under area constraint.

neglected the area required to implement the repair circuit that replaces the faulty columns with the redundant ones. The area required for the repair circuit also increases with an increase in the number of redundant columns. The effect of repair circuit area can be considered in the design flow presented in Fig. 23 by modifying the estimation of maximum area available for the cells in each iteration [i.e., $A_{\mathrm{cellmax}}(i)$] as

$$A_{\mathrm{cellmax}}(i) = \frac{A_{\mathrm{MAX}} - A_{\mathrm{repair}}}{N_{\mathrm{ROW}}(N_{\mathrm{COL}} + N_{\mathrm{RC}})} \quad (46)$$

where A_{repair} increases with an increase in N_{RC} (i.e $A_{\mathrm{repair}} \propto N_{\mathrm{RC}}$.

It should be noted that the consideration of the repair circuit area will further reduce the optimum number of redundant columns. This is due to the fact that the repair circuit area increases with an increase in the number of redundant columns, thereby reducing the area available for the cells (i.e., cell-failure probability increases). It should be noted that in this work, we have considered only column (or row) redundancy. However, the combined row and column redundancy scheme is also used for yield improvements [20]–[25]. Although the tradeoff between the actual and the redundant area will still be valid for the combined redundancy scheme, the exact analysis of the combined scheme is more complex [20]–[25]. In the Appendix, we presented a simplified method to incorporate the combined redundancy scheme in the proposed design flow. The principal modification required is in the method by which the memory-failure probability (P_{MEM}) is estimated from the column (P_{COL}) and row (P_{ROW}) failure probabilities [i.e., (25) in case of either row or column redundancy].

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

VII. Conclusion

In this work, we have analyzed different failure mechanisms in an SRAM cell, namely read, write, access, and hold failures, due to intra-die variation in the transistor threshold voltage. We have developed semianalytical models to estimate the probabilities of different failure events. The derived models can include the correlations of threshold voltages of transistors in the SRAM cell. The cell-failure probability is estimated using the probability of failure of individual events. Using the cell-failure probability, we have developed a set of models to estimate the probability of failure of an SRAM array. The derived memory failure probability model considers the architecture of the array and the use of redundancy. The cell and memory-failure probability models have been used to predict the yield of memory at an early stage of a design. Using the proposed models, we have shown that to predict the performance of an SRAM cell under parametric variations, a failure probability-based analysis is necessary. The proposed models are used for the statistical design and optimization of memory, which is necessary for maximizing yield in nanometer regimes. The developed design approach simultaneously optimizes the transistor sizes and the use of redundancy, to enhance the memory yield. It has been observed that under large parametric variation, increasing the number of redundant elements (at the cost of cell area) may not improve the design yield. The proposed statistical-modeling and design approach provide an integrated circuit and architecture level-design strategy for yield enhancement in nanoscale SRAMs.

Appendix
Estimation of Memory-Failure Probability Considering Combined Row and Column Redundancy

To improve the yield of memory array using redundancy, the use of combined row and column redundancy has been proposed [20]–[25]. Fig. 27 shows the schematic of the memory array considering the combined-redundancy scheme. An exact estimation of the memory-failure probability considering combined redundancy is complex and requires knowledge of the information of fault location [20]–[25]. This is because of the fact that, whether a memory chip with a certain number of faults can be repaired by using redundancy depends on the location of the faulty cells [20]–[25]. In [21]–[24], authors have described different methods for repairing memory array using combined redundancy based on the fault-location information. However, at the design phase the information on fault location is not available and hence cannot be used to estimate memory-failure probability. In [25], authors have proposed a method for the evaluation of memory failure by enumerating the different fixable failure events (i.e., fault maps that can be fixed using combined redundancy), which can be integrated into the proposed estimation models and design strategy. However, it will increase the complexity of estimation of P_{MEM}. Hence, we propose to use simple lower and upper bounds on P_{MEM} (assuming certain fault locations) for the initial estimations of P_{MEM} in the design step presented in Fig. 23.

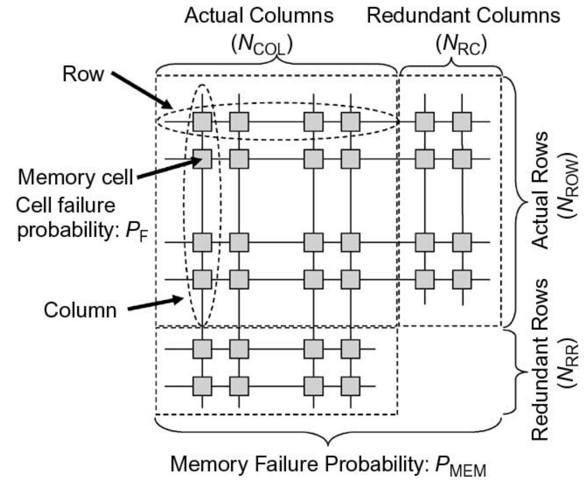

Fig. 27. Combined row and column redundancy and memory-failure probability.

The worst case fault map for the combined-redundancy scheme occurs when none of the faulty cells share any row or column with any other faulty cells (i.e., all the faulty cells are orthogonal [24]). Under this condition, either one redundant column or one redundant row is required to replace one faulty cell. Hence, the upper bound on the memory-failure probability [P_{MEM}(upper)] is given by

$$P_{\mathrm{MEM}}(\text{upper}) = \sum_{i=N_{\mathrm{RR}}+N_{\mathrm{RC}}+1}^{N_{\mathrm{ORTH}}} \binom{N_{\mathrm{ORTH}}}{i} P_{\mathrm{F}}^{i}(1-P_{\mathrm{F}})^{N_{\mathrm{ORTH}}-i} \tag{A.1}$$

where N_{ORTH} $(= \min\{N_{\mathrm{COL}}+N_{\mathrm{RC}}, N_{\mathrm{ROW}}+N_{\mathrm{RR}}\})$ is the total number of orthogonal cells in the array.

On the other hand, the best-case fault map occurs if the faulty cells tend to be concentrated either in a column or in a row. This requires a smaller number of redundant columns (or rows) to correct a large number of faulty cells. A column (or row) definitely needs to be replaced if the number of faulty cells in that column (or row) is more than the total number of redundant row (or column). We define such a column (row) as a must-replace column (row) [23], [24]. The probability that a column (P_{COLMR}) or row (P_{ROWMR}) is a must-replace one is given by

$$\begin{aligned} P_{\mathrm{COLMR}} &= \sum_{i=N_{\mathrm{RR}}+1}^{N_{\mathrm{ROW}}+N_{\mathrm{RR}}} \binom{N_{\mathrm{ROW}}+N_{\mathrm{RR}}}{i} \\ &\quad \times P_{\mathrm{F}}^{i}(1-P_{\mathrm{F}})^{N_{\mathrm{ROW}}+N_{\mathrm{RR}}-i} \\ P_{\mathrm{ROWMR}} &= \sum_{i=N_{\mathrm{RC}}+1}^{N_{\mathrm{COL}}+N_{\mathrm{RC}}} \binom{N_{\mathrm{COL}}+N_{\mathrm{RC}}}{i} \\ &\quad \times P_{\mathrm{F}}^{i}(1-P_{\mathrm{F}})^{N_{\mathrm{COL}}+N_{\mathrm{RC}}-i}. \end{aligned} \tag{A.2}$$

A memory with only must-replace columns or rows is a faulty one if the number of must-replace columns (N_{MRCOL}) is greater than the number of redundant columns or the number of must-replace rows (N_{MRROW}) is greater than the number of

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

redundant rows. This describes the fault map that results in the lower bound on $P_{\rm MEM}$ and given by

$$\begin{aligned} P_{\rm MEM}(\text{lower}) =& P\left[(N_{\rm MRCOL} > N_{\rm RC}) \right. \\ & \left. \cup (N_{\rm MRROW} > N_{\rm RR})\right] \\ \geq & \max\left\{P(N_{\rm MRCOL} > N_{\rm RC}), \right. \\ & \left. P(N_{\rm MRROW} > N_{\rm RR})\right\}. \quad \text{(A.3)} \end{aligned}$$

The individual failure probabilities in the above equation are given by

$$\begin{aligned} & P(N_{\rm MRCOL} > N_{\rm RC}) \\ & = \sum_{i=N_{\rm RC}+1}^{N_{\rm COL}+N_{\rm RC}} \binom{N_{\rm COL}+N_{\rm RC}}{i} \\ & \quad \times P_{\rm MRCOL}^{i}(1-P_{\rm MRCOL})^{N_{\rm COL}+N_{\rm RC}-i} \\ & P(N_{\rm MRROW} > N_{\rm RR}) \\ & = \sum_{i=N_{\rm RR}+1}^{N_{\rm ROW}+N_{\rm RR}} \binom{N_{\rm ROW}+N_{\rm RR}}{i} \\ & \quad \times P_{\rm MRROW}^{i}(1-P_{\rm MRROW})^{N_{\rm ROW}+N_{\rm RR}-i}. \quad \text{(A.4)} \end{aligned}$$

Hence, the overall memory failure probability is bounded as

$$P_{\rm MEM}(\text{lower}) < P_{\rm MEM} < P_{\rm MEM}(\text{upper}). \quad \text{(A.5)}$$

The design flow presented in Fig. 23 can be modified to include the combined-redundancy scheme by minimizing $P_{\rm MEM}$(upper) and $P_{\rm MEM}$(lower) by a proper choice of $N_{\rm RR}$ and $N_{\rm RC}$. A simple design flow can assume $N_{\rm RR} = N_{\rm RC}$.

REFERENCES

[1] S. R. Nassif, "Modeling and analysis of manufacturing variations," in *Proc. Custom Integrated Circuit Conf.*, San Diego, CA, 2001, pp. 223–228.
[2] C. Visweswariah, "Death, taxes and failing chips," in *Proc. Design Automation Conf.*, Anaheim, CA, 2003, pp. 343–347.
[3] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variation and impact on circuits and microarchitecture," in *Proc. Design Automation Conf.*, Anaheim, CA, 2003, pp. 338–342.
[4] A. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.
[5] X. Tang, V. De, and J. D. Meindl, "Intrinsic MOSFET parameter fluctuations due to random dopant placement," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 5, no. 4, pp. 369–376, Dec. 1997.
[6] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998.
[7] D. Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits," in *Symp. VLSI Technology*, Honolulu, HI, Jun. 1994, pp. 15–16.
[8] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling and estimation of failure probability due to parameter variation in nano-scale SRAMs for yield enhancement," in *Dig. Tech. Papers VLSI Circuit Symp.*, Honolulu, HI, Jun. 2004, pp. 64–67.
[9] ——, "Statistical design and optimization of SRAM cell for yield enhancement," in *Proc. Int. Conf. Computer Aided Design*, San Jose, CA, Nov. 2004, pp. 10–13.
[10] S. Mukhopadhyay, A. Raychowdhury, and K. Roy, "Accurate estimation of total leakage current in scaled CMOS logic circuits based on compact current modeling," in *Design Automation Conf.*, Anaheim, CA, Jun. 2003, pp. 169–174.
[11] D. A. Antoniadis, I. J. Djomehri, K. M Jackson, and S. Miller, "*Well-Tempered*," Bulk-Si NMOSFET Device Home Page. Cambridge, MA: Microsystems Technologies Laboratories, Massachusetts Institute of Technology [Online]. Available: http://www-mtl.mit.edu/Well/
[12] MEDICI: 2-D Device Simulation Program. Mountain View, CA: Synopsys Inc.
[13] A. Papoulis, *Probability, Random Variables and Stochastic Process*. New York: MacGraw-Hill, 2002.
[14] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. Piscataway, NJ: IEEE Press, 2001.
[15] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Int. Symp. Quality Electronic Design*, San Jose, CA, Mar. 2004, pp. 55–60.
[16] R. W. Mann *et al.*, "Ultralow-power SRAM technology," *IBM J. Res. Develop.*, vol. 47, no. 5/6, pp. 553–566, Sep. 2003.
[17] B. W. Wah and Y.-X. Chen, "Constrained genetic algorithms and their applications in nonlinear constrained optimization," in *IEEE Int. Conf. Tools Artificial Intelligence*, Vancouver, BC, Canada, Nov. 2000, pp. 286–293.
[18] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*. New York: Wiley, 2001.
[19] A. Agarwal, H. Li, and K. Roy, "A single-$V_{\rm t}$ low-leakage gated-ground cache for deep submicron," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, Feb. 2003.
[20] A. Chen, "Redundancy in LSI memory array," *IEEE J. Solid-State Circuits*, vol. 4, no. 5, pp. 291–293, Oct. 1969.
[21] S. E. Schuster, "Multiple word/bit line redundancy for semiconductor memories," *IEEE J. Solid-State Circuits*, vol. 13, no. 5, pp. 698–703, Oct. 1978.
[22] J. R. Day, "A fault-driven, comprehensive redundancy algorithm," *IEEE Des. Test Comput.*, vol. 2, no. 2, pp. 35–44, Jun. 1985.
[23] W.-K. Huang, Y.-N. Shen, and F. Lombardi, "New approaches for the repair of memories with redundancy by row/column deletion for yield enhancement," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 9, no. 3, pp. 323–328, Mar. 1990.
[24] C.-T. Huang, C.-F. Wu, J.-F. Li, and C.-W. Wu, "Built-in redundancy analysis for memory yield improvement," *IEEE Trans. Reliab.*, vol. 52, no. 4, pp. 386–399, Dec. 2003.
[25] C. H. Stapper, A. N. McLaren, and M. Dreckmann, "Yield model for productivity optimization of VLSI memory chips with redundancy and partially good product," *IBM J. Res. Develop.*, vol. 24, no. 3, pp. 398–409, 1980.

Saibal Mukhopadhyay (S'99) was born in Calcutta, India. He received the B.E. degree in electronics and telecommunication electrical engineering from Jadavpur University, Calcutta, India, in 2000. He is working toward the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.

He was an Intern in the High-Performance Circuit-Design Department, IBM T. J. Watson Research Laboratories, Yorktown Heights, NY, during the summer of 2003 and 2004. His research interests include analysis and design of low-power and robust circuits using nanoscaled CMOS and circuit design using double-gate transistors.

Mr. Mukhopadhyay received the IBM Ph.D. Fellowship award for 2004–2005. He received the "Best Paper Award" at the 2004 International Conference on Computer Design.

Hamid Mahmoodi (S'00) received the B.S. (Hons.) degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 1998, and the M.S. degree in electrical and computer engineering from the University of Tehran, Iran, in 2000. His M.S. research was on low power design of digital systems based on adiabatic switching principles. He is working toward the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.

His major research experiences and interests include low-power, robust, and high-performance design in nanoscale bulk CMOS and SOI technologies, nanoelectronic devices and architectures, design for yield enhancement, and VLSI testing. He has more than 30 publications in journals and conferences.

Mr. Mahmoodi was a recipient of the 2004 ICCD Best Paper Award.

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.

Kaushik Roy (S'83–M'83–SM'95–F'02) received the B.Tech. degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India, and the Ph.D. degree from the Electrical and Computer Engineering Department of the University of Illinois, Urbana-Champaign, in 1990.

He was with the Semiconductor Process and Design Center of Texas Instruments, Dallas, TX, where he worked on FPGA architecture development and low-power circuit design. He joined the electrical and computer engineering faculty at Purdue University, West Lafayette, IN, in 1993, where he is currently a Professor and University Faculty Scholar. His research interests include VLSI design/CAD for nanoscale silicon and nonsilicon technologies, low-power electronics for portable computing and wireless communications, VLSI testing and verification, and reconfigurable computing. He has published more than 300 papers in refereed journals and conferences, holds eight patents, and is a coauthor of the books *Low Power CMOS VLSI Circuit Design* (New York: Wiley, 2000) and *Low Voltage, Low Power VLSI Subsystems* (New York: McGraw-Hill, 2005).

Dr. Roy received the National Science Foundation Career Development Award in 1995, the IBM Faculty Partnership Award, the ATT/Lucent Foundation Award, and the Best Paper Awards at the 1997 International Test Conference, the IEEE 2000 International Symposium on Quality of IC Design, the 2003 IEEE Latin American Test Workshop, and the 2003 IEEE Nano. He is the Chief Technical Advisor of Zenasis, Inc., Campbell, CA, and Research Visionary Board Member of Motorola Laboratories, Schaumburg, IL (2002). He has been a Member of the Editorial Board of *IEEE Design and Test*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS. He was Guest Editor for the Special Issue on Low-Power Very Large Scale Integration in *IEEE Design and Test* (1994) and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS (June 2000), IEEE PROCEEDINGS—COMPUTERS AND DIGITAL TECHNIQUES (July 2002).

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 17:52 from IEEE Xplore. Restrictions apply.