Process Variability Considerations in the Design of an eSRAM

Michael Yap San Min, Philippe Maurine, Michel Robert, Magali Bastian Hage-Hassan

▶ **To cite this version:**

Michael Yap San Min, Philippe Maurine, Michel Robert, Magali Bastian Hage-Hassan. Process Variability Considerations in the Design of an eSRAM. MTDT 2007 - IEEE International Workshop on Memory Technology, Design and Testing, Dec 2007, Taipei, Taiwan. pp.23-26, ⟨10.1109/MTDT.2007.4547609⟩. ⟨lirmm-00275258⟩

HAL Id: lirmm-00275258

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00275258v1

Submitted on 27 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Process Variability Considerations in the Design of an eSRAM

M. Yap San Min[1,2], P. Maurine[1], M. Robert[1]
LIRMM[1]
Montpellier, France

M. Bastian[2]
INFINEON TECHNOLOGIES[2]
Sophia Antipolis, France

Abstract—**Process variation constitutes a serious hindrance to the performance of SRAMs, since memories require bigger design margins for their proper operations. In this paper, we propose a new dummy bit line driver structure and its statistical sizing method to reduce the sensitivity of the memory with respect to process variations, while improving the read timing margin. The dummy bit line driver is an essential component in a self-timed memory during a read operation. It triggers the sense amplifier at the appropriate time when bit line is discharged. We considered a 256kb SRAM in a 90nm technology node.**

I. INTRODUCTION

SRAMs have become a critical component of System on Chips as more than 80% of the surface of the chip is devoted to memories, and this ratio is expected to increase in the coming years [1]. Hence, it can be clearly seen that the overall performances of the chips are closely related to those of the SRAMs. Moreover, the relentless scaling of MOS technology has been responsible for the emergence of process variations at different levels during the manufacturing steps. These variations, whether local or global, have become a serious bottleneck for the proper design of embedded memories in the sub nanometre regime. Some examples of global variations arise due to non uniform mechanical polishing [2], lens aberrations [3] and non uniform temperature [4]. On the other hand, some examples of local variations include gate depletion [5], surface state charge [6] and random dopant fluctuations [7]. These variations are therefore responsible for the increase in the delay of the memory, which requires bigger design margins across process corners to ensure its correct functionality. However, the use of bigger design margins has the disadvantage of increasing power consumption and degrading the speed performance of the circuit [8]. In this paper, a new self timing circuit which is less sensitive to process variations is proposed, compared to a classic one [9]. A statistical sizing methodology of this structure has been developed to improve the read timing margin, while ensuring a high timing yield. The paper is organized in the following way. Section II introduces an approach of computing the required timing margin without being overly pessimistic and the probability of fulfilling this constraint. Section III shows the architecture of the new structure, dubbed dummy bit line driver (DBD). This new structure displays timing performances less sensitive to process variations compared to the original structure. Section IV describes the statistical sizing method of this new DBD used in the critical path of the memory. Section V compares the performances between the proposed and the reference structures.

II. MODELLING APPROACH

Generally, a circuit is characterized across its extreme process corners (worst and best case conditions) to ensure that its timing performances are met under all intermediate conditions. For example, verifications of set up times of data associated with clock frequencies are performed whenever the propagation delays are the greatest (worst case timing corner).

Indeed, we define the worst case delay by considering that the principal parameters of transistors p_i possess values at $\pm m_i.p_i$ ($m_i \in N$) around their mean values μ_{pi}, depending upon their impacts on the propagation delays (σ_{pi} represents the standard deviation of the statistical distribution of parameter p_i). This approach, through an appropriate choice of m_i values, allows the definition of the worst case delay at $n \cdot \sigma_D$, with σ_D being the standard deviation of the distribution delay. Nonetheless, the worst case approach has the drawback of considering that the statistical correlation ρ between all transistors is equal to 1. In other words, it simply means disregarding the local variations between supposedly identical transistors within the same die. This assumption leads to optimistic and pessimistic conclusions involved in worst and best case methods. To overcome the weakness of corner analysis, the most pragmatic solution consists in using a design timing margin $M_T>0$ during worst case design. However, a question arises: What is the read timing margin required in order not to be overly pessimistic? In this section, we will introduce a way of computing this read timing margin.

Consider the following signal racing conditions depicted in Fig. 1 between signals A issued from the control block used in discharging bit line (BL) and signal B, issued from same control block, involved in firing the sense amplifier during the discharged process of BL. Let us also assume that the signal A

should arrive at most 0 ps after signal B for a proper read operation of a selected SRAM cell (denoted by CC in Fig. 1).

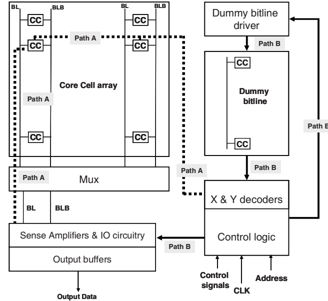

Figure 1. Signal races between paths A and B in the SRAM

In accordance with Fig. 2, let μ_A, μ_B and σ_A, σ_B be the mean values and the standard deviations of the propagation delay distributions of signals A and B. Let μ_D and σ_D represent the mean and the standard deviation values of the path delay difference D between A and B.

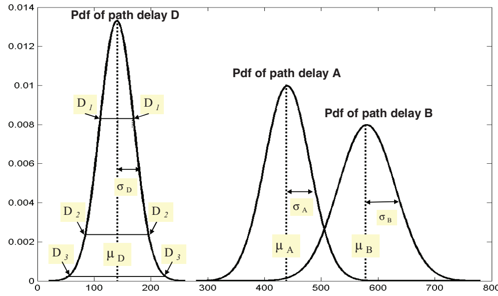

Figure 2. Notations

Let us now evaluate the probability of fulfilling a timing constraint. Assuming that all distributions are normal, the mean value and the standard deviation of distribution D are given by:

$$\begin{aligned} \mu_D &= \mu_B - \mu_A \\ \sigma_D &= \sqrt{\sigma_A^2 + \sigma_B^2 - 2 \cdot \sigma_A \cdot \sigma_B \cdot \rho} \end{aligned} \quad (1)$$

Using the Galton approximation, with the hypothesis that $\mu_D>0$, the probability P^V of satisfying a timing constraint for all values of ρ is computed as follows:

$$P^V = \frac{1}{2} \cdot \left\{ 1 + \sqrt{1 - \exp\left(- \frac{2 \cdot \mu_D^2}{\pi \cdot \sigma_D^2} \right)} \right\} \quad (2)$$

The sensitivities of delays to process variations $V_A=\sigma_A/\mu_A$ and $V_B=\sigma_B/\mu_B$ are known and found to be relatively constant over a wide range values of μ_A and μ_B (±20%). Hence, the value of path delay μ_B and subsequently that of the required read timing margin μ_D^{Yield} (appendix A.1) can be computed as follows to meet a timing yield value defined at $n \cdot \sigma_D$:

$$\mu_D^{Yield} = \frac{-a}{b} \cdot \left\{ \sqrt{1 - \frac{b \cdot c}{a^2}} + 1 \right\} - \mu_A \quad (3)$$

With $a = n^2 \cdot V_B \cdot \sigma_A \cdot \rho - \mu_A \quad b = 1 - n^2 \cdot V_B^2 \quad c = \mu_A^2 - n^2 \cdot \sigma_A^2$

III. Dummy Bit Line Driver with Reduced Variance

In a more specific context, involved in the design of advanced technologies, the corner method seems no longer enough to satisfy the timing constraints without the use of an increasing timing margin, caused by an increase in local variations. This fact brings up a question: Is it possible to maintain, or even reduce the design timing margins through proper design? Hence, we have defined a dummy bit line driver (DBD) having a smaller variance value V_B of path delay B compared to a more classic structure (Fig. 3b)

The DBD is an essential component of self-timed SRAMs. In the absence of an internal clock signal, the DBD coupled with the dummy bit line acts as a metronome to fire the sense amplifier at the appropriate time during a read operation (ΔVDD between BL and BLB is around 10% of VDD in Fig. 1). The proposed and the reference structures are illustrated in Fig. 3a and 3b. The topology of the proposed DBD has been realized in such a way that the discharge characteristics of dummy bit line being discharged by the DBD match those of bit line operated by a 6T SRAM cell (Fig. 3c). In Fig. 3a, transistors PD and PG of proposed DBD have similar functionalities as transistors PDcci and PGcci (i=1, 2) of the SRAM cell. Moreover logic gates g1 and g2 will mimic the signal WEN which controls pass gate PGcci, when WLSDUM is activated. The transistor P1 is used for precharging dummy bit line before any read operation, while transistor N1 sets node Z to 0 V at the beginning of a read cycle operation. We can see that the main difference between the 2 structures is that the reference structure makes use of stacked transistors.

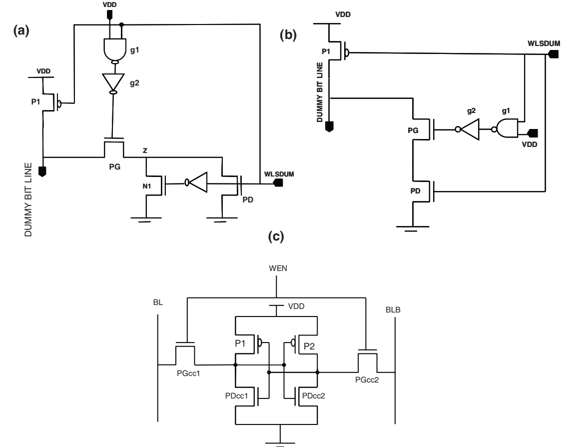

Figure 3. (a) Proposed DBD (b) Reference DBD (c) 6T SRAM Cell

IV. Statistical Sizing Procedure

In order to perform comparisons between the reference and proposed DBDs under constant timing yield, we have developed a statistical sizing methodology. Its main advantage lies in its independency with respect to process corners. In fact, corner methods do not consider local variations and provide no means of determining precisely voltage and temperature conditions under which sizing should be

performed. The sizing procedure of the structure is carried out at a typical process and is composed of 3 basic steps and 2 verification steps.

Step 1 (identification of most critical $(V, T°)_{Crit}$ condition): Starting from an initial solution, the first step involves in identifying the voltage and temperature (V, T°) conditions having the poorest timing yield. In fact, under signal races conditions, the (V, T°) condition leading to the smallest timing yield strongly depends on the delay sensitivities of both paths to temperature and supply voltage. To identify the critical conditions, transient simulations of the timing performances of critical paths A and B in the memory are done under different temperature and voltage conditions covering the whole range to obtain μ_A and μ_B. The most critical conditions correspond to the highest numerical value of the following expression (appendix A.2):

$$\frac{\mu_A \cdot \mu_B}{(\mu_B - \mu_A)^2} \quad (4)$$

Step 2 (variability estimation): The second step requires the estimation of the variability of paths A and B involved in the signal races. To do so, Monte Carlo simulations of the critical path are performed at the critical condition $(V, T°)_{Crit}$ found in step 1. Once these statistical simulations are performed and the values of μ_A, μ_B, σ_A, σ_B and ρ are obtained, the value of the required timing margin μ_D^{Yield} corresponding to a timing yield is computed using (3).

Step 3 (sizing for a given timing yield): The third step consists in sizing the DBD at a typical process and under the voltage and temperature conditions obtained in step 1, $(V, T°)_{Crit}$ in order to obtain the computed μ_D^{Yield}.

Step 4 (first verification step of the timing yield): Once the above sizing procedure is over, the first verification step consists in performing Monte Carlo simulations on the critical path at $(V,T°)_{Crit}$ to obtain μ_A, μ_B, σ_A, σ_B and ρ values. The timing yield is then evaluated using (2). If the computed value fulfills the predefined constraint, we proceed with the second verification step. Otherwise, we reiterate step 3 with the new values of μ_A, μ_B, σ_A, σ_B, ρ and the newly calculated μ_D^{Yield}.

Step 5 (second verification step of the timing yield): It implies verifying that the constraint of the timing yield satisfies all temperature and supply voltage conditions. This is done through Monte Carlo simulations in order to estimate the values of μ_A, μ_B, σ_A, σ_B and ρ for different values of V and T°. Once the statistical simulation has been done, the timing yield is processed. If the values obtained for the various (V, T°) couples are greater than the predefined constraint at $(V, T°)_{Crit}$, the verification step is over. However, if the constraint is not satisfied, step 1 should be repeated with the new sizing obtained.

V. PERFORMANCE RESULTS

The reference and the proposed DBDs have been placed in the critical path of a 256kb SRAM. Both structures have been sized using the method introduced in section IV. The timing yield (predefined constraint) has been set at 99.87% i.e. n = 3σ and the correlation ρ considered was 0.9. The mean values (μ_A and μ_B) and the standard deviations (σ_A and σ_B) of the characteristic delays of the signal races shown in Fig. 1 have been simulated (Monte Carlo: 2000 runs) for both structures. The model card, used in Hspice simulations, is the bsim4.3.0 model which takes into account local and global variations. The results obtained were used to compute the reduction in the delay variance (ΔV_B) of path B between the proposed (prop) and reference (ref) DBDs in table I. In table II, the probability P^V (2) of meeting the predefined timing constraint, the read timing margin μ_D (1) and subsequently the reduction in read timing margin $\Delta\mu_D$ between the proposed and reference structures have been calculated. Tables I and II report the results obtained.

TABLE I. Variability reductions

VDD (V)	T (°C)	μ_A (ps)	V_A (%)	V_B Ref (%)	V_B Prop (%)	$\Delta V_B/V_B$ Ref (%)
1	-40	1919	9	12	10	17
1.08		1534	7	9	7	22
1.26		1061	6	7	5	29
1.32		963	5	6	5	17
1	0	1994	9	11	9	18
1.08		1609	7	9	7	22
1.26		1123	6	7	5	29
1.32		1021	5	6	5	17
1	40	2065	8	10	8	30
1.08		1680	7	9	7	25
1.26		1183	5	7	5	17
1.32		1078	5	6	5	17
1	80	2134	8	10	7	30
1.08		1750	7	8	6	25
1.26		1244	5	6	5	17
1.32		1136	5	6	5	17
1	125	2209	8	10	7	30
1.08		1828	7	8	6	25
1.26		1313	5	6	5	17
1.32		1201	5	6	5	17

In the results presented in table I, the reduction in variability ($\Delta V_B/V_{Bref}$) is significant and lies between 17% and 30%. This reduction has been achieved by using pass gate transistor PG and pull down transistor PD in the proposed DBD which is 2 times the size of the PG and PD used in the reference DBD.

In table II, we can see that as expected, the calculated probabilities of fulfilling the timing constraints are very close to the required value corresponding to 99.87% (3σ). Simultaneously, we observe a reduction in the read timing margin $\Delta\mu_D$, normalized with respect to μ_A, lying between 2.4% and 6.4%. This maximum decrease of 6.4% corresponds to a 4% decrease in the read cycle time of the memory.

VI. CONCLUSION

In this work, we have proposed an easy way of computing the required read timing margin in the design process of an

SRAM, without introducing an excessive read margin. Moreover, we have introduced a new dummy bit line driver structure which is more robust to process variations and a statistical sizing method which enables the comparison between a reference and the proposed structure at constant timing yield. We have shown that the use of the proposed dummy bit line driver and its statistical sizing method improves the reduction in the design timing margins, while guaranteeing a high predefined timing yield.

TABLE II. Reduction of the read timing margin μ_D normalized with respect to path delay A

VDD (V)	T (°C)	μ_A (ps)	$\mu_{D\ Ref}$ (ps)	P^V Ref (%)	μ_D Prop (ps)	P^V Prop (%)	$\Delta\mu_D/\mu_A$ (%)
1	-40	1919	494	99.992	381	99.997	5.9
	27	2041	420	99.987	314	99.997	5.2
	125	2209	302	99.813	241	99.996	2.7
1.08	-40	1534	343	99.999	245	99.999	6.4
	27	1657	336	99.999	244	99.999	5.6
	125	1828	278	99.991	232	99.999	2.6
1.2	-40	1181	250	99.999	175	100.00	6.3
	27	1290	283	99.999	210	100.00	5.7
	125	1448	269	99.999	233	100.00	2.4
1.32	-40	963	216	99.999	158	100.00	6.1
	27	1060	262	100.00	202	100.00	5.7
	125	1201	268	100.00	239	100.00	2.4

APPENDIX

A.1) Estimation of design margin μ_D at $n.\sigma_D$

Suppose that we want to have a read margin μ_D at $n.\sigma_D$, such that:

$$\mu_D - n.\sigma_D = 0 \tag{A.1}$$

As we have seen previously in section II, μ_D and σ_D can be defined using (1). Expression (A.1) can therefore be represented by the following equation:

$$(\mu_B - \mu_A) - n\sqrt{\sigma_A^2 + \sigma_B^2 - 2.\sigma_A.\sigma_B.\rho} = 0 \tag{A.2}$$

Let $V_A = \frac{\sigma_A}{\mu_A}$ and $V_B = \frac{\sigma_B}{\mu_B}$ (A.3)

Hence the value of μ_B in (A.2) is given by:

$$\mu_B = -\frac{a}{b}\left\{\sqrt{1-\frac{b.c}{a^2}} \pm 1\right\} \tag{A.4}$$

with $a = n^2 \cdot V_B \cdot \sigma_A \cdot \rho - \mu_A$ $b = 1 - n^2 \cdot V_B^2$ $c = \mu_A^2 - n^2 \cdot \sigma_A^2$

As the delay of signal B should be greater than that of signal A, delay μ_B in (A.4) becomes:

$$\mu_B = -\frac{a}{b}\left\{\sqrt{1-\frac{b.c}{a^2}} + 1\right\} \tag{A.5}$$

Hence, the required design margin μ_D^{Yield} in (1) is given by:

$$\mu_{D\ Yield} = -\frac{a}{b}\left\{\sqrt{1-\frac{b.c}{a^2}} + 1\right\} - \mu_A \tag{A.6}$$

A.2) Identification of critical condition $(V, T°)_{crit}$

The probability P^V of fulfilling a timing constraint is given by (2). In fact, since $(V, T°)_{crit}$ represents the condition showing the highest probability of the occurrence of a timing constraint violation; P^V should be minimum at this condition. Thus, P^V is minimum if the expression:

$\frac{\mu_D^2}{\sigma_D^2}$ is also minimum. (A.7)

Let $\alpha = \frac{\sigma_A}{\mu_A} = \frac{\sigma_B}{\mu_B}$ (A.8)

By substituting both σ_D by (1) and σ_A and σ_B by (A.8) in (A.7), expression (A.7) can be represented by:

$$\frac{\mu_D^2}{\alpha\sqrt{1+\frac{2.\mu_A.\mu_B}{(\mu_B-\mu_A)^2}(1-\rho)}} \tag{A.9}$$

It can be clearly seen that expression (A.9) is minimum if the expression:

$\frac{\mu_A.\mu_B}{(\mu_B-\mu_A)^2}$ has the highest numerical value (A.10)

REFERENCES

[1] Y. Zorian, “Embedded memory test & repair: infrastructure IP for SoC yield,” IEEE Design and Test of Comuters, vol. 20, no. 3, pp. 58-66 (2003).

[2] D. O. Ouma, D. S. boning, J. E. Chung, W. G. Easter, V. Saxena, S. Misra, and A. Crevasse, “Characterization and modeling of oxide chemical-mechanical polishing using planarization length and pattern density concepts,” IEEE Transactions on Semiconductor Manufacturing, vol. 5, no. 2, pp. 232-244 (2002).

[3] T. A. Brunner, “Impact of lens aberrations on optical lithography,” IBM J. Res. Develop, vol. 41, no. 1/2, pp. 57-67 (1997).

[4] J. A. Croon, W. Hansen, and H. E. Maes, “Matching properties of deep sub micron mos transistors,” Springer International Series (2005).

[5] N. D. Arora, R. Rios, and C. L. Huang, “Modeling of polysilicon depletion effect and its impact on submicrometer cmos circuit Performance,” IEEE Transactions on Electron Devices, vol. 42, no. 5, pp. 935-943 (1995).

[6] J. Shyu, G. C. Temes, and F. Krummenacher, “Random error effects in matched mos capacitors and current sources,” IEEE J. Solid State Circuits, vol. 19, no. 6, pp. 948-956 (1984).

[7] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klassen, “Modeling statistical dopant fluctuations in mos transistors,” IEEE Transactions on Electron Devices, vol. 45, no. 9, pp. 1960-1971 (1998).

[8] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, and R. Mahnkopf, “The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits,” IEEE Transactions on VLSI, vol. 5, no. 4, pp. 360-368 (1997).

[9] B. S. Amrutur, M. A. Horowitz, “A replica technique for wordline and sense control in low power SRAMs’,” IEEE J. Solid State Circuits, Vol. 33, No. 8, pp. 1208-1219 (1998).