Statistical analysis of the impact of within die variations on eSRAM internal signal races

Michael Yap San Min

UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : Microélectronique
Formation Doctorale : Systèmes Automatiques et Microélectroniques
Ecole Doctorale : Information, Structure et Systèmes

Présentée et soutenue publiquement

par

Michael Yap San Min

Le 21 Janvier 2008

Analyse statistique de l'impact des variations locales sur les courses de signaux dans une mémoire SRAM embarquée

---

JURY

- Pr. Serge Pravossoudovitch , Président
- Pr. Michel Robert , Directeur de thèse
- Pr. Régis Leveugle , Rapporteur
- Dr. Jean Michel Portal , Rapporteur
- Dr. Philippe Maurine , Examinateur
- Mrs. Magali Bastian , Examinateur
- Dr. Christophe Chanussot , Examinateur

Acknowledgements

I would like to thank first of all my thesis supervisor Pr Michel Robert, director of the laboratory of computer science, microelectronic and robotic of Montpellier (LIRMM), for his guidance and advice upon my arrival at the laboratory and throughout my stay at the LIRMM.

I am greatly indebted to Dr Philippe Maurine, from University of Montpellier II, and want to express my deepest gratitude for his commitment and precious technical, as well as non technical, advice he has been giving me throughout those three years of my research work. His great enthusiasm and devotion to my work have been for me a serious source of motivation in the complete realization of this work.

This thesis has been the result of an industrial collaboration between Infineon Technologies France and the 'LIRMM' of University of Montpellier II. I want to thank Mr Jean Christophe Vial, the manager of the memory library team (LIB MEM), for having given me the opportunity to form part of the LIB MEM team and start my PhD at Infineon Technologies.

Special thanks to Mrs Magali Bastian and Mr Jean Patrice Coste, both from Infineon Technologies (LIB MEM), for their technical expertise in embedded memories and the enlightenment they have been giving me on SRAMs.

I would also like to thank Pr Régis Leveugle from INPG, Dr Jean Michel Portal from University of Provence, Dr Christophe Chanussot from Infineon Technologies France, and Pr Serge Pravossoudovitch from University of Montpellier II for serving on my thesis committee.

I am also grateful to Mr Jean Yves Larguier, who took me as a trainee at Infineon Technologies in 2004 and gave me the chance to carry on with a PhD.

Thank you also to some designers of LIB MEM who have been providing technical advice and all the PhD students from the microelectronic department of the 'LIRMM' for this great working atmosphere at the laboratory.

Finally, my warm thanks to my parents, my brother and my sister for all their support, encouragement and for believing in me.

Table of contents

Table of contents

List of figures

List of figures

List of tables

List of tables

Introduction générale

Introduction générale

Les systèmes sur puce trouvent leurs applications dans de nouveaux appareils nomades tels que les appareils photo numériques, smart phone, PDA et autres applications mobiles. Ces systèmes sur puce se composent donc d'une multitude de blocs IP, allant des processeurs embarqués à des mémoires embarquées comme les SRAMs, en passant par des encodeurs/décodeurs MPEG et bien d'autres composants. Face à la compétition du marché dans le secteur des semi conducteurs et le temps de mise sur le marché qui reste l'une des principales préoccupations des industriels, ceux-ci font donc plus souvent appel à l'utilisation de plusieurs blocs IP, particulièrement avec l'accroissement de la complexité des puces et de leur coût. Néanmoins, les performances globales et le rendement de fabrication des circuits dépendent en grande partie des performances de ces blocs mémoires, qui peuvent représenter jusqu'à 80% de la surface totale de la puce selon l'ITRS.

Parallèlement à l'accroissement de la part dévolue à la mémoire au sein des circuits, l'évolution technologique s'accompagne d'une augmentation de la variabilité des performances, notamment dues: (a) aux variations de process (P) qui apparaissent lors des étapes de fabrication, (b) aux variations statiques et dynamiques de la tension d'alimentation (V) et (c) aux variations de température (T) dues aux variations de l'activité au sein du circuit. Ces trois paramètres constituent la définition classique de 'PVT'.

De nombreux travaux sont actuellement dédiés à la définition de méthodes de conception statistiques permettant d'anticiper l'impact sur les performances temporelles des variations des procédés de fabrication. Ces variations des procédés de fabrication apparaissent à diverses étapes de fabrication, et constituent un sérieux obstacle lors de la phase de conception des circuits intégrés en technologique fortement submicronique.

Les auteurs de ces travaux distinguent généralement deux catégories de variations des procédés de fabrication : les variations globales et les variations locales. Une variation des procédés de fabrication (P) est dite globale si celle-ci a des conséquences identiques à l'échelle d'un circuit. Inversement, une variation est dite locale si elle ne produit des effets que sur une partie limitée du circuit, et celle-ci peut être de type systématique ou stochastique. Les variations globales (inter-fab, inter-lot, inter-wafer et inter-die) ont de nombreuses origines: planarité des wafers de silicium, aberrations optiques, hétérogénéité de la

température lors de la fabrication … De manière identique, les variations locales (intra-die) peuvent avoir différentes origines comme par exemple la variation de la concentration des dopants, la finesse de la gravure, la variation de l'épaisseur d'oxyde de grille .... Globales, ou bien locales, ces variations affectent, avec la réduction des dimensions des transistors, de plus en plus significativement les performances des circuits intégrés comme la fréquence maximale de fonctionnement, la consommation statique ou encore le rendement de fabrication. Si les variations des procédés de fabrication affectent de plus en plus les performances temporelles des circuits intégrés, la tension d'alimentation et la température demeurent des sources importantes de variations des timings et de la consommation. En effet, les fluctuations de tension sont causées par les chutes de tension RI, l'hétérogénéité spatiale et temporelle de l'activité des blocs, et la non uniformité de la distribution de la tension d'alimentation. Ces chutes de tension, bien souvent localisées, entraînent l'apparition de points chauds et l'existence de gradients de température dans les circuits qui altèrent localement les performances.

En terme de conception, ces variations de procédés de fabrication et de conditions de fonctionnement sont généralement prises en compte en adoptant une approche pire et meilleur cas. Par exemple, l'estimation à priori de la fréquence maximale de fonctionnement est réalisée en effectuant deux analyses distinctes des performances temporelles : l'une en considérant les conditions PVT les plus favorables (best case timing corner) et l'autre en considérant les plus défavorables (worst case timing corner).

Dans ce contexte, l'accroissement de la variabilité des procédés de fabrication conduit à l'accroissement relatif de la fourchette d'estimation des performances, comme la fréquence de fonctionnement d'un circuit. Ceci peut poser des problèmes de convergence du flot de conception. A titre d'exemple, dans certains cas, l'écart entre les estimations meilleur et pire cas peut atteindre 60% (en 90nm) des performances moyennes ou typiques.

Si l'accroissement du degré de pessimisme, conjugué avec à l'utilisation de méthodes pire et meilleur cas, permet de prendre en compte, lors de la conception, l'impact de la variabilité dans de nombreux cas, ce n'est toutefois pas une approche suffisante pour anticiper tous les effets liés à l'accroissement des variations intra-die ou encore à l'apparition de points chauds ou de chutes locales de la tension d'alimentation. Ainsi la seule alternative, permettant de s'affranchir des méthodes pire et meilleur cas, réside dans l'adoption de techniques statistiques et notamment l'analyse statistique des performances temporelles. Cette analyse statistique des performances des SRAMs constitue le cœur de cette thèse.

Le premier chapitre introduit les généralités et les défis des mémoires embarquées, et plus particulièrement les défis liés aux SRAMs tels que la consommation, rendement, fiabilité. Les contraintes liées à la consommation de puissance, la basse puissance et la conception en vue de la manufacturabilité vont également être détaillées. Nous présentons aussi l'architecture de la SRAM, ainsi que la complexité de ses opérations de lecture et d'écriture. Cette complexité est souvent associée aux courses de signaux qui doivent êtres parfaitement synchronisées et ce malgré les nombreuses sources d'incertitudes existantes. Cette synchronisation est réalisée par le 'dummy bit line driver' qui a un rôle essentiel dans une mémoire SRAM embarquée. En effet, celui-ci joue lors des cycles de lecture notamment, le rôle de métronome de la mémoire. Il garantit que les amplificateurs de lecture sont déclenchés après que la différence de potentiel entre leurs entrées ait atteint un niveau suffisant pour que la lecture se fasse correctement.

Le chapitre deux fait un état de l'art des sources de variations de la mémoire causées par des dérives des procédés de fabrication (local ou global), des conditions environnementales (tension d'alimentation, température) et des conditions de vieillissement (NBTI, claquage d'oxyde de grille). L'impact de ces phénomènes de variabilité sur la performance de la SRAM y est analysé, en terme de défaillances paramétriques sur les performances du point mémoire et les divers blocs fonctionnels de la mémoire SRAM. De plus, les techniques les plus courantes pour palier à ces problèmes sont présentées. Parmi les méthodes présentées qui prennent en compte ces variations, nous retrouvons les techniques de pulse et la variation de la tension d'alimentation utilisées pour la conception du point mémoire en présence de la variabilité. Des méthodes telles que la redondance (de ligne, de colonne…) et les codes correcteurs d'erreurs y sont aussi introduits.

Dans le troisième chapitre, nous débutons par l'introduction de la méthode traditionnelle (méthode de corner) très couramment utilisée et qui est basée sur la détermination des conditions extrêmes de fonctionnement d'un circuit : l'une en considérant les conditions PVT les plus favorable et l'autre en considérant les plus défavorables. Nous démontrons les limitations de l'analyse de corner et de son incapacité à considérer les variations locales. Nous montrons que l'accroissement progressif des variations locales peut conduire des analyses de corner, effectuées sur des courses de signaux, à être optimistes, d'où la nécessité de développer des techniques de conception statistique. Afin de faire face à l'optimisme et au pessimisme de l'analyse de corner, nous proposons une modélisation permettant d'évaluer la marge temporelle de lecture requise sans être trop optimiste ou pessimiste dans notre

estimation, et permettant également d'évaluer la probabilité de satisfaire cette contrainte temporelle. Cette modélisation permet donc de prendre en des variations locales dans le calcul des marges temporelles de lecture dans la mémoire.

Les applications de la modélisation introduite au chapitre trois sont présentées dans le chapitre quatre. Les deux applications comprennent : (i) la définition à priori de cartographies du plan mémoire de probabilité d'occurrence de violations des contraintes temporelles, (ii) la mise au point d'une méthode de dimensionnement statistique d'un bloc particulier de la mémoire, appelé 'dummy bit line driver'. Cette structure est un élément essentiel des mémoires SRAMs auto synchronisées. Le 'dummy bit line driver', en l'absence de signal d'horloge interne à la SRAM, joue en effet le rôle de métronome en indiquant à l'amplificateur de lecture quand lire la donnée. Nous introduisons un nouveau 'dummy bit line driver' présentant une sensibilité réduite aux variations de procédés et des sensibilités à la tension d'alimentation. L'utilisation conjointe de la méthode de dimensionnement statistique et du 'dummy bit line driver' permettent de réduire significativement la variabilité des délais des chemins et les marges de conception, tout en garantissant un rendement temporel donné.

General Introduction

General introduction

System on Chip devices, have found their applications in every latest hand-held consumer devices, like smart phones, PDAs, digital cameras and other mobile applications. These SoCs embrace a variety of IP cores, such as embedded processors, MPEG encoders/decoders, DSPs, embedded memories which include SRAMs and more. As time to market has become the main obsession for every company wanting to remain very competitive on the market, the semiconductor companies need to license more and more IPs, particularly with an increase in the complexity of the chip and soaring design costs. However, the global performances and the fabrication yield of the chips are governed in majority by memory blocks, which account for a large percentage of the surface of the chip (around 80% according to the ITRS).

Simultaneously with the rapid increase in memory blocks within the chips, technology evolution is accompanied by an increase in performance variability owing to: (a) process variations (P) which appear due to manufacturing phenomena, (b) static and dynamic variations of the supply voltage (V) and (c) temperature (T) fluctuations due to varying activity levels within the circuit. Those three parameters constitute the classic 'PVT' definitions.

Currently, statistical design methods have been the focus of substantial research in order to anticipate the impact of manufacturing process variations on timing performances. These process variations appear at different levels during the manufacturing steps and have emerged as a serious bottleneck for the proper design of ICs in the sub nanometre regime. They are generally classified into two distinct groups of manufacturing processes, namely global and local variations. Global variations, caused by inter-fab, inter-lot, inter-wafer and inter-die processing variations, originate from several factors which include non uniform chemical polishing (CMP) which occurs due to different pattern densities, lens aberrations, non uniformity of the temperature and more. Global variations are said to affect every element on the chip equally or in a systematic way. On the other hand, local variations or currently known as mismatch result from random dopant fluctuations, line edge roughness, surface state charge, gate depletion and film thickness variation. Local variations are characterized by differences between supposedly identical structures found on the same die, but these variations display either a systematic or a random behaviour. In fact, transistor scaling has

exacerbated the impact of local and global variations, affecting performances of integrated circuits, for instance their maximum operation frequencies and static power consumptions or even the manufacturing yields.

Although manufacturing process variations influence more and more the timing performances of ICs, supply voltage and temperature fluctuations also constitute important sources of timing variations. Indeed, voltage variations are due to IR drop, switching activity of different areas of the chip and non uniform power supply distribution, whereas temperature variations stem from the existence of temperature gradient due to different switching activities in the chip. This condition gives rise to the appearance of hot spots. These hot spots cause important differences in temperature between different areas of the die.

To handle the impact of manufacturing process variations along with the operating conditions in circuit design, corner based methodology is performed by characterizing the circuit under best case and worst case conditions. For instance, the estimation of the maximum operating frequency is carried out at least under 2 distinct timing analysis conditions: one while considering the best PVT conditions (best case timing corner) and the other one by considering the most unfavourable conditions (worst case timing corner).

In this context, the increase of variability in manufacturing processes results in an underestimation of performances in the operating frequency of an integrated circuit. This can therefore impact on the convergence of the design flow. In certain cases, the differences in the estimation between best and worst cases can reach 60% of the mean performances.

If the increase of optimism and pessimism, introduced by best and worst cases, takes into account the effect linked to variability in circuit design under certain classic aspects, this approach is insufficient in foreseeing all the impacts due to an increase of intra-die variations or even with the appearance of hot spots or local IR drop of the power supply.

Thus, statistical analysis method is emerging as the solution to account for these sources of variations, and provides more accurate analysis results of the circuit in terms of timing analysis.

The first chapter introduces the generalities and the challenges of embedded SRAMs. In this chapter, challenges dealing with power consumption, low power and design for manufacturability issues are being discussed. A detailed explanation of the architecture of the SRAM is also provided, describing the functionalities of each of its blocks. We also highlight the complexities of SRAM operations, involved in read and write operations, due to signal

races in the memory. In fact, the memory needs to be perfectly synchronized in the presence of those variability conditions.

Chapter two presents the variability aspects encountered by the memory. The major sources of variations owing to manufacturing processes, environmental conditions and aging conditions are introduced herein. We explain the impact of these sources of variations on the memory performances and give some of the most common techniques, related to the memory cells and to the memory architecture, used to mitigate the effects of variability.

In chapter three, we demonstrate the limitations of the corner analysis method and its inability in capturing local variations. We illustrate that an increase in local variations leads to optimistic conclusions when corner analysis is undertaken during racing conditions in the memory, and the need for developing statistical design techniques. More precisely, to overcome the optimism and the pessimism caused by corner analysis, we provide a simple modelling approach for computing the appropriate read timing margin in the memory and the probability of fulfilling this timing constraint.

Chapter four displays some applications of the modelling approach introduced in the previous chapter. The two applications include: (i) displaying a failure probability map of the memory core which shows its most critical areas that are more likely to meet timing constraint violations during a read operation, (ii) developing a statistical sizing methodology of a particular block of the memory, dubbed dummy bit line driver. This structure plays an important role in an auto synchronized memory during the read operation, since it is responsible in triggering the sense amplifier at the appropriate time when a memory cell is being read. We also introduce a new dummy bit line driver having its timing performances more robust to process and voltage variations. The use of the statistical sizing methodology developed and the proposed dummy bit line driver demonstrate better performances in terms of the reductions in path delay variabilities and read timing margins, compared to the original dummy bit line driver.

Chapter 1

Generalities and challenges of eSRAM

Embedded memories have become increasingly important as they form the major component of SoCs. This chapter focuses on the generalities of SRAMs, their functionalities and the associated complexities involved in SRAM's operations. These complexities arise due to the presence of racing signals, which need to be correctly synchronized in the presence of variability phenomena. We also introduce herein some of the main challenges i.e. yield and reliability issues faced by SRAMs as the transistor is continuing to shrink relentlessly.

INTRODUCTION

The advent of system on chip devices has paved the way to its widespread applications in a myriad of domains, ranging from the automotive sectors to the communication industries, which include numerous latest hand-held consumer devices like smart phones, PDAs, digital cameras and other mobile applications. However, system on chip technology is setting designers the very challenging problem of adopting new techniques to get the SoC operating properly the first time in several embedded applications and in a minimum time to market, through shorter design cycles. This is mainly due to today's rapidly growing number of gates per chip reaching several millions, according to Moore's law which states that "the number transistors on a chip doubles about every two years". To bridge the gap between this fast technology's evolution and the lack of available manpower, designers make use of predefined modules to avoid reinventing the wheel with every new product. These blocks known as intellectual property cores (IP) or Virtual Components (VC), usually come from third parties or are sometimes designed in-house. Among the different existing IP blocks, these include DSPs, microprocessors, mixed signal blocks (ADC, DAC) and embedded memories as shown in figure I.0 below.

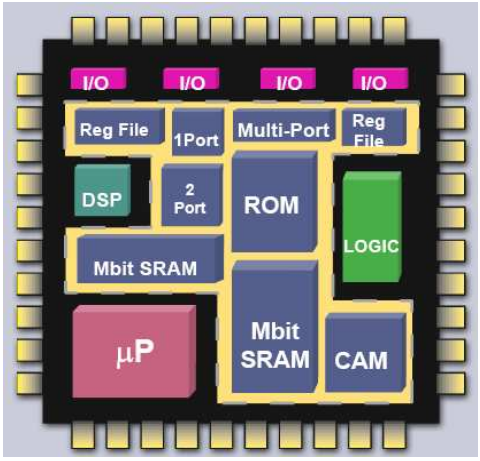

Fig. I.0 SoC block diagram

This chapter focuses on the generalities and challenges of eSRAMs. The first part describes existing types of memories. Next, the different challenges experienced by SRAM will be analysed, followed by a description of its architecture so as to understand its operating mode. Then, in the last section, we will see the complexity involved in the read and write operations

due to signal races and how this condition is being handled, through the introduction of a dummy bit line driver structure.

I.1 Classification of embedded memories

Embedded memories can be categorized as volatile and non volatile memories. This is illustrated figure I.1 that gives evidence of the large choice of memories available.

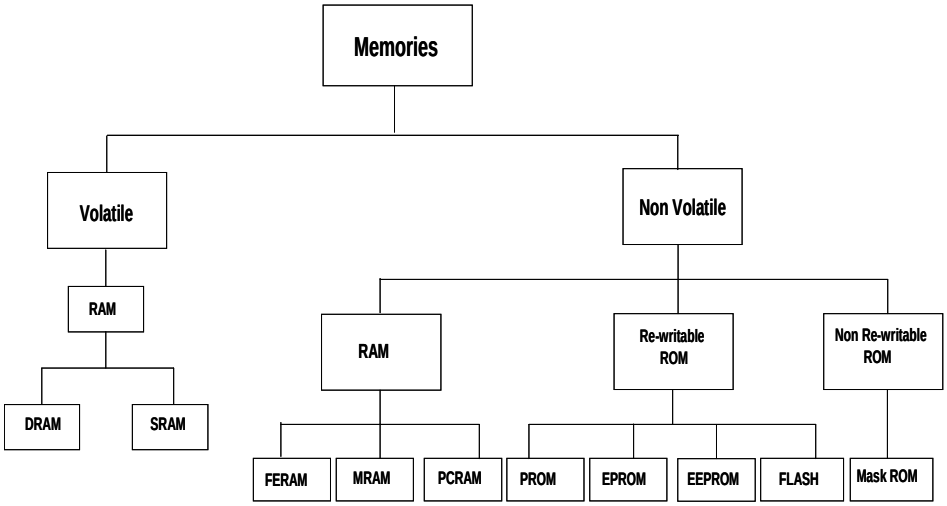

Fig. I.1 Types of memories

I.1.a Volatile Memory (VM)

The volatile memory, as it name implies, loses data when the supply voltage is switched off. It consists only of Random Access Memory (RAM) which can further be split into Dynamic RAM (DRAM) and Static RAM (SRAM). Both DRAM and SRAM allow read and write operations of the cell in the memory chip.

DRAM memory has the advantage of being cheaper and smaller in size than SRAM memory (1T cell instead of 6T for SRAM), offering a higher density. However, this higher density comes with a slower read/write operation owing to a capacitor which needs to be discharged/charged during these operations.

This does not constitute the only drawback of DRAM over SRAM. Indeed, due to leakage currents, the capacitor associated with a DRAM cell needs to be regularly refreshed to avoid loss of stored data. The refreshed logic needed therefore makes DRAM a more complex technology compared to SRAM. As a result, SRAM memory has the advantages of featuring

higher speed than DRAM (6T) and requires no refreshing operation; nonetheless it has a higher cost than DRAM.

I.1.b Non Volatile Memory (NVM)

Non volatile memories can be classified into three distinct categories:

(i) Non Volatile RAM
(ii) Re-writable ROM
(iii) Non re-writable ROM.

Examples of non volatile RAM include:

(i) Ferroelectric RAM (FeRAM) which uses a ferroelectric layer and possesses a similar architecture to DRAM,
(ii) Magnetic RAM (MRAM) which is composed of ferromagnetic materials for storing data
(iii) Phase Change RAM (PCRAM) making use of chalcogenide alloys for switching between crystallized and amorphous states. The PCRAM operation is based on the different resistively of the materials for storing data.

Re-writable ROM includes:

(i) Programmable ROM (PROM) which is a one time programmable memory performed by burning fuses in an irreversible process,
(ii) Erasable Programmable ROM (EPROM) which is programmed electrically and data are erased from the memory cells by using ultraviolet illumination,
(iii) Electrically Erasable PROM (EEPROM) in which the programmed and erased operations are both done electrically
(iv) Flash memory which stores data in a floating gate and is programmed and erased electrically.

Non re-writable ROM, only meant for read operation, consists for its part of Mask ROM in which data are written after the chip fabrication by the IC manufacturer. This is done by making use of a photo mask.

I.2 Challenges of SRAM memory design

In this part, we will detail the specific challenges encountered by circuit designers in designing the memory. Examples of theses challenges include power consumption, low power issues and Design for Manufacturability (DFM) aspects.

I.2.1 SRAM performances in 90nm and 65nm nodes

SRAM memories have become a critical component of SoC devices since they occupy around 80% of chip's surface as shown in figure I.2 below [Zor02]. Hence, the performance of the SoC depends a lot on the performances of these memories, which are meant to operate at low voltage, consume less power and achieve higher manufacturing yield.

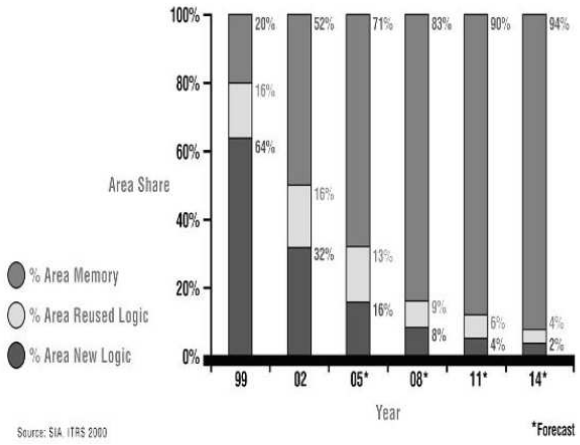

Fig. I.2 Embedded memory usage

In these recent years, the extensive use of low voltage SRAMs in mobile applications has also been driven by the necessity for faster operating and less power consuming memories. Table I.1 below represents different sizes of SRAMs and their respective performance comparisons in 90nm and 65nm technology nodes. In the 65nm node, the memories operate at a higher frequency (around 1.5 times faster than in 90nm node) while their dynamic powers are roughly the same as 90nm SRAM memories. As far as static power is concerned, the power reduction achieved in 65nm process lies between 2 to 6 times compared to the power consumed in a 90nm process. This reduction in static power consumption is achieved by using MOS devices in the input/output (IO) blocks having threshold voltages which have been increased by 25% in the 65nm compared to the 90nm node. As for transistors in the

SRAM cells in 65nm technology, they possess threshold voltages showing up to a 37% increase compared to those of the 90nm process.

Table I.1 Low power SRAM performance comparisons

Memory Size	Power Supply (Vdd)	Frequency (Mhz)		Dynamic Power (mw)		Static Power (µw)	
		90nm Process	65nm Process	90nm Process	65nm Process	90nm Process	65nm Process
128b	1.0	617	666	2.4	2.0	1.2	0.2
128b	1.2	990	1087	5.5	5.0	2.2	0.4
128b	1.32	1136	1315	8.3	7.6	3.3	1.2
64kb	1.0	249	308	3.9	3.2	11.0	2.9
64kb	1.2	431	625	8.7	8.9	22.7	6.2
64kb	1.32	487	719	12.1	12.6	35.0	10.5
128kb	1.0	248	305	6.6	5.0	12.2	3.3
128kb	1.2	431	617	14.3	14.8	24.9	7.1
128kb	1.32	485	709	19.8	20.7	38.3	12.4
256kb	1.0	248	301	11.9	9.1	14.7	4.0
256kb	1.2	427	598	25.1	25.7	29.4	8.9
256kb	1.32	483	689	35.0	35.7	44.9	16.3

I.2.2 Power consumption

Power consumption is an important issue in the design flow of memories, especially in very deep submicron technologies, and this phenomenon will be exacerbated as technology continues in scaling down. This is mainly true as transistor densities increase, leading to an increase in the complexity of the chip, which further needs to operate at a higher frequency. These factors have brought forward the total power dissipation problems in the memory. The total power consumed can be defined as the sum of dynamic and static power dissipated in the memory. Dynamic power refers to the total power consumed by the memory during a read or write operation involving the switching of the logic states in the various memory blocks, and the short circuit power resulting from a current flow between supply voltage and the ground.

$$P_{Dyn} = \eta \cdot C_{out} \cdot V_{dd}^2 \cdot F + I_{SC} \cdot V_{dd} \tag{I.1}$$

where η is the activity rate, C_{out} the output load capacitance, F the operation frequency of the memory, Vdd the supply voltage and I_{SC} is the short circuit current.

On the other hand, static power of the memory is the power consumed when the memory is either in the standby mode or when the power is off. As a result of the memory's state, the resulting leakage currents including [Roy03] reverse biased diode leakage, subthreshold current, gate leakage, Gate induced Drain Leakage current (GIDL) and the punch through current dissipate power.

$$P_{Stat} = I_{leak} \cdot V_{dd} \tag{I.2}$$

where I_{leak} represents the sum of all the leakage components.

The reverse biased diode leakage is due to the drain/source reverse biased conditions, causing pn junctions leakage current. Subthreshold current results from a current flow between the drain and source of the transistor when gate voltage is below the subthreshold voltage. Gate leakage corresponds to the tunnelling of a current from substrate to the gate and vice versa owing to a decrease in the gate oxide thickness and an increase of an electric field across the oxide. The GIDL current is due to the depletion at the drain surface below the gate/drain overlap region, leading to a current flow between that region to the substrate. The punch through current comes from the merging of the depletion region between the drain and source which gives rise to a current between these two regions. As an illustration, figure I.3 features the evolution of the two main leakage current components with the scaling of technology [kim03].

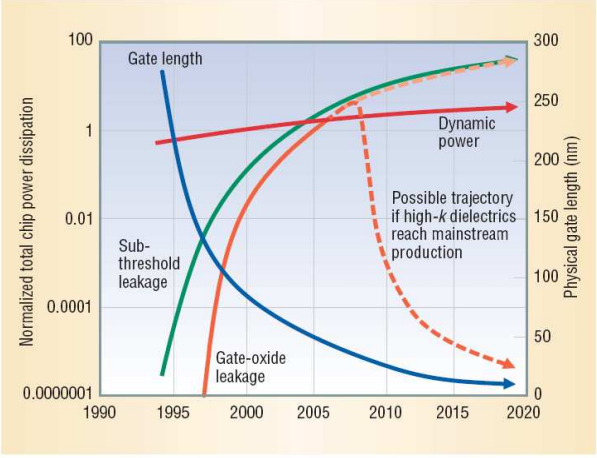

Fig. I.3 Total power consumption on a chip based on 2002 ITRS projection

This figure highlights the 2002 ITRS projected exponential increase of the subthreshold and gate leakage as the gate length decreases. The normalized chip power dissipation corresponds to 2002 ITRS projection normalized to that of 2001. It can be seen that an alternative to control gate leakage as gate oxide thickness reduces is to make use of high K dielectric materials, which can bring gate leakage under control. In [Bor05], the author explains that replacing the gate oxide with a high K material showing the same capacitance as the silicon dioxide but with higher thickness will minimize gate leakage. Nevertheless, the same problem will be encountered as the dielectric thickness will scale down over time.

Circuits showing excessive power dissipation characteristics are more prone to run time failures and present reliability problems. Every 10˚C increase in operating temperature approximately doubles a component's failure rate, thereby increasing the need of expensive packaging and cooling strategies [Mar99]. So as to minimize static and dynamic power dissipations, several low power techniques have been widely used as detailed in the next part.

I.2.3 Low power

Low power SRAM circuit has become a major field of interest, especially in the reduction of power consumption. Several low power techniques so far have enabled the proper operation of memories, though their complexities have kept on increasing to satisfy the high speed demand and throughput computations in various battery-backed applications. Some of the most widely used techniques in reducing power consumption include reduction of capacitance associated to bit lines and word lines through multi banking [Mar99], controlling the internal self-timed delay of the memory to track properly the delay of bit lines across operating conditions [Amr98], Dual Threshold design (DTCMOS) [Roy03] and the sleep mode concept [Ito01].

The largest capacitive elements in the memory are the word lines and bit lines each with a number of cells connected to them. Thus, reduction of the capacitive elements associated to these lines can reduce dynamic power consumption. This is achieved by partitioning the memory into smaller subarrays, such that a global word line is divided into a number of sub word lines. Similarly in multi banking, bit line capacitive switching is reduced when memory is being accessed as those bit lines are always involved in a discharge and precharge process.

Another dynamic power reduction technique consists in using a replica technique for word line and sense control in the SRAM. If this technique mainly aims at sequencing the read operation, it is also extremely efficient for low power design since it eases the use of static or dynamic supply voltage scaling techniques. In read operation, the technique consists in tuning properly the self timing circuitry by using a dummy driver and dummy core cells in order to trigger the sense amplifier when a desired differential voltage has accumulated on the pair of bit lines considered. In doing so, a fast sensing of the minimum differential voltage is performed.

Dual threshold design (DTCMOS design) is also a good approach to reduce the leakage in SRAM memories. Indeed, making use of high and low threshold voltage transistors along uncritical and critical paths allow significant reductions of the overall leakage. This technique is almost interesting as it is fully compliant with the application of the sleep mode concept. The latter implies powering down the periphery logic while the voltage level of the SRAM arrays remain activated for data retention purpose. However, this method is only useful if the memory remains idle for a long period.

Some of the above stated techniques can lower both static and dynamic power but the best mix depends on which types of power dissipation problems a designer wants to minimize and the technology in which the design is being performed.

I.2.4 Design for Manufacturability (DFM)

As the complexity in the density of transistors is skyrocketing along with shrinking of transistors, devices in embedded memories are becoming more sensitive to the disturbances of IC manufacturing processes. Due to this inherent problem, traditional CAD tools which were meant to simplify the work of designers by allowing them to create the nominal design meeting the desired performance specifications are no longer satisfactory [Zha95].

Furthermore, environmental conditions such as temperature fluctuations and voltage variations also induce IC performance variations to increase. The variability effects may for example cause faulty read or write operations in an SRAM due to an increase in access time. To cope with this problem and avoid unnecessary yield losses, designers had to revamp the IC design methodologies and introduce new design techniques.

For instance, statistical techniques can take into account the variability effects which are the root causes of parametric and functional failures in chips. Subsequently, design for manufacturability issue is gaining momentum at each stage of the design procedure in view of improving the yield and reliability of ICs, thereby ensuring stable volume production. The next two sub sections introduce the two main facts related to DFM, yield and reliability challenges.

I.2.4.1 Yield Challenge

I.2.4.1.1 Yield learning during ramp up phase

The improvement of production and cost effectiveness in the semiconductor industry have become an increasing necessity in this very competitive sector, owing to pressure arising from shorter time to market and time to volume (TTV). The period elapsing between the completion of development of the product and its full capacity production is referred as production ramp up [Ter01]. C. Terwiesch and R. E. Bohn [Ter01] specify that two conflicting factors characterize this period, a low production capacity (poor yield) and a high demand due to the "relative freshness" of the product. The pressure faced by the industry between low capacity and high demand is referred to as the "nutcracker". In fact, very often it takes a longer time to achieve a higher yield during ramp up as technology node shrinks.

This is illustrated in figure I.4 which represents the cycles of learning for three different technology nodes and their respective empirical yields. For instance, in order to achieve a yield of 90%, the number of yield cycles learning required doubles as the process technology moves on from 130nm to 65nm. This increase in learning is attributed to yield detracting mechanism which occurs more frequently as devices shrink and become more sensitive to variability aspects.

P. K. Nag and W. Maly [Nag93] defined the cycles of learning or improvement as the total time needed to:

(i) detect and localize a failure, which leads to process intervention (T_f)

(ii) process the correction and for new parameters to become effective (T_e)

(iii) and the time between performing process correction and the time when yield improvement is realized (Tr).

Therefore, the total time for the yield change T_c is expressed by:

$$T_C = T_f + T_e + T_r \tag{I.3}$$

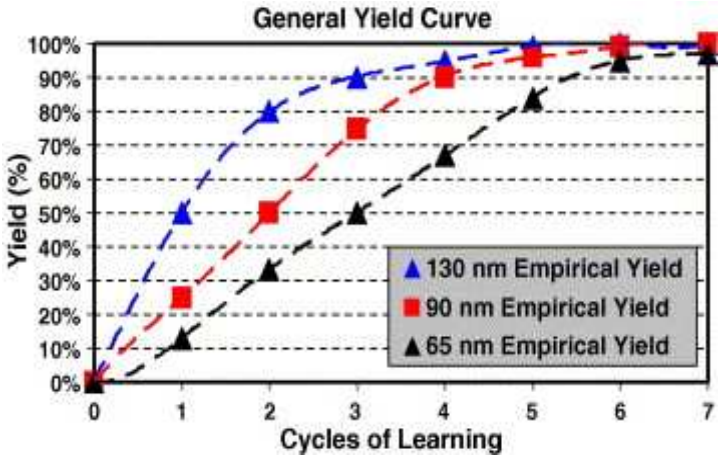

Fig. I.4 Behaviour of learning curve with technology node evolution.

I.2.4.1.2 Yield enhancement during production phase

During the production phase, memory yield is defined as the percentage of fully functional memory chips from all of the fabricated chips or, alternatively by the ratio between the fully functional memory chips and all the fabricated chips [Har01].

However, the large scale integration of memory area on a SoC not only leads to an increase in the size of the die but it can be problematic for the SoC's yield due to its great dependency on the yield of the embedded memory, since the area memory covers around 80% of the total surface of the chip according to figure I.2.

In fact, memory yield is limited by numerous manufacturing defects which stem from process disturbances or non optimal design due to aggressive scaling of transistors. These may cause either inadequate performance for example excessive power consumption, too long delay giving rise to timing constraint violation, or functional failure due to spot defects which produces shorts or opens in the circuit's connectivity [Mal96].

Hence to achieve lower silicon cost i.e. to make multi million transistor systems on a single die both feasible and cost effective, there is an urgent need for providing effective methods to improve memory yield.

In addition to stringent fabrication control (intentional process de-centering) and statistical worst case design, this can be achieved though the use of redundancy [Har01, Kim98, Zor02] added to the circuits, which consists in replacing defective circuitry with spare elements.

Some of the different types of redundancies include word redundancy, wordline redundancy, bit line redundancy and IO redundancy [Rod02]. Word redundancy consists in adding a few redundant flip flop based words, each of which corresponds to a logical address of the RAM. In word line redundancy approach, it is possible to replace one or more word lines with spare rows. Bit line redundancy for its part involves adding redundant column to the memory array whereas IO redundancy consists in replacing the bit lines along with their respective sense amplifiers with redundant elements. In addition to redundancy, error detector and correcting code can also be used to improve the yield. Yield improvement method will be discussed in further details in the next chapter.

I.2.4.2 Reliability issue

Reliability issue is emerging as a design challenge in embedded memories, especially as transistors geometry shrink, thus making SRAM more prone to soft errors. These soft errors are caused by neutrons from cosmic rays or α-particles from radioactive impurities in electronic material that may cause new data to be written in the memory. Figure I.5 [Bor05] shows the random errors occurring in a chip (logic and memory).

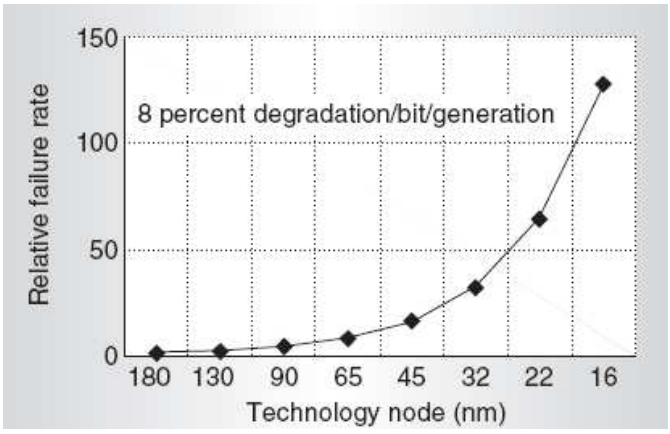

Fig. I.5 Soft error failure of a chip

According to [Bor05], the expected rate of increase of the relative failure per logic state bit per technology node will be around 8%. Moreover in the 16nm generation, the failure rate will be almost 100 times that at 180nm.

In addition to soft errors, physical breakdown and other degradation mechanisms like gate oxide breakdown in NMOS and the Negative Bias Temperature Instability (NBTI) effect

[Mcp06] in PMOS also affect SRAM reliability. For instance, NBTI phenomenon impacts on the read stability of the memory. Commonly, memory reliability is expressed by the probability that the memory performs its designed functions with the designed performance characteristics under the specified power supply, timing, input, output and environmental conditions until a stated time t [Har01].

Reliability specialists often represent the failure rate of a memory (Mean Time to Failure or MTTF) with the time of device usage t by the traditional bathtub curve indicated in figure I.6 below. The bathtub curve possesses three distinct periods i.e. a decreasing failure rate representing the infant mortality period, followed by a constant failure rate which is characterized by a useful device life and terminates by an increasing failure rate due to a wear out period.

Infant mortality problems have been for long a critical issue to both manufacturers as well as to customers receiving memory products lasting for a few hours to few months. These defects arise from left over or latent defects that do not necessarily expose themselves and can skip manufacturing tests [Mak07]. They manifest themselves through intensive electrical and thermal stresses during use, causing a significant functionality problem.

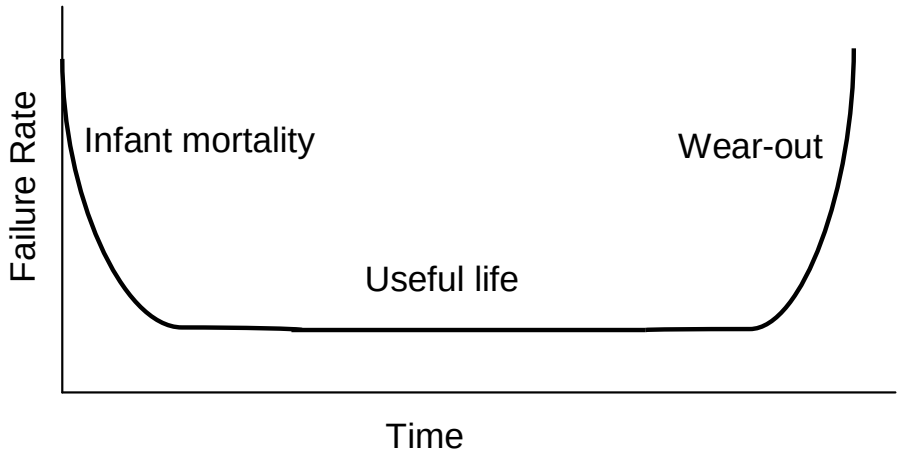

Fig. I.6 The bathtub curve

To counteract infant mortality problems and improve product robustness, several techniques are widely used, for instance the burn-in test. In this process, extreme operating conditions i.e. high temperatures (up to 150˚C) and high operating voltages are applied for example at the wafer level to stress the device and accelerate the memory's failures within some hours. Consequently, failing memory chips can be discarded.

The useful life period corresponds to the time lapse whereby memory products have the lowest failure rate. In other words, it represents the useful device period of the product which has entered a normal life. The wear-out phase occurs after the useful life period as the product ages, resulting in an increasing failure rate. Wear out phase results from electromigration problems (accelerated by high temperature and operating frequency conditions) and oxide break down triggered by high electric field conditions with decrease in gate oxide thickness.
The reliability aspect is undoubtedly closely related to yield performance, since a high reliability is translated onto very high yield related constraints, both in functionality and correct parametric features [Pap07].

I.3 Architecture of SRAM memories

Figure I.7 shows the block diagram of a synchronous single port SRAM. It is composed of:

(i) Memory cores (left and right) containing a matrix of 6T SRAM cells. Each memory cell is connected to a pair of bit lines (BL/BLB) and either BL or BLB is discharged when a row of cells is selected through word line.
(ii) A timing generator block providing the internal signals for activating an operation.
(iii) X and Y decoders to access the required cell as the input addresses are specified.
(iv) A dummy bit line driver, an important block for synchronizing the internal signals in a self-timed memory.
(v) An I/O circuit consisting of :
    a. column multiplexer found in the Y post decoder
    b. sense amplifiers and write circuitry
    c. output buffers found in the post multiplexer

The architecture of the SRAM shown in figure I.7 is referred to as a butterfly memory. It is so called owing to its symmetry with a left and a right memory core. The next sub section details the functionalities of each block.

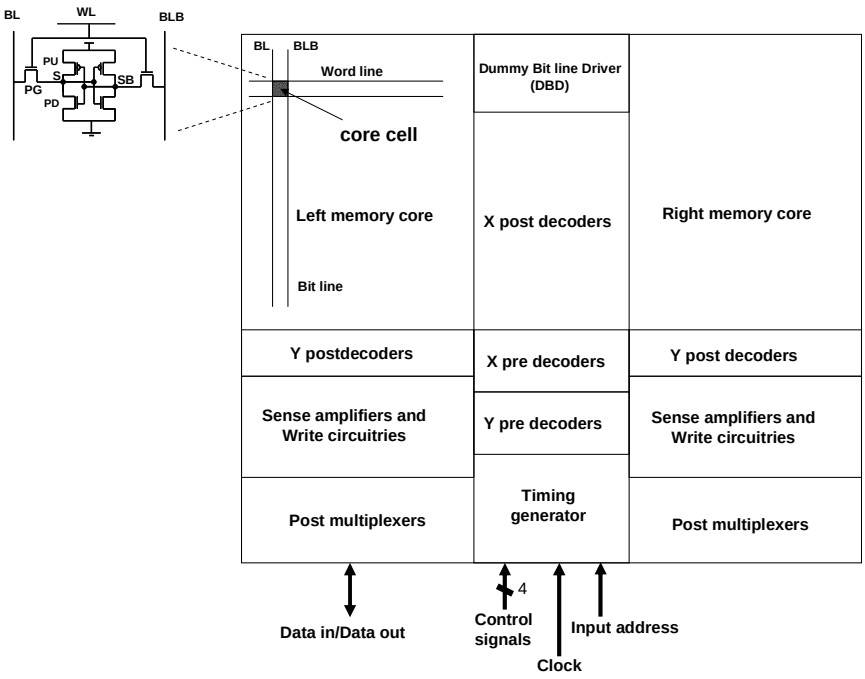

Fig. I.7 Block diagram of an SRAM architecture

I.3.1 Functional blocks of the SRAM architecture

I.3.1.1 Control Block

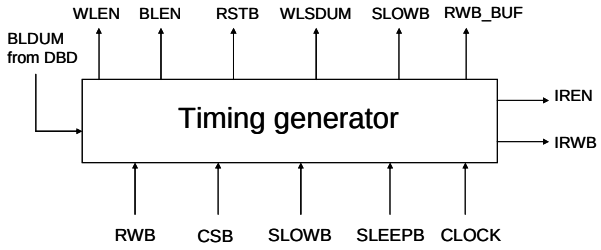

Fig. I.8 Block diagram of control block

The timing generator or control block, as shown in figure I.8, receives the input signals (control signals and clock) and generates the appropriate signals for either triggering a read or write operation or setting the memory in a sleep mode operation. The control signals are composed of 4 signals:

(i) a read/write input signal (RWB) which is low in the write mode and high for a read mode,

(ii) an active low chip select input signal (CSB),

(iii) a slow mode signal (SLOWB) specifying whether the memory is operating in a high performance mode (SLOWB high) under high operating voltages or low performance mode (SLOWB is low) in ultra low power operating conditions,

(iv) and an active sleep mode input signal (SLEEPB) for leakage reduction in standby mode.

I.3.1.2 Pre Decoder

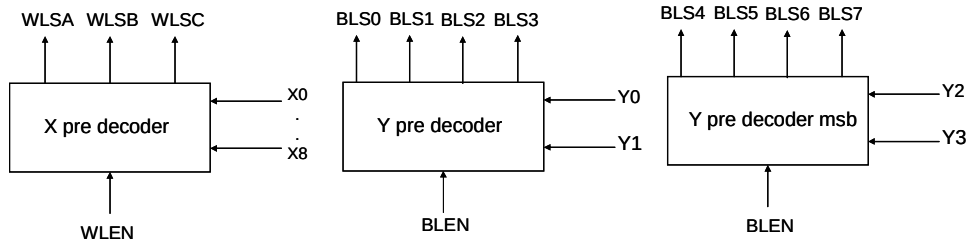

Fig. I.9 Block diagram of X and Y pre decoders

The pre decoder can be either in X or Y depending upon their functionalities. In fact, the X pre decoder receives input X addresses (X0 to X8) and selects the appropriate row decoder among several row decoders in the X post decoder, via signals WLS (A/B/C) (figure I.9) Similarly, the Y pre decoder is meant for receiving input Y addresses (Y0 and Y1) and BLEN after which, the required column multiplexer among several multiplexers is chosen in the Y post decoder by the signals BLSi (i=0 to 3). As for the Y pre decoder msb block, the latter receives input addresses Y2 and Y3 and signal BLEN and produces signals BLSn (n=4 to 7). BLSn signals will be used in the post multiplexer block for selecting the read or write circuitry.

I.3.1.3 Post Decoder

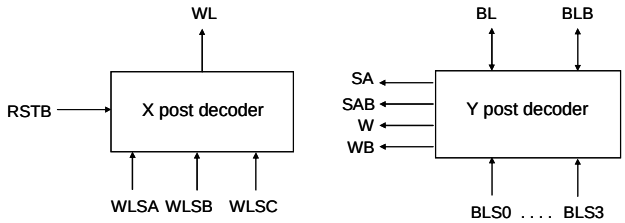

Fig. I.10 Block diagram of post decoders

The X post decoder and Y post decoder receive their respective input signals WLS (A/B/C) and BLSi (i=0 to 3) from the X and Y pre decoders (figure I.10). The X post decoder or more commonly known as row decoder generates the signal word line WL and picks out the

appropriate row of SRAM cells amid several rows, corresponding to the row where a datum is supposed to be read from or written to. The signal RSTB resets WL to 0 after the operation. For its part, the Y post decoder which is composed of a series of column multiplexers selects the corresponding column of SRAM cells (BL/BLB) through 1 multiplexer, after receiving the signals BLSi issued from the Y pre decoder. Its output signals SA/SAB and W/WB are connected respectively to the sense amplifier and the write circuitry block.

I.3.1.4 Memory Core

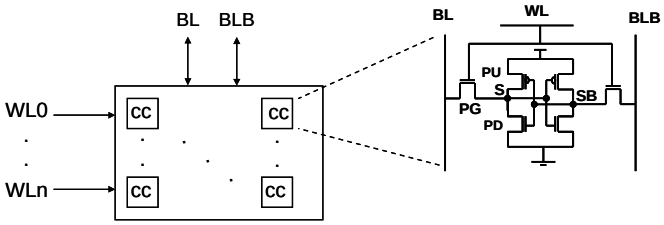

Fig. I.11 Block diagram of memory core and the 6T SRAM cell

The memory core, left or right, is comprised of the elementary SRAM cells or core cells (cc) where data (‘0’ or ‘1’) are stored at nodes S and SB (figure I.11). A 6T SRAM cell is composed of two inverters formed by four transistors (two pull up PMOS transistors PU and two pull down NMOS transistors PD) and two pass gate transistors PG connected to a pair of bit lines (BL and BLB). The flip flop operation is achieved by connecting the input and output of one inverter to the output and input of the other inverter.

I.3.1.5 Dummy Bit line driver

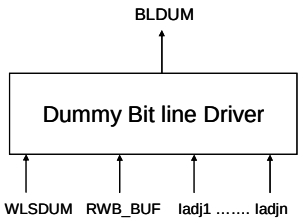

Fig. I.12 Block diagram of Dummy Bit line Driver

Figure I.12 represents the block diagram of the dummy bit line driver (DBD), which is one of the most essential components of the memory. It consists of a series of parallel branches of stacked transistors, with the stacked transistors representing the pass gate and pull down

transistors (figure I.12). The DBD receives WLSDUM and RWB_BUF signals from the control block. RWB_BUF signal sets the use of the DBD in either a read (RWB_BUF=1) or write mode (RWB_BUF=0), input pins Iadji (i=1 to n) control the discharge rate of BLDUM and hence the memory is set either in a fast or slow read mode or write mode. On the other hand, WLSDUM activates or deactivates the transistors PGi (figure I.24) representing the pass gates of the SRAM cell. It should be noted that pins Iadji are either hardcoded or provided by the environment. The most important role of the dummy bit line driver is displayed during the read and write operation. In fact, it is the dummy bit line driver which fires the sense amplifier, through the control block, at the right time when a datum is read from a selected SRAM cell. As for the write operation, the dummy bit line driver ensures that sufficient time is given to a datum to be properly written in the SRAM cell before switching off the word line.

I.3.1.6 Sense amplifiers and Write circuitries

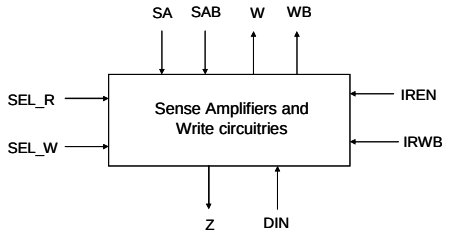

Fig. I.13 Block diagram of sense amplifiers and write circuitries

In figure I.13, signal SEL_R (from post multiplexer block) is used for selecting a sense amplifier corresponding to the column of core cells which has been chosen. The output signals SA/SAB from the multiplexer block correspond to the input signals of a sense amplifier. The sense amplifier consists of a cross coupled latched amplifier connected to a pair of bit lines. It senses the difference in potential between BL (SA) and BLB (SAB) associated with the SRAM cell, when BL or BLB is being discharged by the core cell during a read process. Once the difference in voltage between BL and BLB reaches an appropriate level (around 10% of VDD), the sense amplifier is triggered by signal IREN and amplifies the differential voltage. The output value is collected as Z. Similarly, SEL_W (from post multiplexer block) is involved in the selection process of a write circuitry attached to the column of SRAM cells where a ‘0’ or ‘1’ needs to be written at a predefined memory cell

address. The datum DIN is transmitted to the SRAM cell as output signals W (BL) and WB (BLB) when signal IRWB is turned on.

I.3.1.7 Post Multiplexer

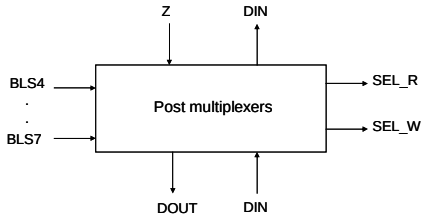

Fig. I.14 Block diagram of post multiplexer

The post multiplexer is the block through which data transit during a read or write operation (figure I.14). It consists of several multiplexers and a latched structure for storing the data read (Z), before being transmitted to the output pin as DOUT. It is also composed of buffers for transmitting input data DIN to a specified address in the memory core during a write operation. Moreover, the post multiplexer produces signals SEL_R and SEL_W from its input signals BLSi (i=4, 5, 6, 7), issued from the Y predec_msb block so as to choose the needed sense amplifier or writing circuitry.

I.3.2 Configurations of the Memory

The memory, with butterfly architecture, shows three available configurations i.e. tall, normal and wide depending on the type of multiplexers used in the post multiplexer block (figure I.15). For instance, a tall configuration uses a 4:1 multiplexer, a normal configuration uses an 8:1 multiplexer and a wide configuration uses a 16:1 multiplexer.

Hence, the choice of the type of column multiplexer influences the size of the memory. To have a better understanding, let us first start by defining the word width (WW) as the number of bits per word and the word depth (WD) as the number of words. The word width varies from a miminum of 4 bits to 128 bits with increments of 1 bit, whereas the the word depth varies from 32 words to 8192 words with increments of 8, 16 and 32 words depending on the size of the multiplexer (Smux) i.e. either 4, 8 or 16.

Therefore, a simple way of defining the number of rows (NR) and columns (NC) in a memory matrix is given by the following operations.

$$NR = \frac{WD}{Smux} \tag{I.4}$$

$$NC = WW \times Smux \tag{I.5}$$

Thus, the memory size (Smem) is given by:

$$Smem = NR \times NC \tag{I.6}$$

Figure I.15 below gives a brief illustration of the available memory configurations with respect to the size of the column multiplexer used. The figure also displays the size of the memory (height and width), depending on its respective configuration.

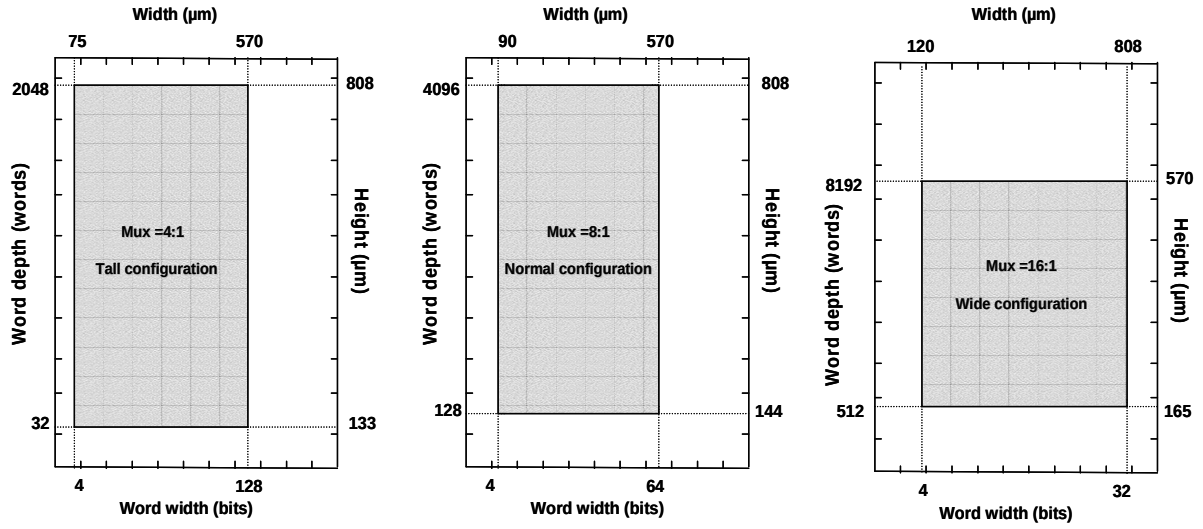

Fig. I.15 Memory Configurations

In our case, we have been working on a low power memory with the maximum available size of 256kb which consists of 8192 words of 32 bits, and having a wide configuration.

I.4 Operating mode of the memory

I.4.1 Read operation

Figure I.16 shows the block diagram of a read operation in the memory. The read operation is triggered by the rising edge of the clock CLK. Signals CSB is set low, RWB high and depending on the operating mode of the memory i.e. fast mode for high performance or slow mode for low performance mode, SLOWB will be set either high or low accordingly. Internal signals word line enable (WLEN) and bit line enable (BLEN) are generated in parallel,

whereas signal WLSDUM is issued from internal signal BLEN. WLEN going through the X predecoder block along with the required X addresses (X0 … X8), makes a proper selection of the appropriate row decoder found in the X post decoder.

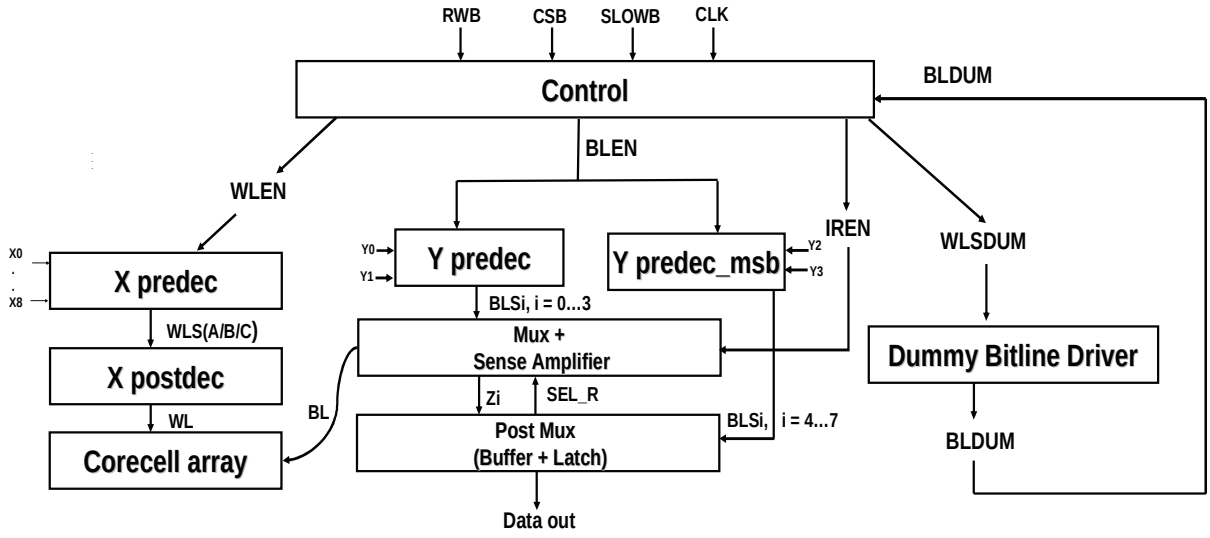

Fig. I.16 Block diagram of a read operation

The rising edge of BLEN triggers both blocks of Y pre decoders. Specified couples of input addresses (Y0, Y1) for block Y predec allow the selection of the required column multiplexer in the Mux/Sense Amplifier block via signals BLSi (i=0 to 3), thereby choosing the column of corecells needed through signal BL. Similarly, input couple (Y2, Y3) of the predec_msb block picks out the appropriate post multiplexer via signals BLSi (i=4 to 7) in the post mux block. The post mux block provides the signal SEL_R to the block Mux/Sense Amplifier. SEL_R selects the desired sense amplifier corresponding to the column of corecells chosen above and also stops the precharge of BL/BLB associated with the sense amplifier before any read operation. Simultaneously, the X post decoder chooses the row of core cell by activating word line (WL) and subsequently discharges BL or BLB, depending whether a '0' or '1' is being read from an SRAM cell. In parallel WLSDUM signal, issued from the control block, turns on the dummy bit line driver which discharges the line bit line dummy (BLDUM). BLDUM going through the control block fires the specified sense amplifier at the appropriate time through signal read enable (IREN). In this way, the differential signal voltage between BL and BLB is detected and amplified before being transmitted as the output data.

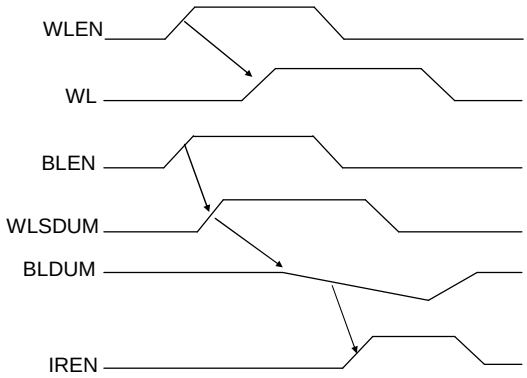

Fig. I.17 Timing diagram of a read operation

Figure I.17 illustrates the timing diagram of the read operation with the main signals involved. The arrows represent the respective edges responsible for triggering the subsequent signals.

I.4.2 Write operation

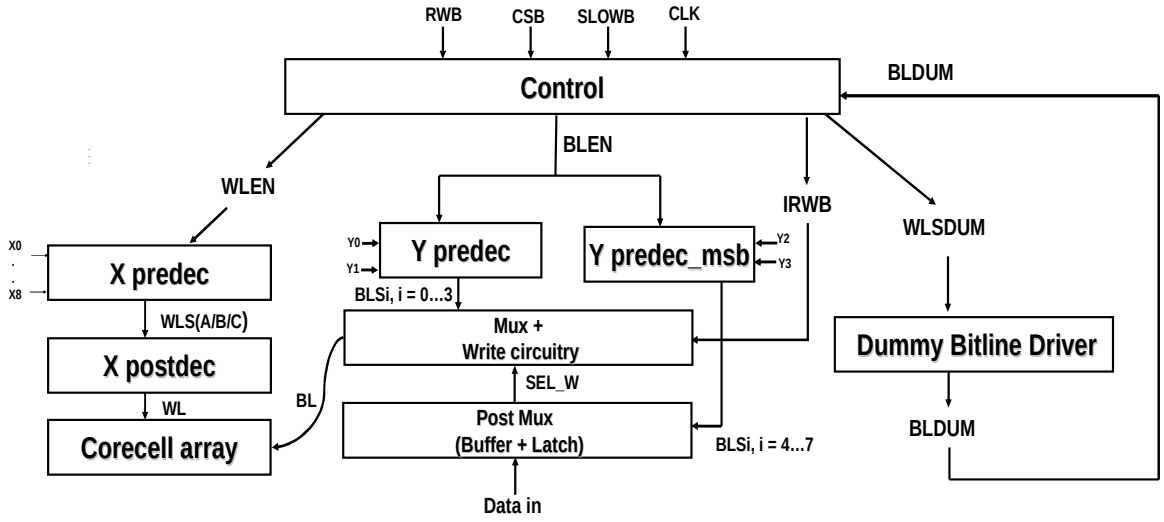

Fig. I.18 Block diagram of a write operation

Figure I.18 depicts the block diagram involved in a write operation. In this write process, signal CSB is set at its active low level and RWB is put at a low level. Similarly, as in the read operation, the memory can either operate in a slow operating mode at a low supply voltage in the case when operating speed is not crucial or in a fast mode when a maximum operating speed is required at high supply voltage. This is achieved by setting pin SLOWB at '0' or '1'. The rising edge of the clock starts the write operation cycle. Signals WLEN and BLEN are generated simultaneously from the control block, whereas both signals IRWB and

WLSDUM are generated internally in the control block from signal BLEN. Input signals at the X predec block, which are composed of WLEN and X addresses (X0 to X8), generate signals WLS (A/B/C) in the selection process of the appropriate row decoder found in the X postdec block. The rising edge of BLEN triggers both the Y predec and Y predec_msb blocks. Couple of input Y addresses (Y0, Y1) in the Y predec block choose the required column multiplexer, via signals BLSi (i=0 to 3), among several multiplexers corresponding to the column of SRAM cells where data need to be written. In the same way, the couple of Y addresses (Y2, Y3) provide the signals BLSi (i=4 to 7) to the post mux block in order to pick out the specified writing circuitry through signal SEL_W. Consequently, signal IRWB fires the writing block in view of discharging BL/BLB. The choice of the discharge of BL/BLB depends on the initial value stored in the chosen SRAM cell and the value which is carried by the input datum that needs to be written. As soon as BL/BLB has been discharged at an appropriate voltage level, WL signal issued from the x postdec block activates the array of core cells. In this way, a particular SRAM cell is chosen. Depending whether a '0' or '1' has to be written, the node S/SB of the SRAM cell will be discharged accordingly. The time during which WL is maintained activated strictly depends on the discharge time of signal BLDUM, generated by the dummy bit line driver. Indeed, dummy bit line driver is triggered by signal WLSDUM coming from the internal signal BLEN. Therefore, BLDUM signal is an important signal since it specifies the time interval required for a datum to be written in the SRAM cell.

Figure (I.19) illustrates the timing diagram of the write operation with the main signals involved. As previously, the arrows represent the respective edges responsible for triggering the subsequent signals.

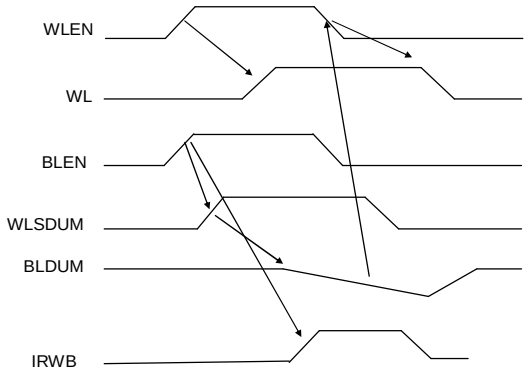

Fig. I.19 Timing diagram of a write operation

I.5 Metrology of SRAM

Following the previous descriptions of the read and write operations involved in the memory, it can be clearly seen that these operations display a certain complexity in the presence of numerous signals, particularly when signal racing conditions exist. The first part investigates the difficulty appearing during the operating mode of the memory, particularly in order to ensure a proper synchronization between racing signals. To allow a correct read and write operation in such conditions, a structure called a dummy bit line driver is normally used. The following part describes its architecture and its associated functionality.

I.5.1 Complexity of Read/Write operation

Read and write operations involve the use of hundreds of signals that are triggered simultaneously, and these render the operations quite tedious. A representation of the operations is depicted by the signal transition graph (STG) in figures I.20 and I.21.

In those figures, signal races are said to occur between any parallel branches issued from the same signal. The '+' sign corresponds to the rising edge of a signal whereas the '-' sign corresponds to its falling edge. The shaded oval means that the signal is back to its original off state. The STG has been simplified to show only the main signals implied in the operations.

Figure I.20 represents a clear illustration of the STG of the memory in a read mode. The two most important racing paths which need to be considered are shown by the two single arrows, in this case between T_{SAON} and T_{BL}. T_{BL} (around 940ps) is the time elapsed between the activation of the signal WLEN and the discharge of BL at Vdd-Δ, where Δ is around 10% of Vdd. T_{SAON} (around 1130ps) is the delay between the activation of BLEN and the firing of the sense amplifier. It should be noted that signals BLEN and WLEN are identical signals generated from the control block. Therefore, both path delays have to be perfectly synchronized such that the activation of the sense amplifier occurs when bit line has reached an appropriate discharge voltage level. Indeed, triggering the sense amplifier too early or too late might result in a faulty read operation or excessive dynamic power consumption in the memory.

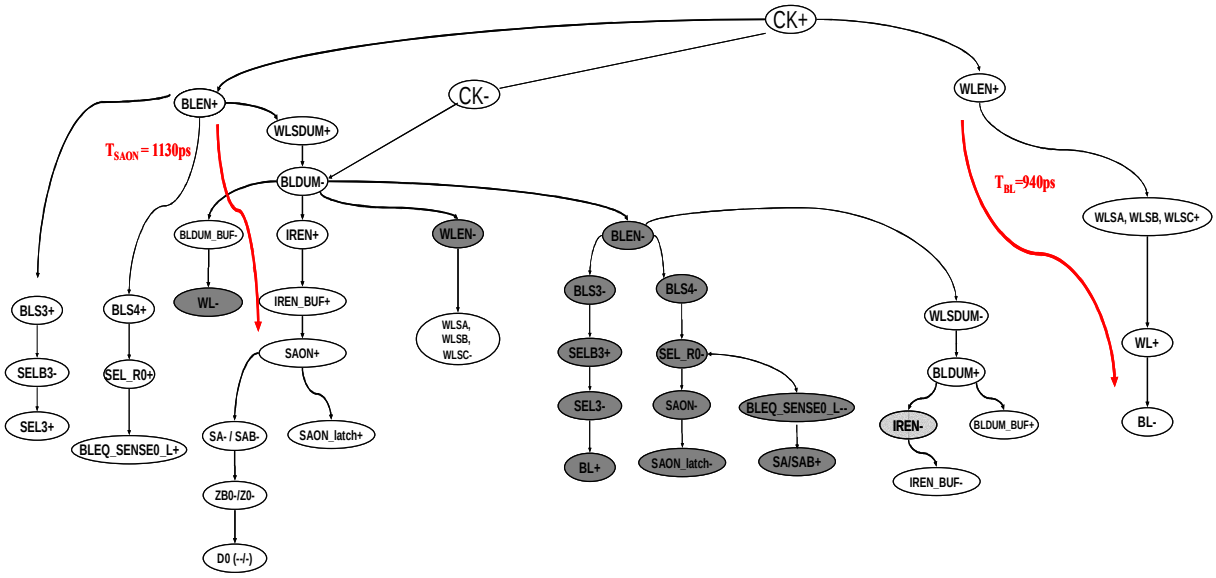

Fig. I.20 STG of a Read operation

Figure I.21 illustrates the STG of a write operation. The two racing paths which we will consider are also represented by the two single arrows, T_{WL} and $T_{node\_discharge}$. $T_{node\_discharge}$ (around 740 ps) represents the delay between the activation of signal $WLEN^+$ and the instant when the SRAM's cell internal (s^-) node voltage crosses Vdd/2 i.e. either when a '0' or '1' is being written. T_{WL} (around 1360 ps) is the delay between the activation of signal $BLEN^+$ and the falling edge of world line WL^- at Vdd/2.

Data can be properly written in a selected SRAM cell if only sufficient time is given to the internal node voltage of the SRAM cell to flip before deactivating word line. If word line is switched off too early or too late, the data might not be written properly or this might result in excessive dynamic power consumption. In either a read or write operation, the proper discharge of signal BLDUM ensures that these operations are carried out correctly. Signal BLDUM triggers the sense amplifier in the case of a read operation or the closure of word line in the case of a write operation. The control of the discharge of dummy bit line is performed by a dummy bit line driver. The next part describes in details its architecture and operating mode.

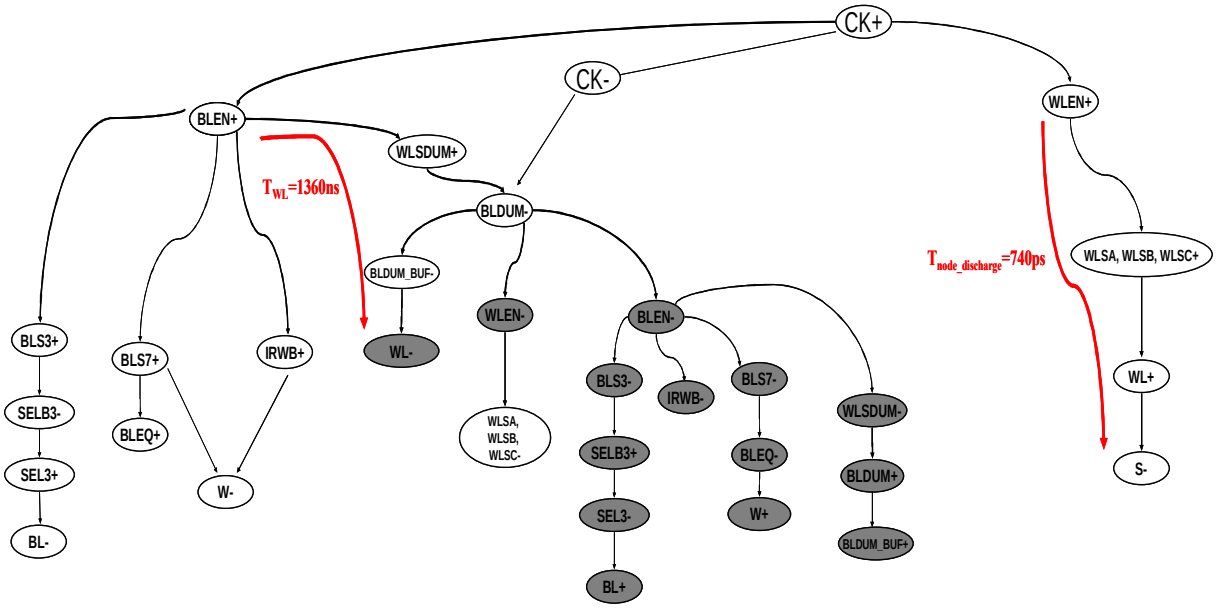

Fig. I.21 STG of a Write operation

I.5.2 Dummy Bit line Driver

The dummy bit line driver is an essential component of the embedded SRAM during a read or write operation. To grasp the real importance of this structure, let us first start by defining the read margin and write margin. The read margin can be defined as being the differential voltage ΔSa between input signals bit line BL and its complementary BLB of the sense amplifier when the latter (SAON) is fired at that instant by the signal BLDUM. This is shown in figure I.22 below.

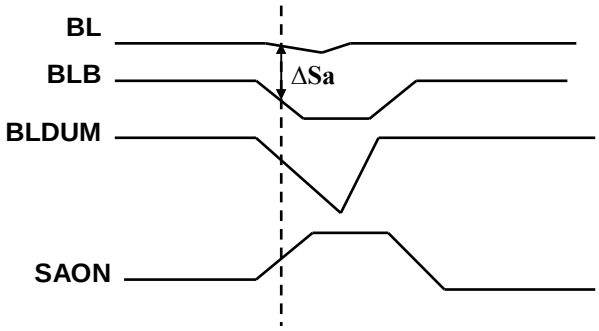

Fig. I.22 Definition of the read margin

On the other hand the write margin is defined by the time difference ΔTwm when the internal node of the SRAM cell (S or SB) reaches Vdd/2, and when the falling edge of wordline WL crosses Vdd/2 represented by figure I.23.

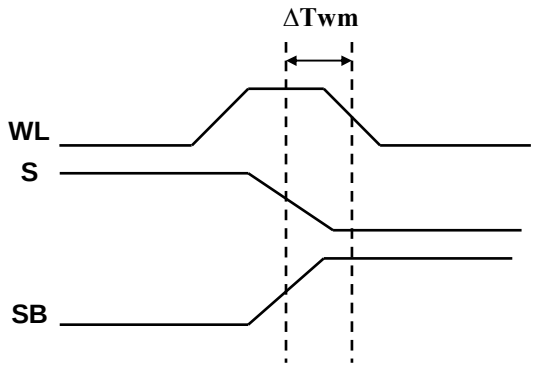

Fig. I.23 Definition of the write margin

Therefore, adjusting the read or write margin with respect to actual supplied voltage of the memory is necessary such that the read or write operation is performed correctly.

In fact with decreasing supply voltage, during a read access, the sense amplifier requires a larger voltage swing for detecting the differential voltage between BL and BLB i.e. it requires a larger read margin. In the same way during a write cycle at low Vdd, the selected SRAM cell necessitates a higher voltage swing before switching off the write driver and WL, implying the need for a bigger write margin. This means that an increase in read or write margin at low voltage allows the memory to operate in a slow mode. Moreover an increase in read or write margin is required with decreasing supply voltage owing to the impact of mismatch parameters of transistors like threshold voltage on memory's performance. Proper adaptation of these margins is achieved through the dummy bit line driver structure. On the other hand, when the memory needs to operate at high supply voltage, a proper adjustment of the current supplied by the dummy bit line driver is possible so as to decrease these margins. Let us now see in greater details the structure of this dummy bit line driver (figure I.24). It is mainly composed of stacked transistors PGi and PDi representing the pass gate and pull down transistors of a 6T SRAM cell. Signal rwb is used in specifying whether a read or write operation is performed whereas input pins Iadji (i=1, 2, 3, 4) are provided either by the environment of the memory or individually hard coded. Logic gates g1 and g2 are used to mimic the signal WL which controls the pass gate of the SRAM cell. Transistor Pr is used in the precharge of dummy bit line to Vdd before any read or write operation. As for the dummy bit line, it displays the same characteristics as a bit line by coupling an identical number of dummy SRAM cells to dummy bit line as the number of 'true' SRAM cells found on a bit line, thereby replicating the load of the SRAM cells on bit line. In addition to the replica loads, the

matching between dummy bit line and bit line is realized by ensuring that they both have identical properties including same length, width, thickness etc.

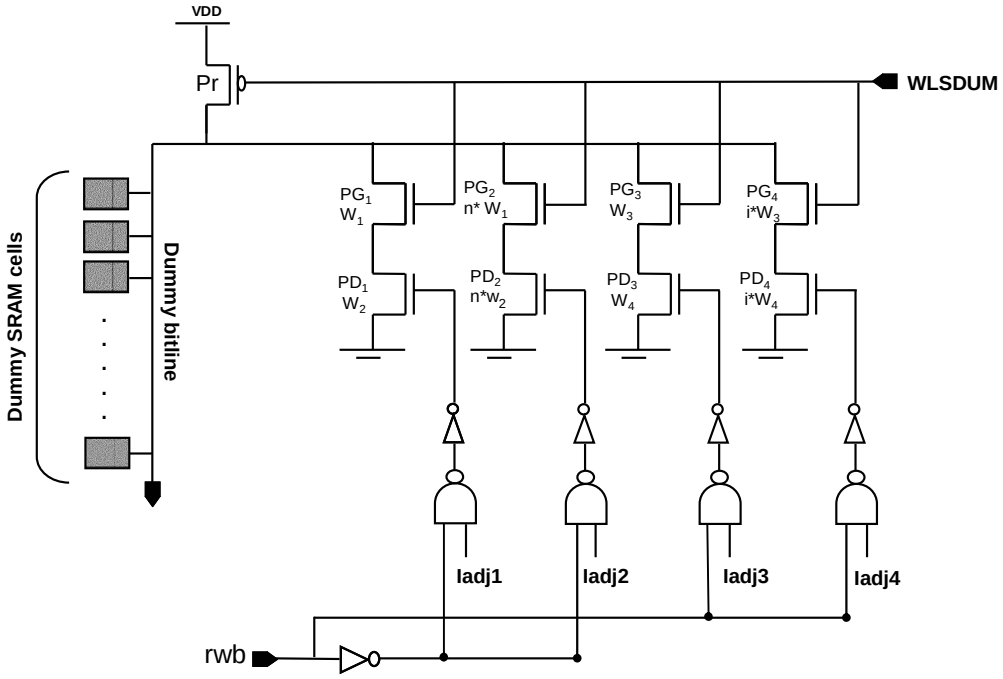

Fig. I.24 Architecture of Dummy Bit line Driver

During a read operation, the third and fourth branches of stacked transistors are selected by the signal rwb which is at '1'. Depending on memory operation mode i.e. either at low or high supply voltage, pairs of transistors (PG3, PD3) or (PG4, PD4) will be activated accordingly by pins Iadj3 or Iadj4. At low supply voltage, (PG3, PD3) could be activated to provide a slow discharge of dummy bit line whereas at higher supplied voltage, (PG4, PD4) could instead be triggered. The widths of transistors (PG4, PD4) are i times greater than (PG3, PD3) to provide a faster discharge rate of the dummy bit line. Moreover, each branch of stacked transistors can also be composed of several parallel stacked MOS transistors which are not shown here for clarity purposes. Furthermore, the number of branches is not limited to only four. It should be noted that generally not only one branch of stacked transistors is assigned for a high voltage or low voltage operating mode, but two or more branches can be chosen in either operating mode.

In a write cycle, rwb is set low. Stacked transistors (PG1, PD1) and (PG2, PD2) are selected and as in the read mode and depending on supply voltage, the first or second branches will be activated through pins Iadj2 or Iadj3. The second branch of stacked transistors is n times

greater than the first branch in order to discharge dummy bit line faster in high speed mode in figure I.24. Likewise in the write mode, the number of selectable branches is not limited to one in both slow and fast mode.

Conclusion

In this chapter, we have seen that embedded memories are key components of SoCs as they are meant to occupy larger surfaces in the coming years-as much as 80% or more devoted to memory elements. Owing to this high density occupation of memory on SoC, the overall reliability and yield of the chip are mainly dominated by the performance related to these memories. This is why a fundamental understanding of the operating mode of the memories, in this case SRAM, with all the signal races conditions involved is necessary. Indeed, a proper synchronization of the racing signals is required to avoid a faulty read or write operation. This is achieved through the proper use of the dummy bit line driver which triggers the sense amplifier at the appropriate time in the case of a read operation or which deactivates word line after writing data to the SRAM cell. However, performances related to SRAMs are not only impaired by variability effects appearing during the manufacturing process steps but also during operating mode, caused by environmental conditions i.e. supply voltage and temperature conditions. Therefore in this research work, after a general introduction to SRAM, the second step has been in understanding the causes of variability effects.

Chapter 2

Variability aspects in eSRAM

The performance of MOS devices and interconnects are becoming unpredictable due to an increase in the variability phenomena appearing as transistors continue to shrink. The major sources of variations are attributed to manufacturing processes, environmental conditions and aging conditions. In this chapter, we will detail a non exhaustive list of their main causes and their resulting impacts on the performances of embedded SRAMs. Some of the most widely used techniques to maximize the yield of the memory in the presence of variability conditions will also be presented.

INTRODUCTION

The aggressive scaling of MOS devices has introduced numerous challenges that were so far quite familiar to designers, namely operating speed of the IC, power consumption and surface area of the chip. Nevertheless, the trend towards very deep submicron region has been accompanied by the surfacing of new phenomena that were once imperceptible relative to the larger feature sizes in older technologies. Such phenomena, more commonly known as variability effects, constitute a severe problem as they influence the process parameters of transistors as shown in figure II.1. These sources of fluctuations impede the overall performance of the chip, for instance its operating frequency and overall static power consumption [Eis97, Rao06].

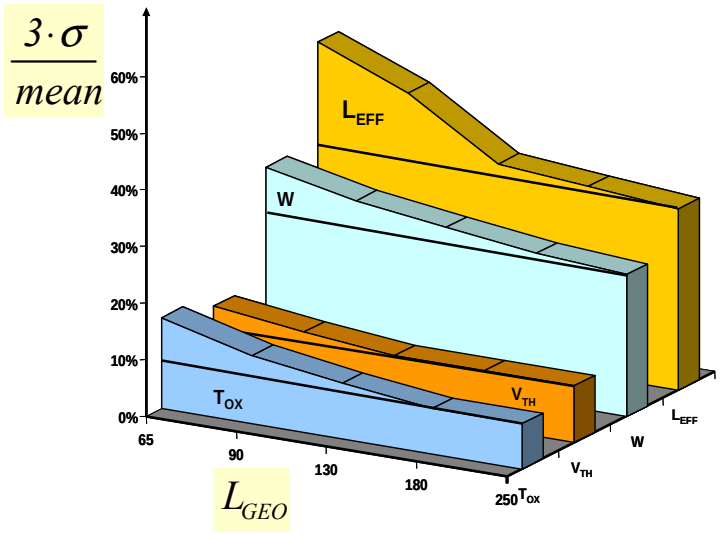

Fig. II.1 Variability trend in process parameters with technology evolution

Those variability effects include problems linked to environmental factors (voltage and temperature fluctuations), to manufacturing process variations and aging process. In this chapter, we will detail these phenomena so as to have a better understanding of their origins and see how they can be classified. Next, we will analyse their impacts on performances of SRAMs. Finally, we will see the solutions that are actually being used to improve the performances of the memories in the presence of variability conditions.

II.1 Variability aspects

Fluctuations in the performances of ICs seem to be a challenging problem for designers, especially when dealing with the uncertainty associated with variability. Indeed, these variations have caused circuits to deviate from their original specifications at which they were meant to operate. As stated previously, this can be due to manufacturing process conditions (P), to environmental conditions i.e. voltage and temperature factors (V, T) and to time factor. Figure II.2 summarizes the classification of parameter variations

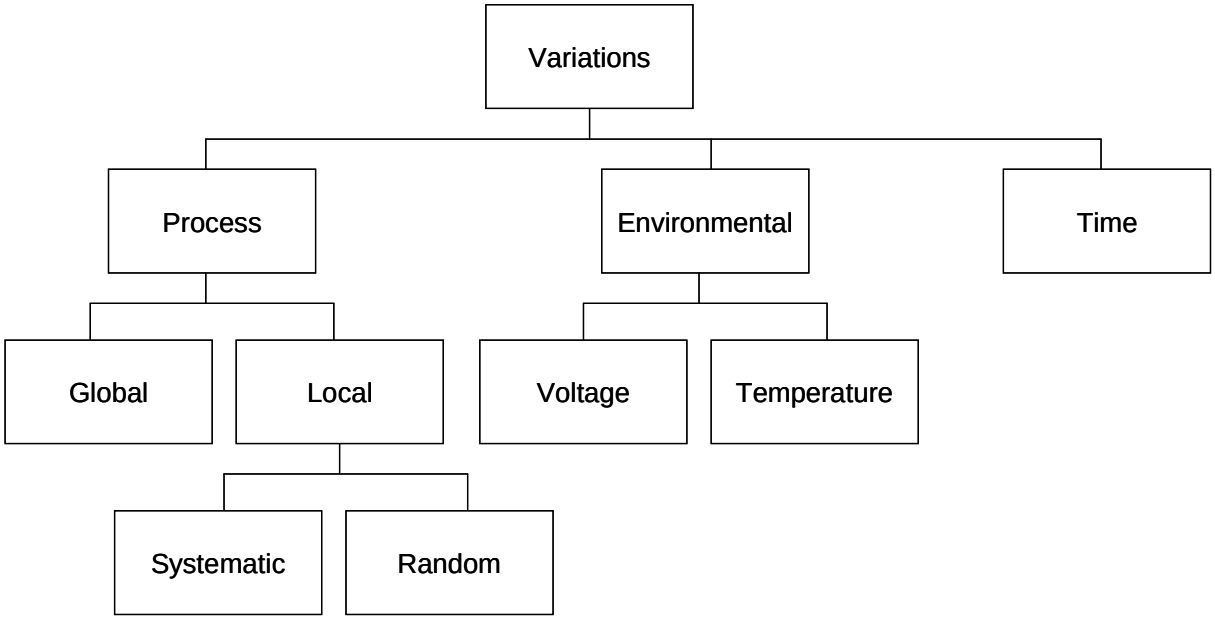

Fig. II.2 Classification of parameter variations

II.1.2 Environmental factors

II.1.2.1 Temperature variations

Very often, different blocks on the chip have their own specific functionalities and show different switching activities, resulting in higher heat flux and temperature variations. This condition gives rise to the appearance of hot spots. These hot spots are in turn responsible for the existence of temperature gradients across the chip, which cause important differences in temperature between different areas of the die as depicted in figure II.3 [Bor03]. The different regions in the chip can reach 40˚C to 50˚C differences, for example between the cache and the core.

Hence, higher temperature variations mean that the performance of the chip is degraded. For instance, an increase in temperature enhances leakage current like subthreshold current and power consumption, leading to functionality and reliability problems [Bor03].

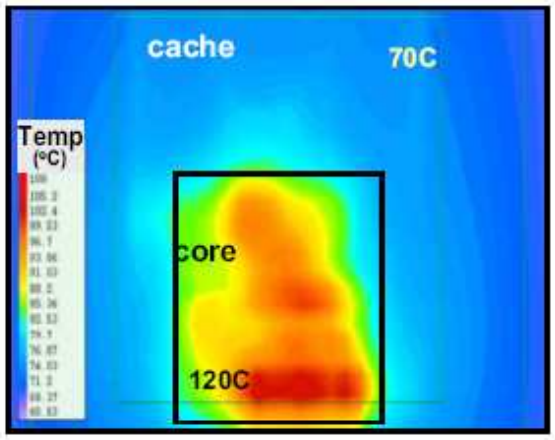

Fig. II.3 Within die temperature variations

Temperature fluctuations also account for interconnect delay of different signal and clock paths (clock skew) on the critical path signal distribution. Clock skew can severely limit the maximum operating frequency of the chip and give rise to catastrophic race conditions, ending up with setup or hold time violations [Tsa93].

The impact of test induced hot spots should also be considered [Zho06], since the thermal map distribution between normal and test mode operations is different. A higher activity during test mode is responsible for significant temperature fluctuations. Considering this condition, non critical path may slow down due to test induced hot spots making the die to fail delay testing for a good part. Significant yield loss is inevitable as normal operation thermal map impact on path delay is often different from that during testing mode.

As far as temperature variation effect is concerned in the increase of propagation delay of the circuit, this factor originates from the decrease of carrier mobility with temperature rise. In fact, the two main parameters in MOS devices that are mostly affected by temperature are the effective mobility (μ) of the electrons and the threshold voltage (V_T) given by equations (II.1) and (II.2) [Seg04]:

$$\mu(T) = \mu(T_r)\left(\frac{T}{T_r}\right)^{-m} \qquad \text{(II.1)}$$

where T is the absolute temperature, T_r is the room absolute temperature and m is a constant lying between 1.5 and 2.0.

$$V_T(T) = V_T(T_r) - K(T - T_r) \quad \text{(II.2)}$$

where K is the threshold voltage temperature coefficient usually between 0.5mV/K and 3mV/K.

An illustration of the effects of temperature on effective mobility and threshold voltage on the drain current I_{DS} of the MOS device are clearly understood by analyzing the drain current of the device in saturation mode [Tsi99]:

$$I_{DS} \propto \mu(T).[V_{GS} - V_T(T)]^2 \quad \text{(II.3)}$$

At high V_{GS}, an increase in temperature leads to a simultaneous decrease in the effective mobility and the threshold voltage parameters. The decrease in carrier mobility tends to decrease I_{DS} whereas a decrease in V_T tends to increase I_{DS}. However, at high V_{GS}, the carrier mobility is predominant and therefore leads to a decrease of drain current and an increase in the gate delay as temperature increases. This is why the delay associated with the read access time of an SRAM is more important at higher temperatures than at lower temperatures [Che06].

On the other hand at low V_{GS}, when temperature increases, the effective mobility and threshold voltage both decrease. The decrease in carrier mobility will tend to decrease drain current, whereas a decrease in threshold voltage will tend to increase the drain current. At low V_{GS}, the threshold voltage is predominant and hence the drain current increases as temperature increases.

In certain conditions a certain value of V_{GS} can be found at which the current becomes practically temperature independent over a large temperature range [Tsi99]. This condition is known as the temperature independence point (TIP).

Figure II.4 [Las07] displays the temperature delay sensitivities of several paths, composed of several inverted cells (> 5), with respect to voltage variations. In [Las07], the authors represented 3 distinct regions of the delays measured, depending on the temperature delay sensitivities of the cells under certain operating voltage conditions.

In the first domain (I), between 0.8V and 0.94V, the cells exhibit a negative temperature coefficient. This means that in I, temperature inversion phenomenon occurs for N and P devices i.e. the worst case performances occur at -40˚C.

In the second domain (II), between 0.94V and 1.05V, positive temperature coefficient values are displayed by N devices whereas negative temperature coefficient values are shown by P devices. This range is characterized by weak positive and negative temperature coefficient values. Thus, domain II is less sensitive to temperature variations such that when temperature values are altered, the timing performances of the circuits are not significantly affected.

In the last domain (III), from 1.05V onwards, the temperature coefficient values are all positive. This implies that worst case operating temperature condition occurs at 125˚C.

Another interesting conclusion [Las07], which can be drawn between regions I and III, is that the absolute value in the temperature delay sensitivity in III is weaker than in I. In fact, the propagation delay path at 1.2V does not vary by more than 15% between -40°C and 125°C, whereas at 0.8V, the propagation delay in the cells can reach 30%. It can be inferred that high operating voltages minimize the delay sensitivities of the cells to temperature gradients.

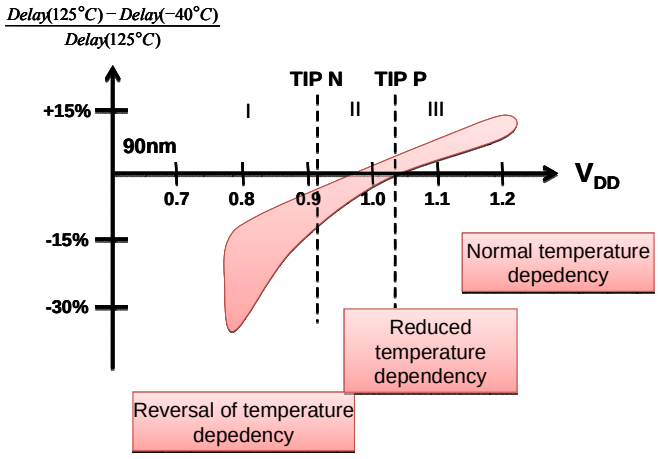

Fig. II.4 Temperature delay sensitivity variations with respect to voltage variations

II.1.2.2 Voltage variations

Voltage variations originate from IR drop, switching activity of different areas of the memory and non uniform power supply distribution [Hum06]. The resistance of the power network in a chip causes rail voltage drop or IR drop due to the current and resistance associated with the power network. In fact, IR drop is more important in areas found at the centre of the chip

compared to areas found near the periphery of the die, owing to the length of the wires. As a result, blocks found at the centre operate slower than those found near the periphery. Hence, this effect will impact on the overall timing performances of a circuit and cause uneven power dissipation. For e.g. a memory cell found at the centre of the memory core will experience a larger IR drop and operate at a slightly lower voltage, compared to the same cell found at the periphery, although a uniform power supply voltage is being delivered.

II.1.3 Aging process

The aging of devices and the progressive degradations of their electrical properties appear over time because of the extreme conditions under which these transistors operate for instance high temperature conditions and dynamic voltage fluctuations. Indeed, two important examples that underpin the degradation performances include the negative bias temperature instability (NBTI) [Kum06] and the gate oxide break down [Pap07].

NBTI happens in p-channel device, due to the generation of interface traps, when PMOS devices are stressed with negative gate voltages at elevated temperatures [Sch03]. It results in an increase in the threshold voltage with time and a reduction in the drive current.

The origins of NBTI can be explained as follows [Kum06]. When Si is oxidized, Si atoms bond to oxygen atoms. It can happen that Si atoms also bond with hydrogen, giving rise to weak Si-H bonds. When PMOS devices are biased in inversion, holes in the channel dissociate the Si-H bonds. The hydrogen diffuses into the oxide, generating interface traps. These interface traps are electrically active physical defects and manifest themselves as an increase in the threshold voltage of the PMOS device. In [Luo07], the authors report that NBTI worsens exponentially with thinning gate oxide, and threshold voltage shifts of the order of 20-50mV are serious for devices operating at 1.2V or below. Threshold voltage degradation can be expressed as [Pau07]:

$$\Delta V_{th}(E_{ox}, t) = (1 + m)\frac{qN_{IT}}{C_{ox}} \qquad \text{(II.4)}$$

where m is a constant representing the mobility degradation caused by interface traps, q is the electronic charge, N_{IT} is the trap density and C_{ox} is the oxide capacitance.

NBTI effect on PMOS leads to performance problems and functional failures in circuits due to the timing variations of the signals upon their arrivals at different blocks. For example, a read failure can occur if the word line in the row decoder is activated for a shorter time, such

that sufficient time is not allowed for the appropriate bit line to get discharged. In [kum06], the effect of NBTI has been studied on SRAM stability in 100nm and 70nm devices. It has been shown that after approximately 3 years of device stressing, NBTI induces around 8% degradation in the SRAM cell stability. However, NBTI phenomenon seems negligible for a positive gate voltage applied to the PMOS and for either a positive or negative voltage applied to the NMOS device [Sch03].

Gate oxide break down, characterized by the time-dependent dielectric breakdown (TDDB), is a problem which is more frequently encountered in devices as the gate oxide thickness scales down, while the supply voltage is not scaling accordingly [Pap07]. This condition generates an electric field across the oxide and hence leads to gate oxide tunnelling current (soft oxide breakdown). An investigation realized by H. Wang et al. [Wan06] on SRAM components illustrated that gate oxide break down, occurring in NMOS devices found in the sense amplifier, causes a speed degradation of this block by as much as 22% and a higher power consumption amounting to 36%.

II.1.4 Process Variations

Process variations appear at different levels during the manufacturing steps and have emerged as a serious bottleneck for the proper design of ICs in the sub nanometre regime. Usually, these process imperfections result from factors such as processing temperature, equipment properties [Bor03], and at the device level they are caused by random dopant fluctuations and line edge roughness [Bhu07]. These conditions lead to variations in transistor parameters like threshold voltage fluctuations, thereby influencing the overall performance of the chip. Generally, process variations can be classified into 2 distinct groups i.e. global and local variations which affect the chips differently. In this sub section, we will explain the existing differences between global (inter die to inter fab) and local variations (intra die) as summarized in figure II.5, which shows that variation between circuit increases as their distance at process time increases [Cro05].

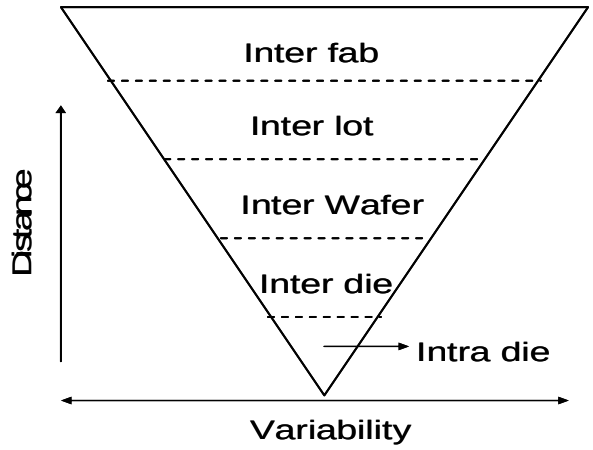

Fig. II.5 Variability at several levels

Transistor matching has been extensively studied by M. J. Pelgrom et al. [Pel89] on threshold voltage, current factor and substrate factor and their dependence on the area, distance and orientation of the device. They showed that the parameter variance ΔP between two rectangular devices for long and short correlation distance variations can be expressed as:

$$\sigma^2(\Delta P) = \frac{A_p^2}{WL} + S_p^2 D_x^2 \tag{II.5}$$

A_p is the area proportionality constant for parameter P, S_p describes the variation of P with spacing, D_x describes the spacing between the devices and WL represents the device area.

II.1.4.1 Global variations

Global variations deal with inter-fab, inter-lot, inter-wafer and inter-die fluctuations (figure II.6) and are said to affect every element on a chip equally [Bow02] or in a systematic way.

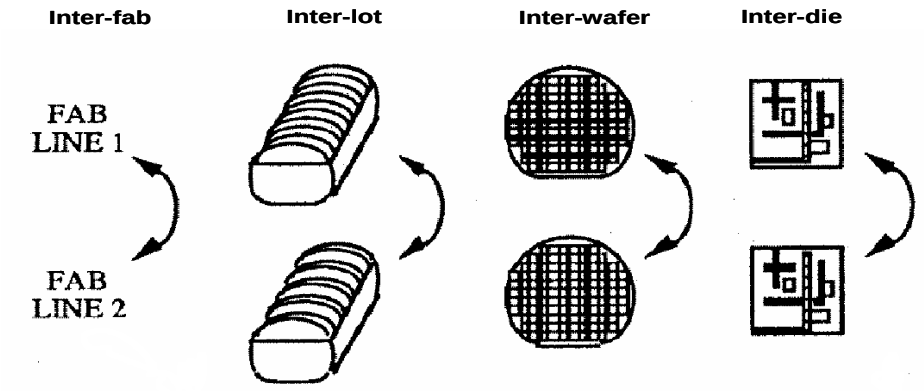

Fig. II.6 Types of global variations

Process parameters of all transistors on a die will either have for example a high threshold voltage (Vth) or a low Vth. These variations originate from a myriad of factors. For instance, different lots are not processed using the same machines, or on a single wafer differences might exist between dies due to a slightly different processing temperature between the edge of the wafer and its center [Cro05]. Inter die variations can also stem from variations in oxide thickness, channel length, channel width and device orientation.

II.1.4.2 Local variations

Local variations or mismatch, also known as intra-die, within-die, on-die or across chip variations, are characterized by differences between supposedly identical structures found on the same die [Cro05] but this time, these variations display either a systematic or a random behaviour. Systematic variations result from mechanical strain variation, lens aberrations, photo mask size differences, corner rounding and non uniform chemical mechanical polishing (CMP) that occurs due to different pattern densities. On the other hand, a non exhaustive list of random errors include random dopant fluctuations, line edge roughness, surface charge effect and ion implantation.

As in systematic global variations, systematic local variations means that neighbouring transistors are spatially correlated such that if a device experiences a shift in its Vth, all other transistors on the same die will have that parameter moving in the same direction. For its part, random variations will cause neighbouring transistors to behave differently, implying that the shift in a parameter of two neighbouring transistors is completely independent [Muk05].

II.1.4.2.1 Systematic within die variations

II.1.4.2.1.1 Lens aberrations

The systematic within die variations, caused by lens aberrations, appear during the IC lithography process and induce a deviation of an image from its ideal one. The aberration can be represented by figure II.7 [Cho07], whereby an aberration plate is introduced to model the distortion. In the IBM Journal of Research and Development, T. A. Brunner [Bru97] sets the problem by first defining the optical path of a point object as the distance along the ray emitted by a light source multiply by the local refraction index of the lens. By definition, he stated that lens aberrations are said to occur when different rays have different optical paths

and these aberrations generate an optical path difference (OPD) for each particular ray. The different types of lens aberrations that have a deleterious impact on lithographic imaging include for example image shift, which corresponds to the positional shift of the image in the plane of a wafer

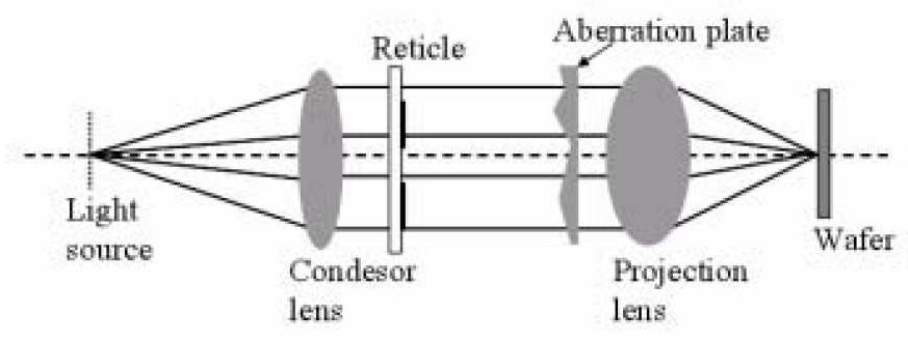

Fig. II.7 Schematic diagram for representing an aberrated lens

Moreover, lens aberrations have also demonstrated to have a critical influence on the timing of a cell and on the overall circuit. Due to lens aberrations, the average delay in a 90nm cell can fluctuate between 2% to 8% [Kha06].

II.1.4.2.1.2 Chemical Mechanical Planarization

Chemical Mechanical Planarization (CMP) is a process used for the realization of multilevel interconnects in high density CMOS circuits [Ouy00]. The procedure is typically carried out using the rotary CMP tool as depicted in figure II.8 [Ick].

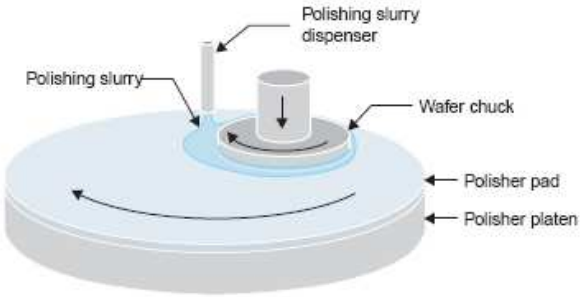

Fig. II.8 A rotatory CMP tool

The wafer is held on a rotating carrier while the surface that requires the polishing is pressed against the polisher pad fixed to a rotating polisher. The slurry, used as a chemical abrasive, is allowed to flow through the dispenser, onto the pad and it reaches the wafer through the porosity of the pad before chemically attacking the wafer surface.

We see that the patterned metal layer beneath shapes the volume of the protruding dielectric material which needs to be polished. The figure features two regions of metal patterning, one with closely spacing metal lines representing a denser region compared to the region where metal lines are spaced further apart [Ouy00]. In the denser regions, there is a larger deposit of oxide than in sparse area, such that the removal rate of the oxide during CMP for a specific time gives rise to ILD thickness variations.

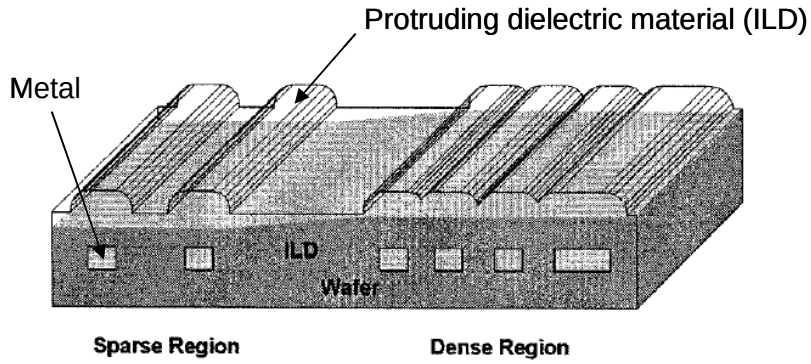

Fig. II.9 Non uniform deposit of inter layer dielectric due to the underlying metal pattern density

II.1.4.2.2 Stochastic within die variations

II.1.4.2.2.1 Random errors due to surface state charge

In MOS devices, parasitic charge exists within the oxide (SiO2) and at the interface of the oxide/semiconductor as shown in figure II.10. These charges can be described as follows [Tsi99]:

- Interface State Charge existing at the oxide/semiconductor interface results from defects present at that interface.
- Oxide fixed charge appears close to the oxide/semiconductor interface during the mechanism of oxide formation.
- Oxide trapped charge which is acquired through radiation, photoemission, and injection of carriers from the substrate.

- Mobile ionic charge is due to the contamination by alkali ions (sodium) introduced during manufacturing steps.

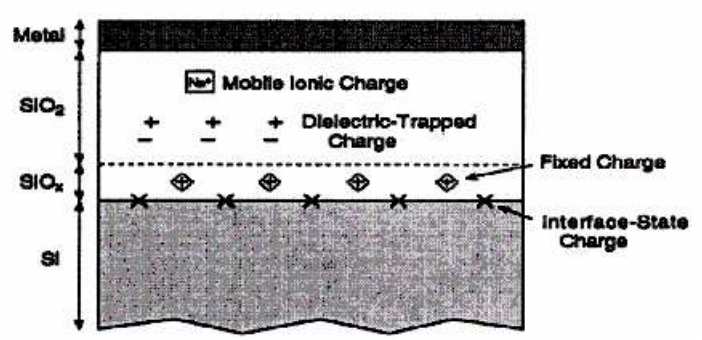

Fig. II.10 Parasitic charges within the oxide and at the oxide/semiconductor interface

Among these charges, the interface state charge (Qss) can affect the threshold voltage as the latter deviates from its mean value causing a shift in the nominal threshold voltage (Vth). The dependency of Vth on Q_{ss} is given by the following equation [Shy84]:

$$V_T = \Phi_{MS} + 2|\phi_F| + \frac{\sqrt{2q\varepsilon_{Si}N_a(2|\phi_F| + V_{BS})}}{C_{ox}} - \frac{Q_{ss}}{C_{ox}} \quad \text{(II.6)}$$

where Φ_{MS} is the potential difference of the work functions of the gate and the substrate, Φ_F is the built in potential in the bulk of the substrate, ε_{Si} is the permittivity of silicium, N_a is the bulk density, V_{BS} is the source to bulk body bias voltage, Q_{ss} is the surface state charge density per unit area, C_{ox} is the gate capacitance per unit area.

II.1.4.2.2.2 Random errors due to line edge roughness

Line edge roughness (LER), which can be considered as the random deviation of the gate line edges from their ideal forms, is normally induced by imperfections appearing during the lithography process. For many years, the semiconductor industry has been able to print feature sizes properly without worrying about LER, as far as the silicon feature size was above the lithography wavelength (figure II.11) [Mcp06].

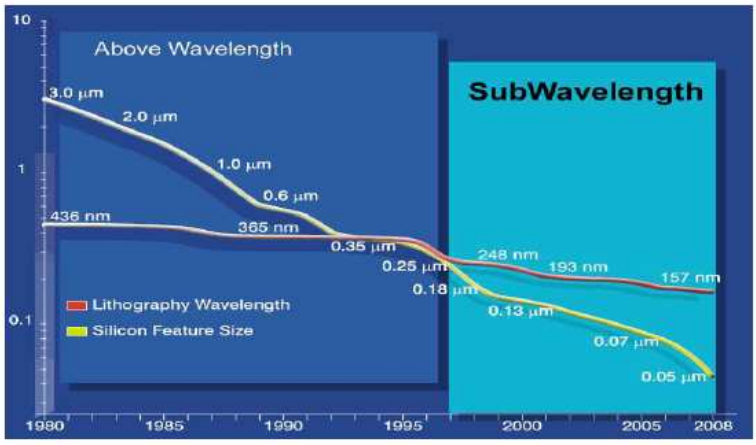

Fig. II.11 Evolution comparison between lithography wavelength and silicon feature size

However, as transistor sizes started to shrink below the wavelength used i.e. moving towards the sub wavelength region, LER started to become a severe problem and is expected to worsen. In fact, LER does not scale accordingly thereby representing a significant proportion of the gate length. This is shown in figure II.12 [Ase03]

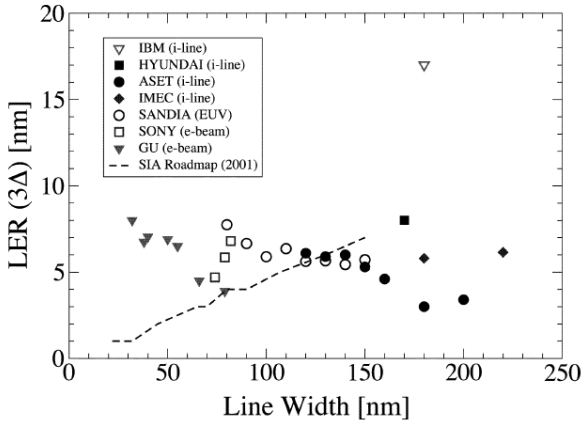

Fig. II.12 Data from various advanced lithography processes reported by different labs

Figure II.12 reports the variations of the LER with respect to the decrease in the line width of transistors, with Δ representing the rms amplitude of the edge roughness from different labs. It can be clearly seen that the value of the LER does not decrease with line width reduction according to the semiconductor industry roadmap (SIA) requirements, but oscillates around 5nm, independently of the type of lithography used in production or research.

Investigations [Ase03] have shown that LER induces a fluctuation (σV_T) in the threshold voltage V_T and a lowering of the threshold voltage ($<V_T>-V_{T0}$) compared to the threshold

voltage V_{T0} of a generic device with straight gate edges. This is represented in figure II.13 for an effective gate length (L_{eff}) of 30nm and 50nm, with the inset figure representing the average threshold voltage lowering. It can be seen that the threshold voltage variations increase with drain voltage (squares represent V_D=1.0V and circles represent V_D=0.1V). Moreover, for a given LER, both the threshold voltage fluctuations as well as the threshold voltage lowering increase as transistor dimensions are reduced. Consequently, LER fluctuations affect the electrical parameters like the drive current and the subthreshold current [Old00, Kay01].

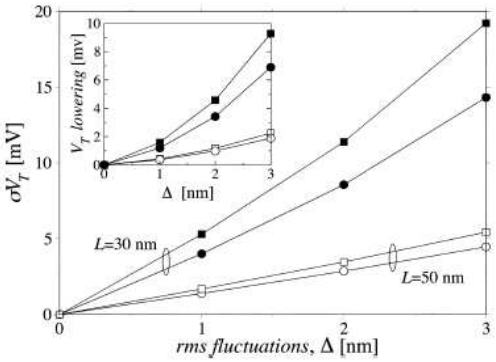

Fig. II.13 Standard deviation of V_T for 2 different L_{eff} and V_T lowering with increase of LER at drain voltage V_D= 1.0V (squares) and V_D=0.1V (circles)

II.1.4.2.2.3 Random errors due to random dopant fluctuations

The study of random dopant fluctuations and their impacts on the electrical behaviour of MOSFETs have been the subject of numerous publications [Won93, Miz94]. This intrinsic behaviour of MOS devices is due to the random nature of ion implantation, dopant diffusion and other processes involved in the doping of silicon device areas, which leads to stochastic variations of the channel dopant number [Sto98]. In fact, random dopant distributions account for the variations in the threshold voltage σV_T of devices and can be expressed analytically, according to [Miz94] by the following expression:

$$\sigma V_T = \left(\frac{\sqrt[4]{4q^3 \varepsilon_{Si} \phi_B}}{2} \right) \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \frac{\sqrt[4]{N}}{\sqrt{W_{eff} L_{eff}}} \qquad \text{(II.7)}$$

where q represents the elementary charge, ε_{Si} and ε_{ox} are the permittivity of silicon and oxide, $\Phi_B = 2K_B T \ln(N/n_i)$ (with K_B representing the Boltzmann's constant, T the absolute

temperature, N the channel dopant concentration and n_i the intrinsic carrier concentration), T_{ox} is the gate oxide thickness and W_{eff} and L_{eff} are the respective effective channel width and length.

However, a better understanding of the threshold voltage shift can only be achieved by using 3D atomistic simulation study as it has been performed by A. Asenov in 1998 [Ase98]. For the first time, the latter provided a systematic analysis of random dopant effects down to an individual dopant level in 3D. In his study, A. Asenov explained that the analytical model takes only into account the fluctuations of the number of dopants in the depletion region but the atomistic simulations display a higher level of fluctuations in the threshold voltage. To analyse this discrepancy, the correlation between the threshold voltage and the number of dopants in the depletion region for a sample of 2500 transistors with an effective length L_{eff} of transistor of 50nm has been plotted as shown in figure II.14.

It can be clearly seen that even with an equal number of dopants in the depletion region, the threshold voltage (V_T) varies over a wide range. This difference can only be due to the microscopic arrangements of dopants within the depletion layer, which in fact is not taken into account by the analytical model.

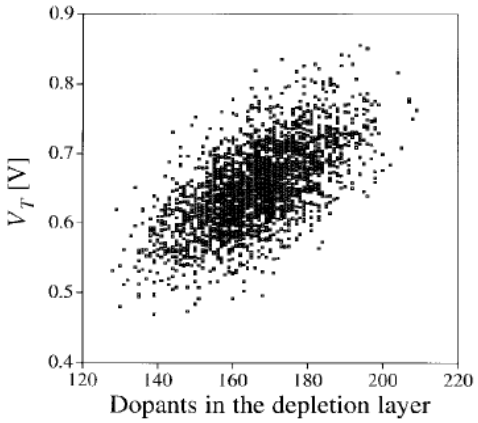

Fig. II.14 Correlation between V_T and the concentration of dopants in the depletion layer

A clear illustration of this situation is given in the figures II.15(a) and II.15(b), where the author compared the potential distributions at the Si/SiO_2 interface of two microscopically MOSFETs, each having the same number of dopants (170) in the channel depletion region but with different threshold voltages. Transistor in figure II.15 (a) has a threshold voltage of 0.78V whereas the device in figure II.13 (b) has a threshold voltage of 0.56V. As transistor in

figure II.13 (a) has six to seven dopants in the middle of the channel which blocks the current path, this results in a higher V_T whereas device in figure II.13 (b) has virtually no dopants at the surface, hence a smaller V_T.

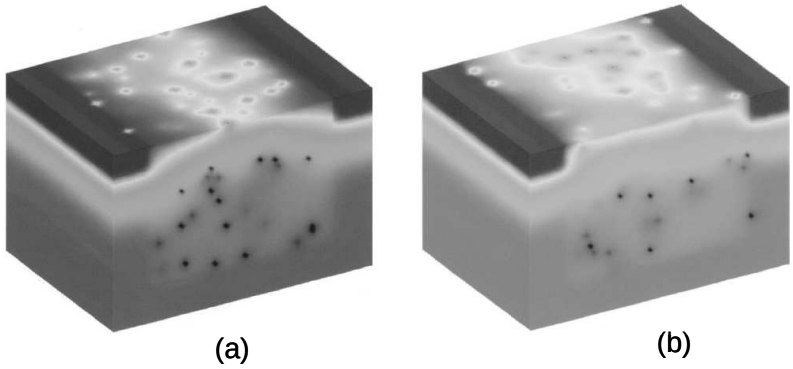

Fig. II.15 Potential distribution at Si/SiO_2 interface of a MOSFET with (a) V_T=0.78V (b) V_T=0.56V

II.1.4.2.2.4 Random errors due to gate depletion

Doping the polysilicon gate, at different concentration levels, is responsible for the degradation in the performance of MOS devices. This is due to the formation of a depletion layer near the polysilicon/SiO_2 interface which is referred to as the polysilicon depletion or polydepletion effect [Aro95]. The depletion layer introduces a depletion capacitance in series with the oxide capacitance of the SiO_2, causing an increase in the total equivalent capacitive oxide along with a fluctuation in the effective gate oxide thickness. An approximation of this equivalent increase in oxide thickness (t_{GD}) has been established by [Cro05] and can be expressed as follows:

$$t_{GD} = \sqrt{\left(\frac{t_{ox}}{2}\right)^2 + \frac{\varepsilon_{ox}{}^2\left(V_{GS} - \phi_{MS} - \phi_B\right)}{\varepsilon_{si} q N_p}} - \frac{t_{ox}}{2} \qquad \text{(II.8)}$$

where t_{ox} is the oxide thickness, ε_{ox} and ε_{si} are the permittivity of oxide and silicon , V_{GS} is the gate bias, Φ_{MS} and Φ_B represent the work function and the surface potential in strong inversion, N_p is the doping concentration in the poly gate at the interface of the oxide and q is the elementary charge. J. A. Croon et al. [Cro05] then represented the induced threshold voltage variation caused by fluctuations in the oxide thickness by the following equation:

$$\sigma V_T = \frac{\left(t_{ox} + t_{GD}\right)\sqrt[4]{2q^3 \varepsilon_{si} N_A \psi_S}}{\varepsilon_{ox}\sqrt{3WL}} \qquad \text{(II.9)}$$

where Ψ_S represents the surface potential and N_A is the channel doping concentration.

II.1.4.2.2.5 Random errors due to halo implants

In deep submicron devices, threshold voltage reduction due to short channel effect is partly compensated by the use of local channel implants or pockets. Halo doping reduces the charge-sharing effects from the source and drain fields. In the presence of such condition, the width of the depletion region in the drain-substrate and source-substrate regions is decreased [Muk05]. Consequently, this mechanism diminishes the threshold voltage degradation effect, although the sizes of transistors decrease (Reverse Short Channel Effect).
These halo implants, which are generally located near the source and drain regions as shown in figure II.16, are not supposed to interfere with the matching properties of devices.

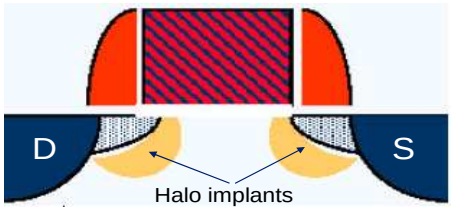

Fig. II.16 Halo implants at drain/source regions

However, J. A. Croon et al. [Cro05] observed that the mismatch of transistors increases with halo implant realizations. A rational explanation they attributed to this effect would be that, during the implant process, the gate does not act as a perfect mask and part of the halos are implanted through the gate. Implant ions used like boron or arsenic found them trapped at the gate side or the channel side as depicted in figure II.17.

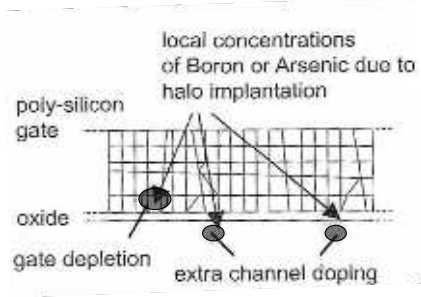

Fig. II.17 Schematic drawing of a MOSFET with localized regions of charge due to halo implants

The localized regions of charges in the polysilicon area give rise to gate depletion as explained previously, whereas charges found in the channel account for the variation of dopants found in the channel. Hence, the overall effect corresponds to variations in the gate oxide thickness and threshold voltage variation.

It should also be noted that when halo implants are performed, the implantation angle of the ions seems also to be a critical factor in the determination of the sensitivity of the transistor parameters. A small tilt angle variation could be responsible for a significant portion of the total allowed V_T variation [San03]. Simulation results carried out by [Adi01] reported that in a 0.18µm device, a change by ± 2° tilt angle variation during halo implant process induces a V_T shift of 5% and an increase of 60% in the leakage current.

II.2 Failures in SRAM

II.2.1 Types of defects

The sources of variations describe in the previous section alter the performances of IC and can lead to failures in the memory. These failures are classified as being either catastrophic or parametric.

II.2.1.1 Catastrophic failure

Most catastrophic or functional failures are due to material flaws and local disturbances, such as particle wafer contamination and spot defects [Zha95]. Spot defects result in opens or shorts in the circuit's connectivity. Opens and shorts refer to a region with a missing or an extra material in one of the conductive or semi conductive layers as shown in the figure II.18 [Mal87].

Figure II.18 (a) depicts a break which occurs in 7 metal lines caused by the interaction between a contaminating particle and the photoresist during the lithographic operations. On the other hand, figure II.18 (b) represents a short caused by a particle deposit on the surface of the photoresist before the exposure step.

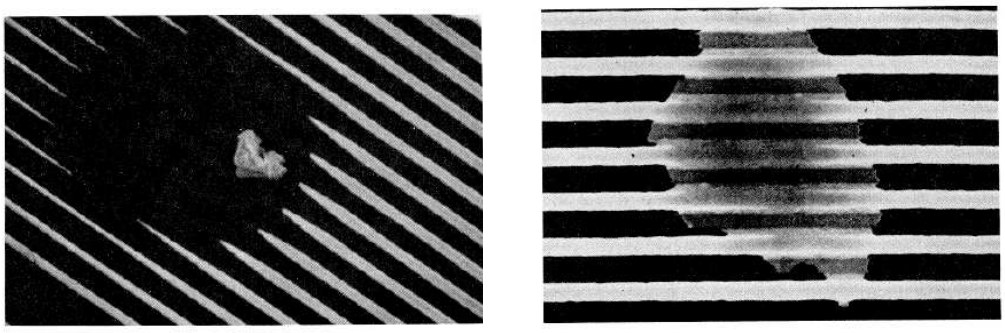

Fig. II.18 (a) Break of 7 metal lines (b) Short of 7 metal lines

As it can be seen, the 'footprint' of the particle is replicated in the metallization layer. Globally, catastrophic failure changes the basic functionality of the circuit.

II.2.1.2 Parametric failure

Parametric failures can be defined as failures due to the variation in one set or a set of circuit parameters such that their specific distribution in a circuit makes it fall out of specification [Seg04]. In other words, the components appear to function but may not be within the desired tolerant limits. These types of failures originate from environmental factors which include voltage and temperature variations as explained in section II.1.2 and from physical variations inherent in the circuit manufacturing steps as detailed in section II.1.4. Parametric failures are very common in SRAMs and they affect significantly the performance of the memory by causing failures in the memory cell, offset problems in the sense amplifier block and failure mechanisms in the address decoder block. This is why we will focus on parametric failures and analyze their effects on SRAMs.

II.2.2 Impact of parametric failures on SRAM performance

Transistor parameter variations, for e.g. channel length and width variations, effective gate oxide thickness variations and threshold voltage fluctuations, result in the mismatch of transistors within the 6T SRAM cell. Thus, the SRAM cell is prone to several types of parametric failures. Some of which can be resumed as follows [Muk05, Aga06]:

(i) Read stability failure

(ii) Write failure

(ii) Hold failure

Mismatch of transistors found in other blocks for e.g. in the sense amplifier can also result in offset problems within the structure and lead to parametric failures. The next subsections describe each of the above stated effects.

II.2.2.1 Read stability failure

Read stability failure occurs during the discharge process of BL or BLB, depending whether a '0' or '1' is read from the cell. Generally, the stability of the memory cell is expressed by its static noise margin. It corresponds to the maximum static noise, caused by disturbances such as offsets and mismatches due to processing and variations in operating conditions, which can be tolerated by the flip flop before changing states [See87]. Consider the following memory cell in figure II.19 storing a '0'.

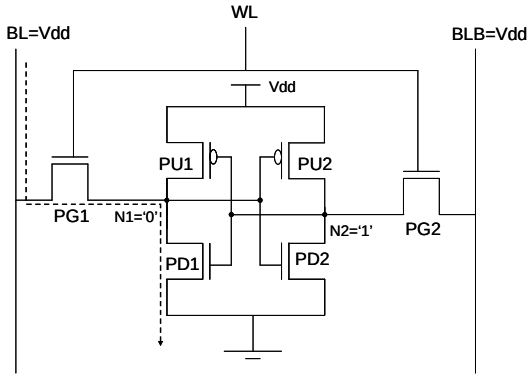

Fig. II.19 Reading a '0' from an SRAM cell

When word line WL is activated and BL begins its discharge, internal node N1 is exposed to a voltage disturbance caused by the resistive voltage division action between pass gate transistor PG1 and pull down transistor PD1. Typically, if voltage node N1 is above the trip point voltage of transistors PU2 and PD2, the state of the cell flips and leads to a read failure. This is represented in figure II.20 by the dotted lines.

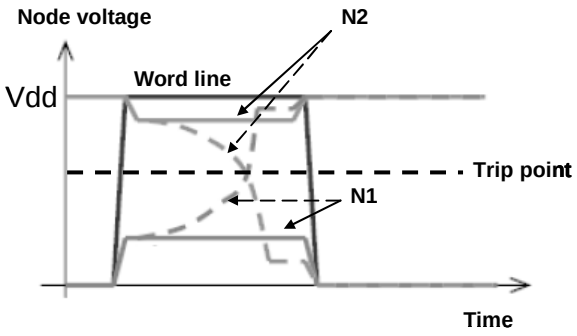

Fig. II.20 Read failure of an SRAM cell

This is why the driving strength of PD1 is made stronger than PG1, and it is characterized by a β ratio [Muk05] given in (II.10), which determines the cell stability:

$$\beta = \frac{\beta_{PD}}{\beta_{PG}} = \frac{\frac{\mu_{eff} C_{ox} W_{PD}}{L_{PD}}}{\frac{\mu_{eff} C_{ox} W_{PG}}{L_{PG}}} \qquad \text{(II.10)}$$

where μ_{eff} represents the effective mobility, C_{ox} is the oxide capacitance, W_{PD} and W_{PG} are the widths of the pull down and access transistors and L_{PD} and L_{PG} stand for the lengths of the pull down and access transistors. In general, the value of β is around 1.5 as the sizes of PG is made smaller compared to PD.

However, effects due to threshold voltage variations can alter the driving strength of PG and PD transistors by decreasing the threshold voltage of PG and increasing the threshold voltage of PD. This causes node N1 to increase from its nominal value, thereby flipping the contents of the cells.

II.2.2.2 Write failure

Write failure is said to occur when the memory cell is unable to write properly the desired state in it. To illustrate this operation, let us assess what happens in the SRAM cell when a '0' needs to be written in the cell storing a '1' (figure II.21). During a write operation, BL is set at '0' before activating WL. Once WL is triggered, node N1 which is at VDD gets discharged through PG1. This time, the discharge process depends on the voltage division action established between PU1 and PG1. If the voltage node N1 cannot be discharged below the trip point voltage of PU2 and PD2 within the duration of WL, a write failure is said to occur (the dotted lines in figure II.22 represent the write failure). This condition is achieved by making

PU1 weaker (increase in threshold voltage) than PG1, so that the internal node N1 can be brought closer to '0' during the write operation.

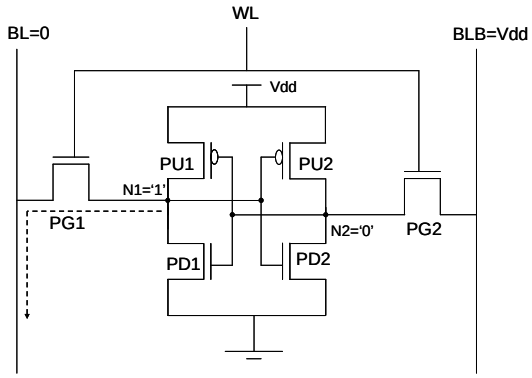

Fig. II.21 Writing a '0' to an SRAM cell

Threshold voltage variations can cause the threshold voltage of PG1 to increase and that of PU1 to decrease. Hence, node N1 is not given sufficient time to decrease below the trip point voltage of PU2 and PD2 before the deactivation of WL, leading to a faulty write operation. The write failure condition is shown by the dotted lines in figure II.22.

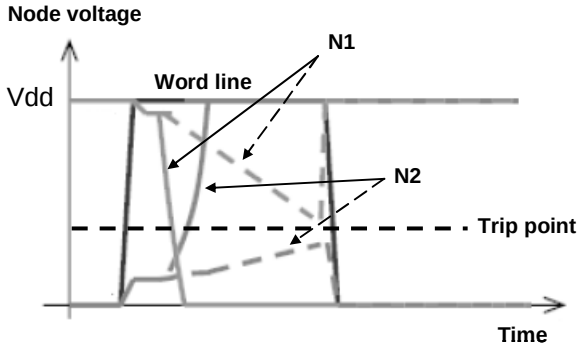

Fig. II.22 Write failure of an SRAM cell

II.2.2.3 Hold failure

Hold failure occurs in standby mode of the memory [Muk05]. Stand by mode, which consists in lowering the supply voltage Vdd of the memory by a certain voltage Δ, is adopted in order to minimize leakage current in the SRAM cell when the memory is not being accessed, with the aim of maintaining the stored data. Nevertheless, when vdd is being reduced, node N1 is now at a lower potential. The internal node voltage also starts to degrade due to the presence of leakage currents represented in figure II.23, namely gate leakage and subthreshold leakage.

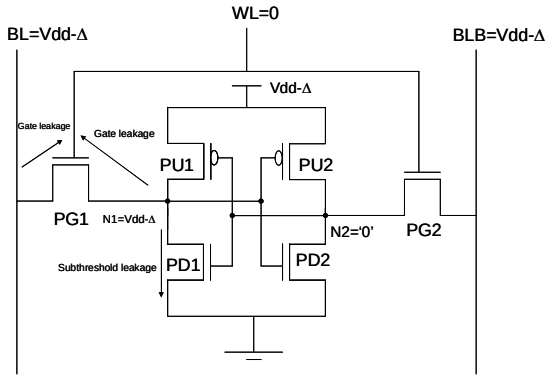

Fig. II.23 Leakage currents in stand by mode degrading voltage node N1

Moreover, random dopant fluctuations can cause the threshold voltage of PD1 to decrease and that of PU1 to increase. If voltage node N1 decreases below the trip point voltage of PU2 and PD2, the state of the cell flips and a hold failure is said to occur.

II.2.2.4 Mismatch in sense amplifier

Transistor mismatch in the sense amplifier is an important factor which results in the parametric failures of memory. In fact, the mismatch between the two transistors MN0 and MN1 (figure II.24) has a dominant impact on the offset voltage of the sense amplifier. The offset voltage can be defined as the minimum differential voltage between BL and BLB that is required for the correct output of the latch [Sin04].

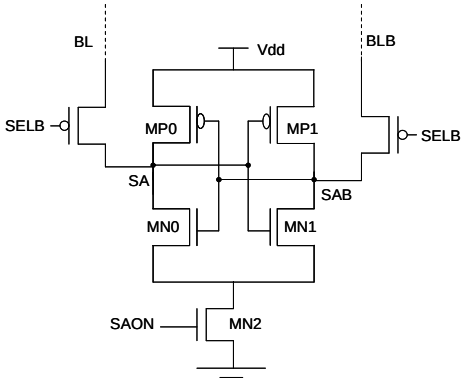

Fig. II.24 Latch type sense amplifier

Unbalanced higher threshold voltages in MN0 and MN1 impact on the discharge time of node SA or SAB and increase the offset voltage. For example, in reading a '1' from a memory cell, BL is being discharged before activation of SAON. As explained previously in chapter 1, the

voltage difference between BL and BLB should be around 10% of VDD before triggering MN2 and discharging node SA, initially at a voltage less than VDD. If MN0 has a higher threshold voltage, the discharge rate of SA will be longer than expected, such that the delay for the required differential voltage to develop between BL and BLB increases. This condition slows down the sense amplifier and the overall read operation of the memory. In this case, the memory consumes more dynamic power than expected.

II.2.2.5 Failure mechanism in address decoder

Row and column address decoders can suffer from process variations. The variation in threshold voltages, due to local variations in the transistors found for example in the row decoders, can maintain word line activated for a shorter time. This means that during read or write operations, shorter time will be allocated for reading or writing data in the memory cells, resulting in read and write failures.

II.3 Solutions for controlling variability in the memory

In chapter I, some of the main challenges encountered by SRAMs in VDSM technologies have been described, one of which includes memory yield challenge. As we have seen in section II.1, embedded memories are vulnerable to variability conditions (PVT variations) which decrease the memory's yield. To circumvent the problem of variability and achieve a better functionality of the memory, numerous techniques have been adopted related to the memory cells but also to the memory architecture. Some of the prevalent techniques used in the memory cells include pulsed technique and dynamic variation of supply voltage. On the other hand, non related memory cells design techniques are composed of replica technique, redundancy scheme and error correcting code (ECC).

II.3.1 Memory cell design techniques

II.3.1.1 Pulsed techniques for read and write margin

Pulsed word line (PWL) and pulsed bit line (PBL) techniques have been proposed by M. Khellah et al. [Khe06] to improve SRAM cell stabilities and failures at low operating voltages (0.8V) in 65nm process. They also used the pulsed read-modify-write (RMW) scheme with the pulsed word line method so as not to degrade the write margin.

In the pulsed word line scheme for read operation, word line is activated only for a short period of time so that the internal nodes of the cells are isolated from bit lines before the nodes begin to flip. Moreover, sufficient care is taken so that the minimum required input differential voltage of the sense amplifier has been developed before switching off word line.
Nevertheless, the problem with pulsed word line method arises during a write operation. In fact, the minimum activation time of word line degrades the write margin. To cope with this problem, the read stability and the write margin constraint need to be balanced. This is achieved by using a pulsed read-modify-write scheme (RMW), which is a two step procedure in a write operation, as depicted in figure II.25 [Khe06]:

(i) A selected column is read by the sense amplifier (SA), connected to each column of bit lines, using the PWL technique.

(ii) Next, the word line is maintained high for long enough to provide the appropriate write margin.

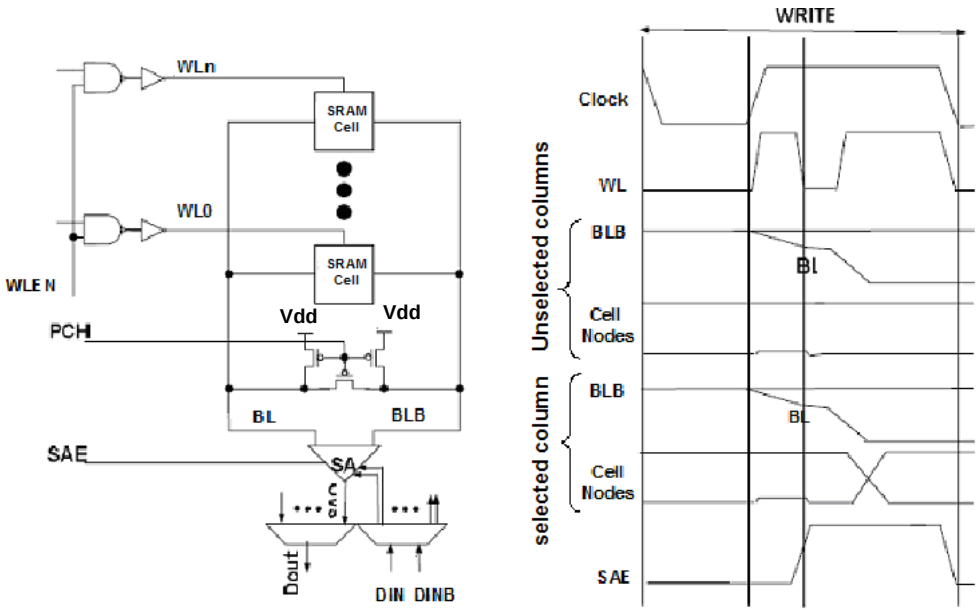

Fig. II.25 Pulsed word line scheme and write operation timing diagram

The RMW technique uses a modified sense amplifier which acts both as a sense amplifier in a read operation, and as a write driver in a write operation. The read operation is similar to the PWL method but post sense amplifier muxing is used.
In the pulsed bit line method, the precharged bit lines (bit line and its complementary) are pulled down (ΔBL) by 100mV to 300mV before turning on word line in a read operation.

With bit lines found at a lower voltage, the discharge read current flowing through the pass gate and pull down transistors decreases. Hence, the accessed cell is less likely to flip during the read operation.

Their results have shown that for PWL technique at 0.8V, the cell failure has been reduced by 33 times compared to a conventional design, whereas at 0.7 V, cell failure has been improved by 26 times with the RMW method. On the other hand, the PBL scheme has improved the failure rate by 10 times when ΔBL is equal to 100mV. However, one drawback comes from the RMW method which incurs an extra area penalty of 4 to 8%.

II.3.1.2 Vdd variation for accessed memory cell

A concept of supply voltage variation called the column based dynamic power supply has been introduced by K. Zhang et al. [Zha06] at Intel to improve the read and write margin of the memory cell. In section II.2.2.1, we saw that the pull down transistor has to be stronger compared to the pass gate transistor to prevent the node voltage developed at N1 from becoming higher than the trip point of the inverter PU2 and PD2 during the read operation. On the other hand, for a proper write operation, a stronger pass gate transistor is preferred along with a weaker pull up PMOS device (section II.2.2.2). The sizing of these transistors, to meet the requirements for a good balance between read and write margin, turns out to be even more difficult as the design window is getting increasingly narrower [Zha06]. Therefore, varying dynamically the voltage between the word line voltage and the supply voltage attenuates the problem of static noise margin (read stability) and write margin problem. The structure proposed is shown in figure II.26 [Zha06].

When a read operation is performed, the supply voltage of the SRAM cell (SRAM_VCC) is connected to a higher voltage (Vcc_hi) than the word line voltage through the column voltage multiplexer R. By doing so, the gate drive strength of the pull down device (PD1 in figure II.19) increases and improves the cell read stability (30% improvement of the static noise margin with 100mV higher voltage on SRAM_VCC). The other cells found in the same row, and experiencing a “read stress”, are also set at Vcc_hi to maintain good read stability.

During the write operation, only the accessed memory cell with its column is connected to a lower voltage (Vcc_lo) than the word line voltage, resulting in a weaker gate drive of the pull up (PU1 in figure II.21) relative to the pass gate transistor (PG1 in figure II.21). With a weak pull up device (PU1) and a strong pass gate (PG1), internal node N1 can be driven more

easily at 0V through the discharge in bit line path (figure II.21). In the write process, the cells found in the same row will be maintained at Vcc_hi to experience “dummy read” stress.

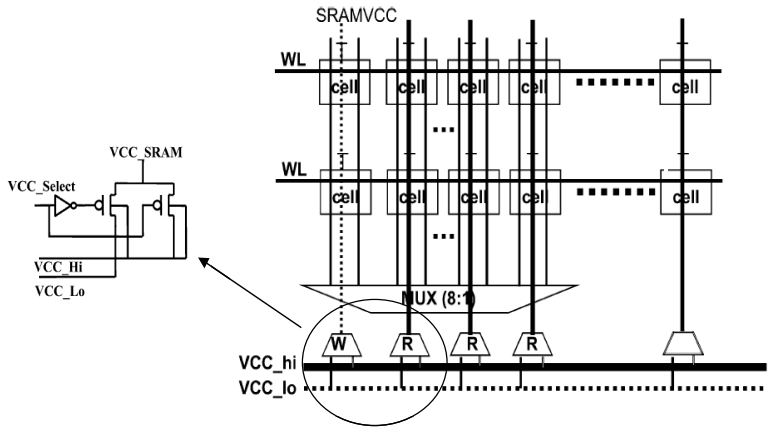

Fig. II.26 Muxing power supplies to Vcc_hi or Vcc_low based on read or write operation

The authors claimed that the write margin has been improved by approximately 20%, by reducing the supply voltage of the SRAM cell by 100mV and keeping word line at 1.1 V. This corresponds to a reduction factor of 10 folds in bit cell failures, if no voltage reduction scheme had been chosen

M. Yamoaka et al. [Yam05], at Hitachi, proposed a writing margin expanding concept for improving the write margin by lowering the supply voltage of the accessed memory cell as displayed in figure II.27.

Each supply voltage Vddm (M) of a memory cell found in the memory column N is connected to supply voltage line Vdd, through transistor MSW (N). During a write operation, MSW is switched off so that Vddm (N) becomes a floating line. The write current flowing between the pull up PMOS and the pass gate transistor lowers the value of Vddm. This will weaken the pull up PMOS relative to the pass gate transistor such that the cell flips easier, thereby improving the write margin.

The word lines of the unselected rows of this column are inactive; hence the other cells remain stable. Moreover, this architecture only lowers the supply voltage of the column being accessed, while the non selected columns remain at Vdd. Hence, data destructions in the non selected columns are eliminated

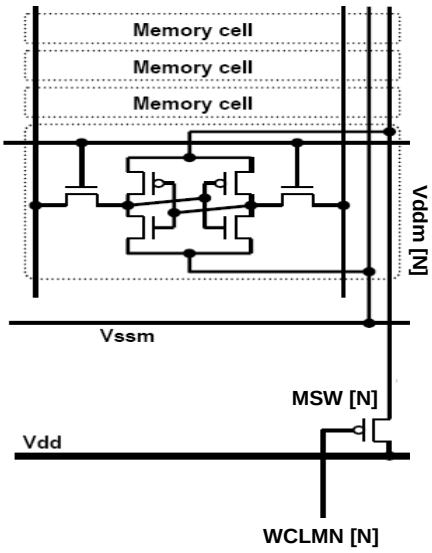

Fig. II.27 Writing margin expanding scheme

II.3.2 Non Memory cell design techniques

II.3.2.1 Replica Technique

Replica technique [Amr98, Gou06] consists in using dummy memory cells connected to a dummy bit line to mimic the delay of bit line path across process conditions. The dummy bit line and the dummy memory cells are controlled by a dummy bit line driver as shown in figure II.28 [Gou06].

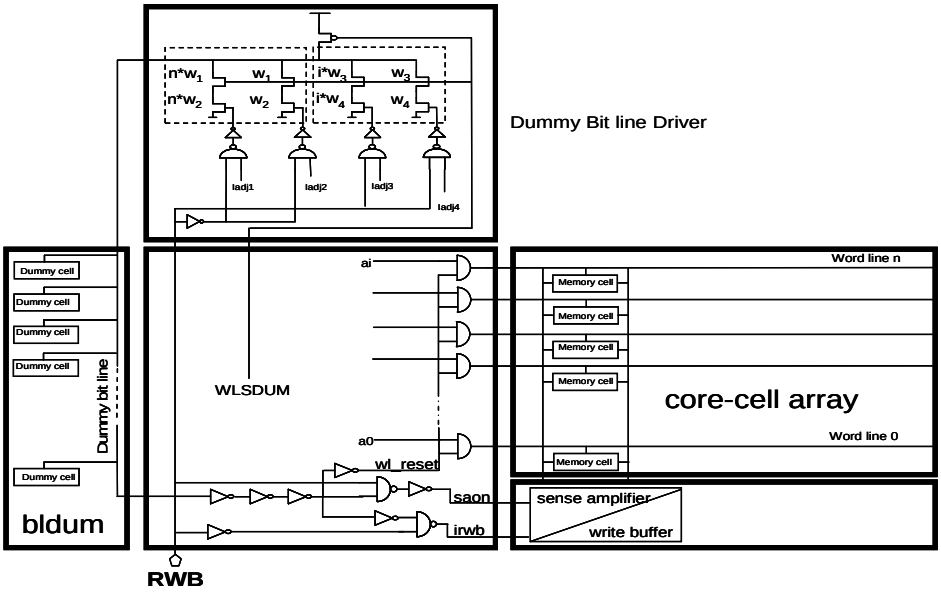

Fig. II.28 Control circuits for sense amplifier activation

II.3.2.2 Redundancy method

Redundancy consists in using spare elements in the memory to replace faulty circuit elements. In other words, it is a good way to improve the wafer yield and to reduce the test cost per good die by fixing potentially repairable defects [Ram97]. The need for redundancy can be justified by the following reasons [Har01]:

(i) Reducing cost per bits in large capacity memories.

(ii) Increasing memory bit capacities for immature processes during the ramp up phase. The use of redundancies on immature processes may allow the use of fully functional memories, using those processes, which would otherwise result in a very low manufacturing yield due to processing defects.

(iii) Provide fully functional parts for memories being manufactured in small volumes.

Memory repair, using redundancy, can be performed through the traditional method. It consists in performing external test and repair, but this method displays serious limitations and is being jeopardized by cost production. This is why self repairable memory i.e. Built-in-self-repairable memory (BISR) are preferred. BISR memory is provided in the form of infrastructure IP (IIP). Compared to functional IP cores such as DSP and microprocessors, which are dedicated to special functionalities in SoC, infrastructure IP cores do not add to the main functionality of the chip [Zor02]. IIP cores are meant to optimize yield and reliability during manufacturing procedure. In the next sub sections, we will introduce the concept of the traditional approach for memory repair, BISR methods and error correction code methods (ECC).

II.3.2.2.1 Traditional repair method

The traditional approach, for memory undergoing reparation, is the external test and repair method. This procedure is processed in several steps as described in [Zor02]. In his explanations, the author describes the steps as follows:

(i) Test algorithms are applied through BIST or by an ATE (Automatic Test Equipment) to the embedded memory with redundancy. Results collected are used in building a failed bit map which is then stored in a memory on the ATE. If memory failures are detected, the test program on the memory tester uses the failed bit map to determine whether the failures are repairable and the best way to allocate the redundant elements.

(ii) Once the appropriate configuration is chosen, the results are then fed into the laser repair equipment which blows the fuses found in the laser fuse box to activate the redundant resources.

(iii) After the repair action has been done, the memory is again tested by the memory tester to check if the repair has been successful.

Figure II.29 resumes the procedure involved in external test and repair of the memory. Nonetheless, this method is prohibitive as it relies on the extensive use of equipment, which represents around 40% of the global manufacturing cost of a semiconductor chip [Zor02].

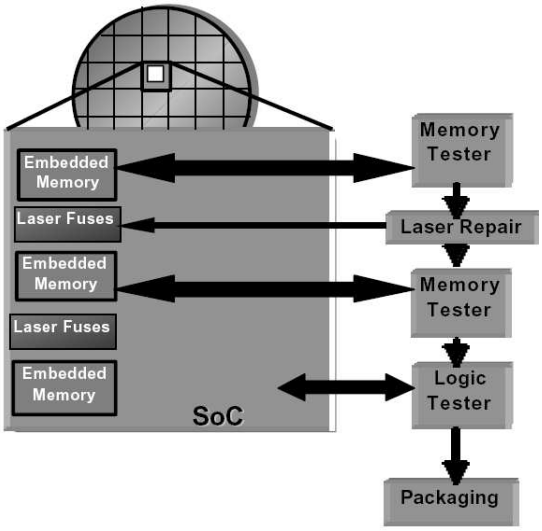

Fig. II.29 External test and repair

II.3.2.2.2 Built in Self Repair method

To deal with manufacturing costs, built in self test (BIST) of the memory is provided in the form of IIP as an alternative to ATE for memory testing. BIST generates testing vectors and applies them to the memory to verify its functionality. In addition to self testing, the memory has to be made self repairable so as to optimize cost production. This is achieved again through IIP in the form of BISR, whereby redundant elements are used as replacement for defective parts of the memory. Figure II.30 resumes the general organization of a diagnosis and repair architecture in a SoC. It is made up of:

(i) A BIST block which generates algorithms (March test, checker board etc) for detecting failures

(ii) A diagnosis block which stores the failure and proposes a repair solution to the repair block

(iii) A repair block which chooses the type of redundancy to be used (word redundancy, word line redundancy etc)

(iv) Fuses (laser or electrical types) which activate the redundant elements in the memory.

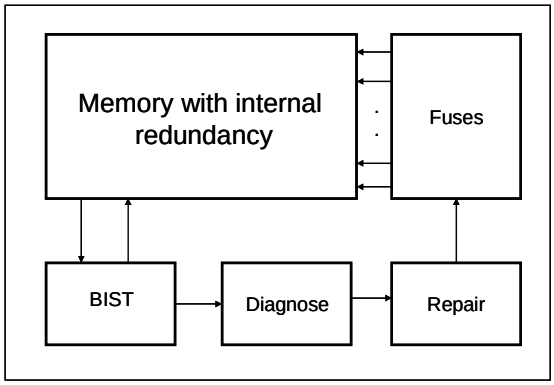

Fig. II.30 General diagnose and repair structure in a SoC

However, T. P. Haraszti [Har01] specifies that the repair efficiency and the memory yield improvement are confined by the number of redundant elements incorporated in the chip. Redundant elements increase the surface area of the chip, leading to a decrease in memory yield. The solution put forward by him for yield improvement, while using redundancy, is to:

(i) Optimize the number of on chip redundant elements i.e. to identify the optimum redundancy configuration for yield maximization.

(ii) Investigate the impact of a given number of redundant elements on the memory yield by computation.

One way of identifying the optimum redundancy configuration is based on statistic methods. Possible defects, which are likely to occur, are inserted in a large number of embedded memories found inside the chip, according to pre defined statistics. Defect statistics, inserted in memory cells, are obtained from process failure history of a reference memory which comes from a specific foundry. For instance, typical defects encountered are single cell bit failure, word line failure, bit line failure and IO failure. Each faulty memory is then analyzed and various possible redundancy configurations are applied to the memory for defect repair.

Data are collected into yield figures and the total area consumed by redundancy is estimated. Hence, the best redundancy choice can be carried out afterwards.

Some examples of memory elements which can be replaced through redundancy include memory cells, rows, columns and sense amplifier block. In fact, there exist several types of redundancies and the most common ones are word redundancy [Rod02, Sch01], word line redundancy and IO redundancy as described in [Rod02].

II.3.2.2.3 Word redundancy

Word redundancy is based on the use of a few redundant flip flops based words, each of which corresponds to a logical address of the RAM [Rod02, Sch01]. Figure II.31 shows the SRAM redundancy wrapper [Sch01] making use of word redundancy which is meant primarily for repairing single bit faults. The wrapper consists of redundancy logic, fuse boxes and the single port SRAM block.

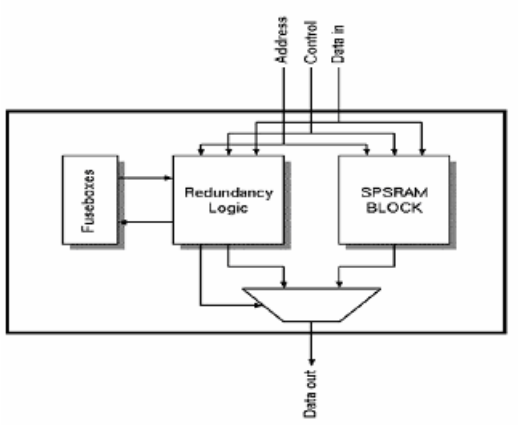

Fig. II.31 SRAM redundancy wrapper

In this approach, V. Schober et al. [Sch01] describe the use of word redundancy in combination with memory BIST (MBIST) and fuses. The memory BIST identifies the address of the faulty memory cells during production test. The faulty addresses of the memory cells are stored in the fuse box which contains a bank of electrically programmable elements i.e. electrical fuses (efuses) which are made up of polysilicon or metal resistor. The efuses are electrically programmed or blown by activating the fuse programming voltage for a certain period of time for e.g. 200 µs in a 90nm technology node using a programming voltage of 3.5V. The fuses then activate the redundant elements in the redundancy logic. In fact, the

redundancy logic block consists of the spare memory words and logic which are used for overwriting defective memory locations when the spare memory words are enabled. It can be seen in figure II.31 that the redundancy logic and the memory block are accessed in parallel. In fact, each time the memory is accessed during a read or write operation, the redundancy logic compares the addresses of the memory to the addresses stored in the fuse box. If the addresses are identical, the word which is read or written is switched towards the appropriate registers assigned to that faulty address.

In larger memory blocks for e.g. a 1Mb SRAM comprising of 4 distinct 256Kb SRAM blocks as depicted in figure II.32 [Rod02], the redundant words can be shared among the 4 blocks to replace one or more faulty locations anywhere in those 4 blocks. The surface penalty in 180nm technology is around 2.4% for 6 redundant words each consisting of 64 bits.

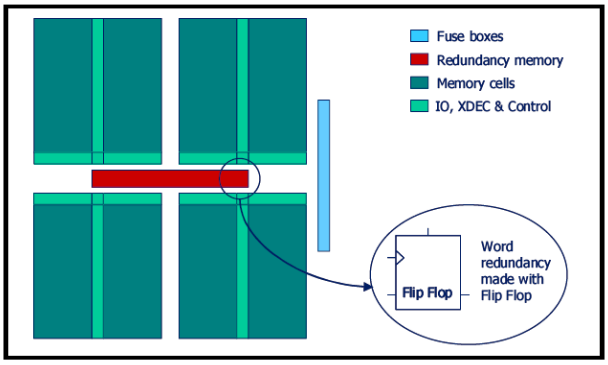

Fig. II.32 1Mb SRAM using word redundancy

II.3.2.2.4 Word line redundancy

Word line redundancy technique consists in replacing one or more rows with spare elements by using appropriate logics around the memory array. This method also uses a combination of MBIST, redundancy blocks and fuse boxes. The MBIST is used in the detection of faulty memory locations during production test, the redundancy logic determines the faulty word lines, whereas the fuse boxes are involved in the selection process of the redundant word lines. Figure II.33 [Rod02] represents the same 1Mb SRAM consisting of 4 blocks of SRAMs, each of which makes 256Kb in size but in this case, redundant X decoders have been added. As far as the increase in area is concerned, the authors reported that for a 256Kb memory block i.e. 512 word lines by 512 bit lines, the use of two couples of redundant word lines with their respective logics yields an increase of less than 3% in the surface area.

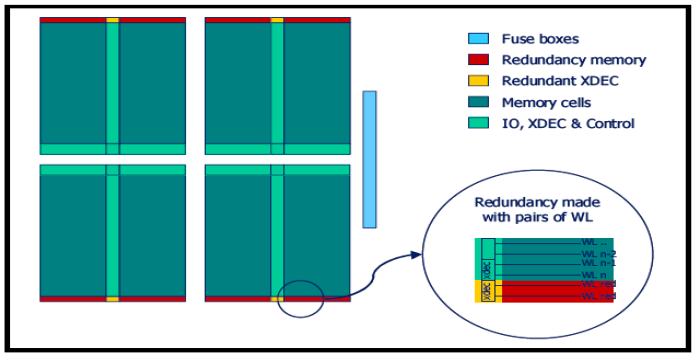

Fig. II.33 1Mb SRAM using word line redundancy

II.3.2.2.5 Bit line and IO redundancy

Bit line redundancy implies the use of redundant columns in the memory array. However in [Rod02], the authors explain that the realization of bit line redundancy is complex, particularly during the choice of a bit line among several bit lines for each IO. The IO (Input/Output) of a memory is comprised of the Y post decoder block (column multiplexer), the read/write circuitry block and output buffers. In fact, sharing the redundant bit lines among the various IO blocks renders the design of the memory architecture too complicated. This is why, they recommend IO redundancy.

In IO redundancy, the method involves the substitution of part of the memory's IO with the redundant IO block as shown in figure II.34.

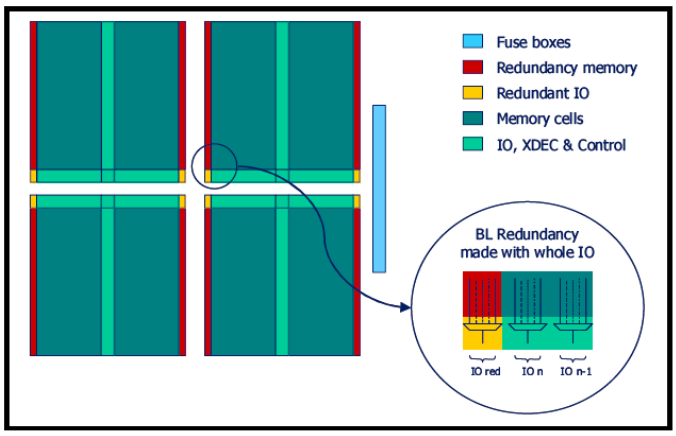

Fig. II.34 1Mb SRAM with I/O redundancy

The memory slice which is being replaced is made up of 4, 8 or 16 bit lines depending on the memory organization, as it has previously been seen in chapter I (section I.3.2), and the IO elements. In IO redundancy, allocation of the redundant columns is also done by programming fuse boxes.

E. Rodney et al. [Rod02] realized IO redundancy on a 256Kb SRAM in 180nm technology with 2 redundant IOs and logical words of 32 bits. They found out that the area penalty is around 6%. Instead of using words of 32 bits, they used words of 64 bits for the same memory and with the same number of redundant IOs. They noticed that the extra surface area of the redundant elements was nearly halved, concluding that IO redundancy shows a greater efficiency for memories with larger word width (number of bits per word). Hence, the area consumption not only depends on the memory size but also on the memory shape.

II.3.2.3 Error correction code

Error correction code (ECC) is a logical correction used in SRAMs to repair soft errors and hard errors. Soft errors are temporary errors caused for example by α particles, cosmic rays, cross talk effect, whereas hard errors are caused for example by stuck faults or permanent physical damage to the memory devices. In fact, redundancy cannot be used to counteract or repair soft errors since these errors are random with respect to time and space [Dup02]. This is why ECC technique is adopted. In ECC, redundant bits are added to the data to form code words which are part of a code space. When an error transforms the code word transmitted, error detection identifies any read word outside the code space and the error correction associates an out of code space read word with the originally written code [Har01].

The simplest approach for error detecting code is called a parity check where a bit is added to the word to have either an even (parity) or odd (imparity) number of ones. An encoder adds the bit to the words transmitted, and a decoder performs a modulo-2 operation on the received data to check the parity of the data. In this way, errors are detected but not corrected.

To perform correction of errors in the memory, other codes have been so far used. One of which is the Hamming code [Bos80, Gray00], known as the single error correction and double error detection (SEC-DED). Popular in memories in the 1960s and 70s, the Hamming code is still widely used today. It is capable of detecting and correcting errors in the memory and is

easily implemented. The basic idea of SEC-DED code is to use enough extra parity bits to identify and correct any single bit error in a word and identify two errors.

Figure II.35 shows how memories incorporate a traditional SEC-DED system for correcting erratic data [Dup02, Gray00].

The data entering the memory is used by the ECC block, called the parity generator to compute the parity bits. In fact, the parity generator is made up of XOR circuits. The resulting code bits or parity bits are fed in the memory along with the data. When the data are read out of the memory, they are captured by the syndrome generator and the bit correction circuitry. The syndrome generator receives as input the uncorrected data and the code bits to generate check bits or syndrome bits. The check bits are then transferred to the error logic block to detect the occurrence of any bit errors i.e. none, single or double and generate them as output. The other output of the error logic block is transmitted to the bit correction circuitry and if single or double errors are detected, the corrections are carried out by the block.

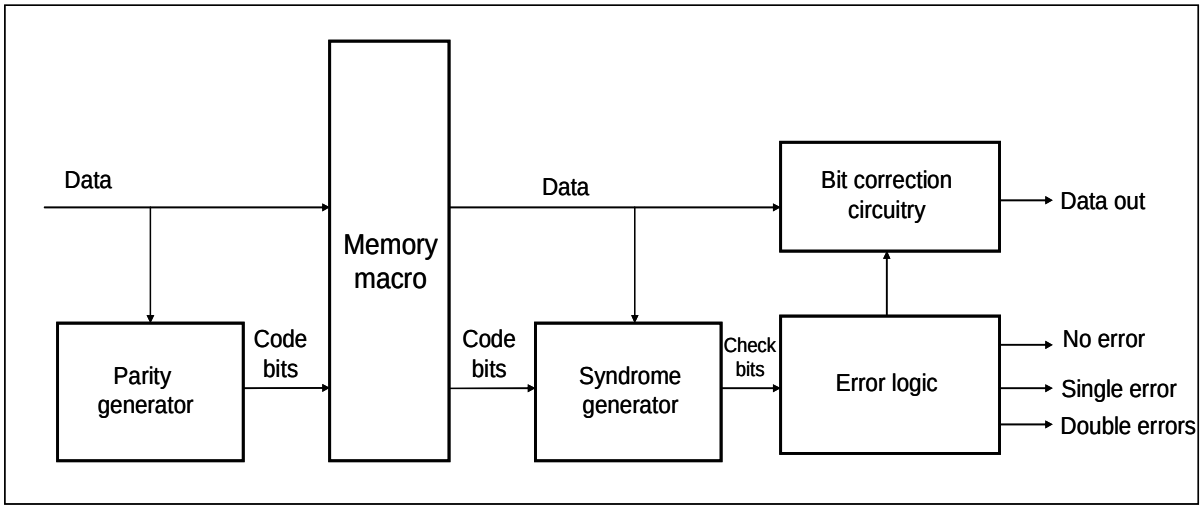

Fig. II.35 Block diagram of a SEC-DED system

Conclusion

The effects of process variations, environmental fluctuations and time factor have become dominant issues for SRAM designers, who need to come up with a variety of design techniques to mitigate these problems.

Indeed, variability significantly affects the performance of SRAMs by causing critical problems, including excessive static power consumption, timing constraint violation and parametric failures in memories. To continue in meeting the performance demands of integrated circuits while improving the memory's yield, numerous design techniques related to memory cells and to the memory architecture have so far been used. However, traditional characterization methods, which were traditionally meant for taking into account global variations, have become obsolete in ensuring robust design of memories in the presence of local variations. The third chapter introduces the traditional corner analysis method, its limitations and the need for statistical methods.

Chapter 3

Corner analysis and statistical method

Manufacturing process variations constitute a major bottleneck to the performances of integrated circuits. Traditionally, to deal with those variability aspects, the corner based method has been extensively used to guarantee that the circuit is fully operative over a wide range of process, voltage and temperature conditions. However, the emergence of local variations renders corner analysis method unsuitable to account for such types of variations. To worsen matters significantly, designers choose to over design their circuits to compensate for local variations, so as to meet the targeted performance. However, using excessive guardbanding also ensues in ICs which are more conservative than necessary in power, area and other performance specifications. For these reasons, designers are revamping the corner method and developed statistical design techniques and tools.

INTRODUCTION

Manufacturing process variations have led to drastic variations in process and electrical parameters of transistor devices, as a result of which the performances of integrated circuits become severely degraded. As it has been previously enumerated in chapter 2, process variations are composed of global and local variations. In the past, global variations were only the main concerns of designers in digital circuits, but with the move downward in process geometries, local variations can no longer be bypassed. To face the variability of fabrication processes along with the operating conditions in circuit design, designers initially adopted the corner based methodology to characterize the circuit under best case and worst case conditions. For instance, the estimation of the maximum operating frequency of a circuit is carried out at least under two distinct timing analysis i.e. one while considering the best PVT conditions (best case timing corner), and the other one while considering the most unfavourable conditions (worst case timing corner). Nevertheless, although the corner approach has been the mainstay of designers for many years, this method constitutes a source of pessimism in the estimation process of circuit performances. This chapter focuses on the principles of corner based analysis, its advantages and limitations. It also introduces the importance of statistical method, as an alternative to corner analysis. Moreover, a modelling approach for computing both the probability of fulfilling a timing constraint in the memory and the appropriate design margin in the presence of process variations will be provided.

III.1 The corner analysis method

In [Zha95], the corner based method or worst case approach is described as the determination of two worst case combinations of the random variables Θ. Usually, Θ represents variations of device model parameters and environmental effects, such as temperature and supply voltages. One combination, Θ^+ represents the maximum value of the circuit performance y^+, whereas Θ^- corresponds to the minimum value of the circuit performance y^-. These two performances can be expressed as follows:

$$y^+(x) = y(x, \theta^+) \tag{III.1}$$

$$y^-(x) = y(x, \Theta^-) \tag{III.2}$$

with x being a set of circuit designable parameters, which includes nominal transistor mask dimensions, process control parameters etc.

Hence, an approximation to the performance mean y^{mean} for a particular design x, can be defined as:

$$y^{mean} = \frac{y^{+}(x) + y^{-}(x)}{2} \tag{III.3}$$

The worst case approach is extensively adopted in the characterization of digital circuits. The effects of process disturbances, which appear during the manufacturing steps, affect the performance of IC for e.g. its operating speed. Typically, those disturbances are represented in terms of transistor performances by classifying the devices into fast transistors (+n.σ), mean transistors and slow transistors (-n.σ) [Zha95]. Fast devices operating at high voltages and low temperatures will lead in general to the best case corner, with a high operating speed y^{-}, whereas slow devices operating at low voltages and high temperatures constitute the worst case corner y^{+}. Hence, the two worst case models Θ^{+} and Θ^{-} of delay circuits are given by:

Θ^{-} = (low temperature, high voltage, +n.σ transistor model parameters) (III.4)

Θ^{+} = (high temperature, low voltage, -n.σ transistor model parameters) (III.5)

During circuit characterization, if the circuit is operational at the worst and best case corners, then it can be concluded that the design will function properly at any intermediate condition. A direct application of worst case approach, based on the ±σ CMOS transistor model parameter extraction, has been given by [Zha95]. In this simple modelling approach, the authors refer to the extraction of I-V curves (I_D vs. V_{DS} and V_{GS}) as the major's device characteristics.

J. C. Zhang and M. A. Styblinski [Zha95] explain that if M transistors of the same type are collected from different chips, lots and wafers and the drain currents of the devices are measured for some biasing voltages, the I-V curves will differ owing to process disturbances (figure III.1 (a)). Indeed, if the drain current I_D of each device is measured N times for different V^{i}_{DS} (i=1 to N), then I_D^{ji} represents the i^{th} drain current for the j^{th} transistor. The value of the drain current for each transistor is actually a random variable, with mean current I_{Dmean}^{i} and standard deviation σ^{i} i.e.

$$I_{Dmean}{}^{i} = \frac{1}{M}\sum_{j=1}^{M} I_D{}^{ji} \tag{III.6}$$

$$\sigma^{i} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left[I_D{}^{ji} - I_{Dmean}{}^{i}\right]^2} \tag{III.7}$$

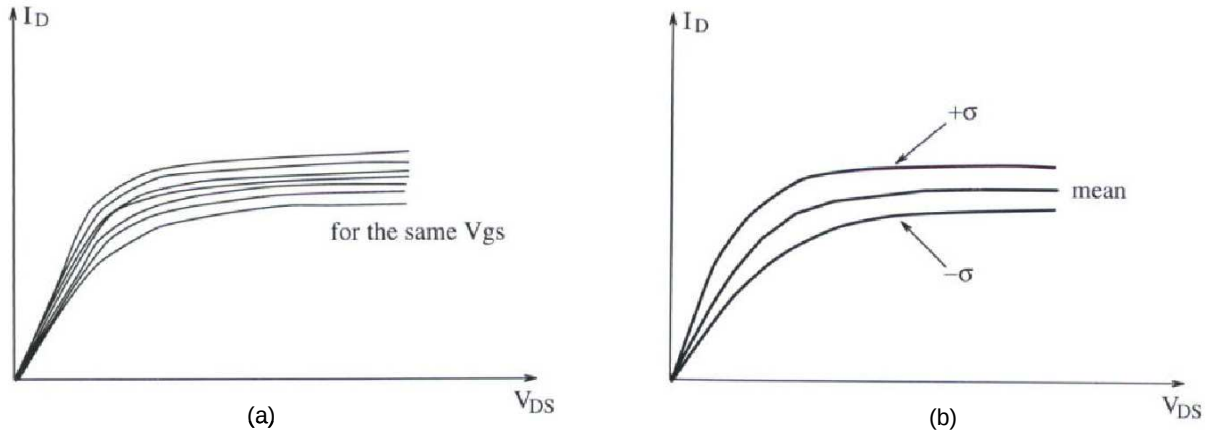

Fig. III.1 (a) I-V curve variations (b) Statistical I-V curve variations

The mean value of all the drain currents can be approximated using (III.6), resulting in the mean curve shown in figure III.1 (b). As the standard deviation σ^{i} represents the variability of the drain current, the variation of the current can be approximated by the two ±σ curves, depicted in figure III.1 (b). The +σ drain current is the sum of the mean current and the + σ drain current variations, whereas the -σ drain current is the difference between the mean current and the -σ drain current variations. This is why, transistor models extracted from the +σ curve are termed +σ transistor models, and transistor obtained from the -σ curve are termed -σ transistor models. Transistor models at ±2.σ and ±3.σ can also be extracted in this way.

III.2 Advantages and limitations of corner analysis method

Traditional corner analysis has the advantage of being a simple and fast method as it has been seen in III.1, compared to statistical methods. Statistical methods rely on the availability of accurate statistical models, which is particularly difficult when the number of model parameters involved is large [Zha95]. The corner method also ensures a high manufacturing

yield, since designers often over design their circuits by using excessive safety margins to account for manufacturing variations.

Nonetheless, though the corner based method is still being used in the characterization of circuits, it displays several disadvantages:

(i) Corner analysis cannot provide quantitative yield information, but provides instead a go/no go answer to the question of whether or not the circuit will function at the extreme case process corners [Gat01]. Hence, this approach does not give the designer any quantitative feedback on the robustness of the design [Sin99].

(ii) The traditional corner approach offers no possibility to designers to detect paths which are more or less sensitive to process variations [Gat01].

(iii) Corner based method does not take into account the correlations between device parameters. Hence, an independent combination of process parameters may lead to unrealistic pessimism, causing valid designs to be rejected [Koc02]. Consequently, design parameters need to be adjusted to meet worst case constraints by increasing the guardbanding. This condition favours unnecessary large chip area, excessive power dissipation and increases design effort [Koc02]. Moreover, it also means a reduction in the product competitiveness and an increase in its cost.

(iv) In the past, the worst case analysis could handle global variations by performing a handful of corner analysis, but as transistor is sizing down, the number of corners needed to account for an increasing number of parameter variations is exploding, up to 64 or more corners are being used by designers [Ele05]. This fact renders the corner analysis method slow and cumbersome. In other words, these sources of variations (process, voltage and temperature) are too numerous and complex to be captured within a small set of process corners [Sap04].

III.3 Statistical modelling

Corner analysis method is viewed as a serious limitation for characterizing digital circuits, due to the interaction between inter die and intra die variations, and the pessimistic standpoint it entails. For this reason, statistical design techniques and tools appear as solutions for taking into account process variations, especially in 90nm design and below. This is mainly true for instance in a 100nm process node, where a 0.01µm variation accounts for only 1% of the

nominal whereas the same 0.01µm variation at the 65nm process node is greater than 15% of the nominal [Pra06].

The use of statistical tool, like statistical static timing analysis (SSTA) tool, is prevalent in achieving timing sign off in IC chips. Indeed, SSTA tool can not only receive as input information the statistical distributions of process parameter fluctuations like channel length variations, oxide thickness variations, but it also handles interconnect variations. The SSTA tool then treats the delays not as fixed numbers, but as probability density functions (pdfs), taking the statistical distribution of parametric variations into consideration while analyzing the circuit [Cha03]. Ultimately, the SSTA tool combines the individual pdfs to achieve an overall distribution for a given node in the circuit.

Indeed, this paradigm shift in timing analysis towards SSTA offers several advantages, particularly in the subnanometer regime where overly pessimistic assumptions threaten to negate many of the inherent benefits that smaller process geometries offer [Pra06]. Foremost benefits of SSTA include:

(i) A more realistic estimation of timing relative to actual silicon performance [Soc07], compared to traditional sign off methods, whereby design passing this standard method might still fail in silicon due to process variations.

(ii) A much faster assessment of all process, voltage, and temperature effects on design timing in order to estimate the expected yield in just one or two runs[Soc07, Stat07]. This might in turn give rise to other potential benefits [Stat07]:

- Pessimism can be greatly reduced, for example possible reduction of arrival times of signals by 10 to 15% to mitigate power consumption.
- Faster analysis might result in faster timing closure
- A quicker exploration of the different scenarios and implementation to grasp yield, performance, and cost trade offs.

In the corner method, tens of analysis are often required that may end up in several hours of runs to achieve timing sign off [Soc07].

In the light of the above advantages derived from SSTA, designers would thus be able to answer the following question: “Given my target specification and my timing report, where can I make improvements?” [Stat07]

Several CAD tools have been developed so far, based on process oriented approaches for the statistical design of ICs. The first complex statistical simulator called FABRICS for IC

fabrication processes, based on this approach, has been developed by researchers at Carnegie Mellon University at the very beginning of the 80's [Nas84, Zha95]. In this approach, as described in [Zha95], a process simulator, receiving its input from process description and IC layout, computes the geometry and the physical characteristics (dopant fluctuations, oxide variations etc) of a given IC. The output is then fed into a device simulator for calculating the electrical parameters of the IC devices. The interaction between process and device simulators allows a better study of the relationship between process parameters and electrical parameters. Then, the circuit simulator outputs the IC performances after receiving a description of the circuit and device models. This is shown in figure III.2 [Zha95].

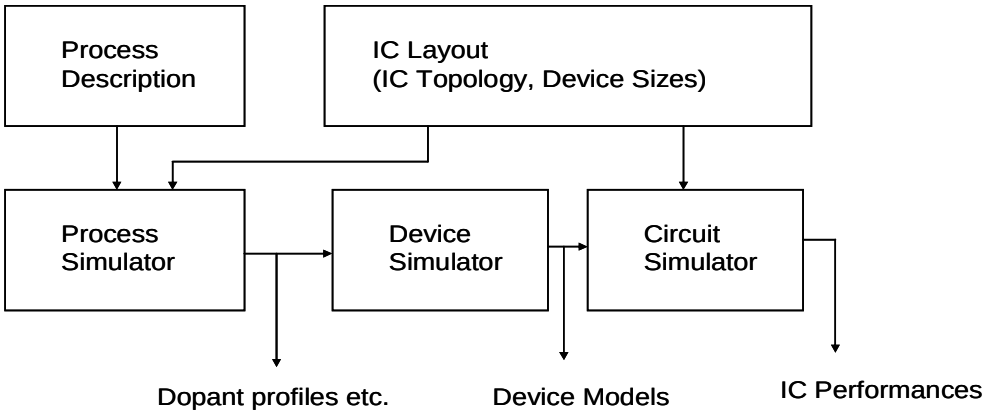

Fig. III.2 Process-oriented nominal IC design

Some of the most common EDA tools actually developed by EDA vendors like Magma Design Automation, Solido Design Automation and Extreme DA, for performing statistical timing analysis and verification include Quartz SSTA (by Magma), SolidoSTAT (by Solido) and GoldTime (by Extreme DA).

For example, Quartz SSTA tool [Mag] provides useful information to designers about their design performance sensitivities with respect to process, metal and environmental variations using statistical methods. Hence, designers are given the opportunity to make appropriate tradeoffs in the design for ensuring an adequate robustness to variability aspects. Quartz SSTA, which is fully integrated in the Magma IC implementation flow, enables a faster and simpler timing closure and sign off. Figure III.3 [Mag] illustrates a simplified Magma IC implementation system which incorporates the SSTA tool.

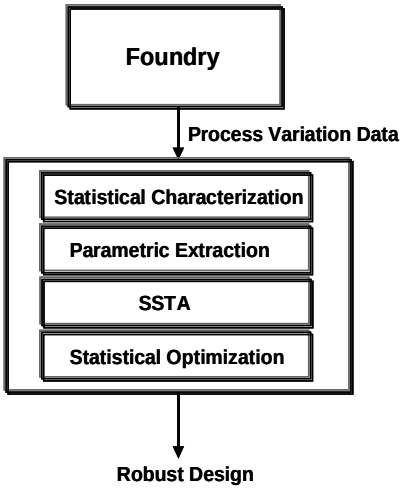

Fig. III.3 Incorporation of SSTA tool in Magma IC implementation system

III.4 Corner analysis and local variations

Conventionally, characterization of a circuit involves performing several simulations across best and worst case corners to verify whether its performances and timing constraints are met under all conditions. For example, verification of set up times of data, associated with clock frequency, is performed whenever propagation delays are the greatest (worst case timing corner). In fact, worst case delay is defined by considering that principal parameters p_i of transistors have their values at $\pm m_i \cdot \sigma_{pi}$ ($m_i \in N$) around their mean values μ_{pi}, depending upon their impacts on propagation delays (σ_{pi} represents the standard deviation of the statistical distribution of parameter p_i). The set up of such a simple approach, through a proper choice of m_i values, allows the worst case to be defined at $n \cdot \sigma_D$, with σ_D being the standard deviation of the distribution delay. σ_D is generally measured on silicon on several qualification process structures of the distribution delays

However, one of the drawbacks of the worst case approach lies in its consideration that the statistical correlation ρ between all transistors is equal to 1, i.e. to ignore local variations or even the differences between supposedly identical transistors found on the same die, since all transistors possess the same characteristics: the worst or the best. In fact, failing to account for local variations across process corners (best and worst case timing corners) is not a serious problem as far as simple data paths are considered. However, the issue is far more complicated with complex data path showing race conditions, and which are quite numerous in auto synchronized SRAM memories as shown in figure III.4. The set up of such an

approach can drive towards optimistic and pessimistic estimations of the worst and best cases respectively.

As a simple illustration, let us consider the signal races represented in figure III.4. Signal A represents the signal, issued from the control block, used in triggering the word line of a particular memory cell (cc) being read so as to discharge bit line (BL) when a '0' is stored in the SRAM cell. On the other hand, signal B generated from the same logic block is involved in turning on the sense amplifier at the appropriate time when BL is discharging (generally when the difference between BL and BLB is around 10% of vdd). Let us also assume that the signal A should arrive at most 0 ps after signal B for the proper read operation of a selected SRAM cell. In accordance with figure III.5, 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, representing the read timing margin, between A and B. Finally, let D_n be the difference in delay between paths A and B obtained during a corner analysis performed at $n \cdot \sigma_{A,B}$

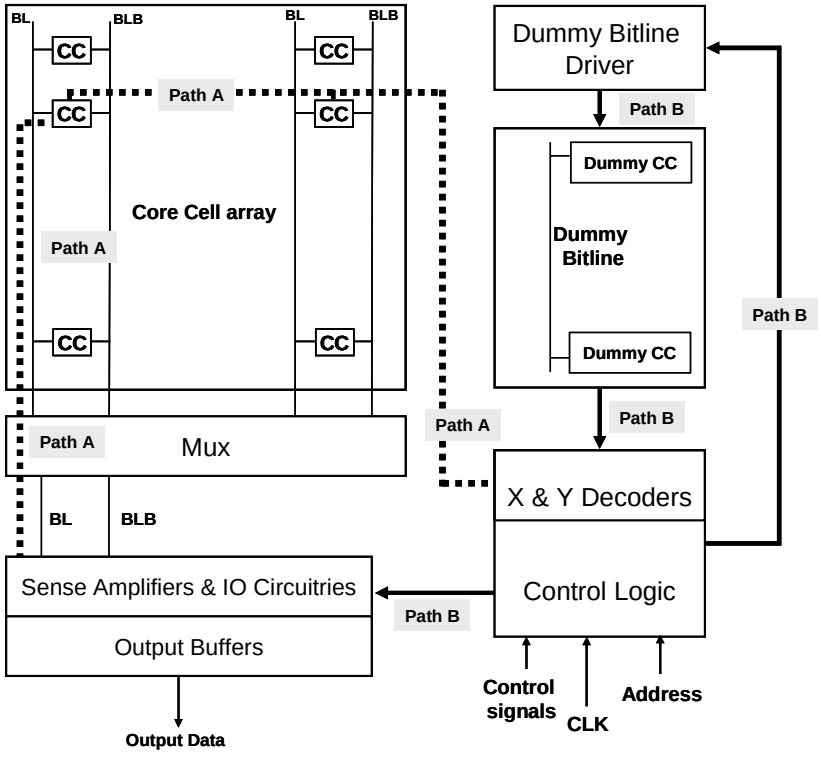

Fig. III.4 Signal races between paths A and B

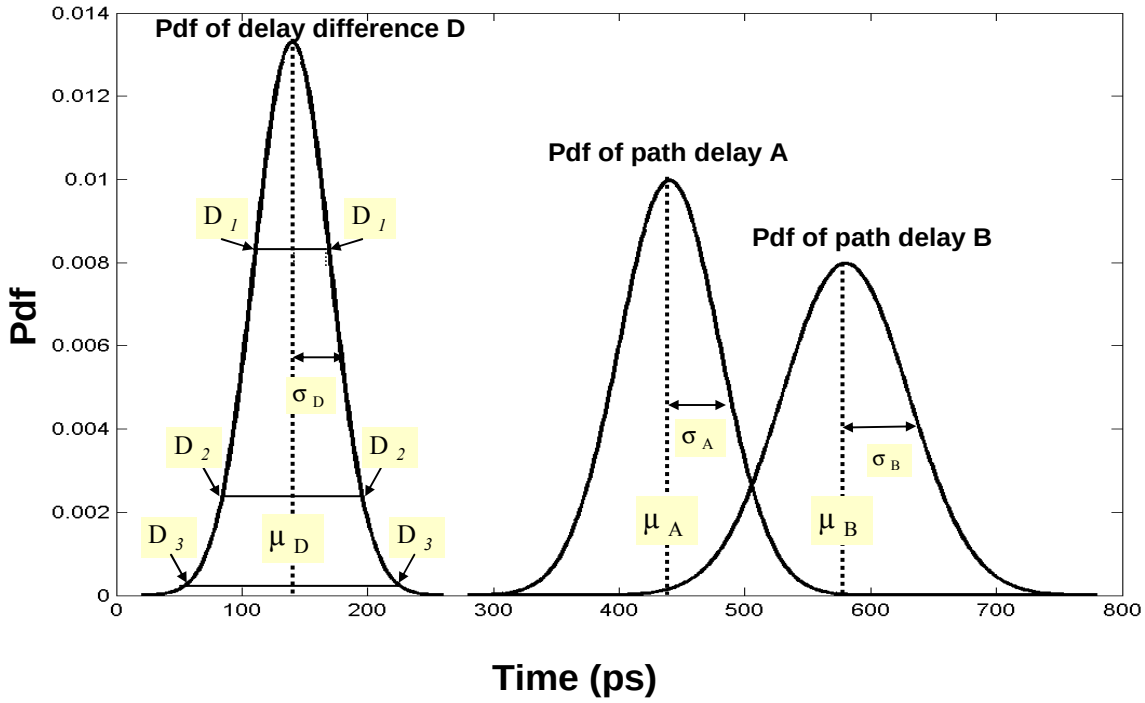

Fig. III.5 Notations

Based upon worst and best case analysis methods, choosing $D_n \geq 0$must guarantee a proper read operation of an SRAM cell with a probability of 99.87% for n=3, as the corner approach ignores local variations i.e. it considers a correlation of 1 between path delays A and B (figure III.6). FF_PA and FF_PB represent the fast corner process of paths A and B, whereas SS_PA and SS_PB stand for the slow corner process of paths A and B.

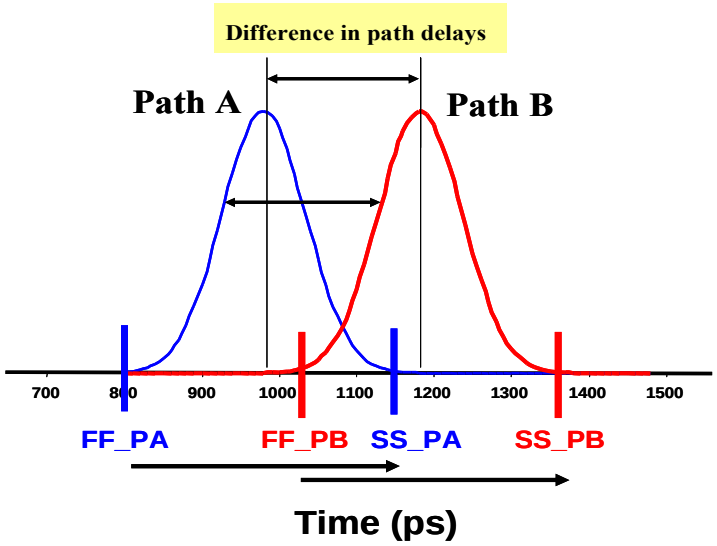

Fig. III.6 Path correlation of 1 between paths A and B

However, as aforementioned, this approach could be faulty since the traditional corner analysis method does not take into account local variations. In fact, the relative increase of local variations with respect to global variations, i.e. the decrease of the correlation coefficient ρ between path delays A and B, can lead to an underestimation of the probability of a timing constraint violation. This is shown in figure III.7, which represents the probability density function (pdf) of path delays A and B, and the cumulative distribution function (cdf) of D for different values of ρ.

Consider the worst case path delay difference between A and B at which the memory performs a read operation i.e. when the difference in delay (D_n) at n.σ is 0 (for n=3 in figure III.7). At this condition, we can see in the cdf that the probability P^V of fulfilling the timing constraint decreases rapidly with the decrease of the correlation coefficient ρ, showing a serious limitation of the corner analysis approach in a context wherein local variations become significant. Note that the mean (1ns) and standard deviation values of the propagation delays (5%) are representative of data obtained by simulation during the read operation of an SRAM cell in a 256kb SRAM design in 90nm technology process.

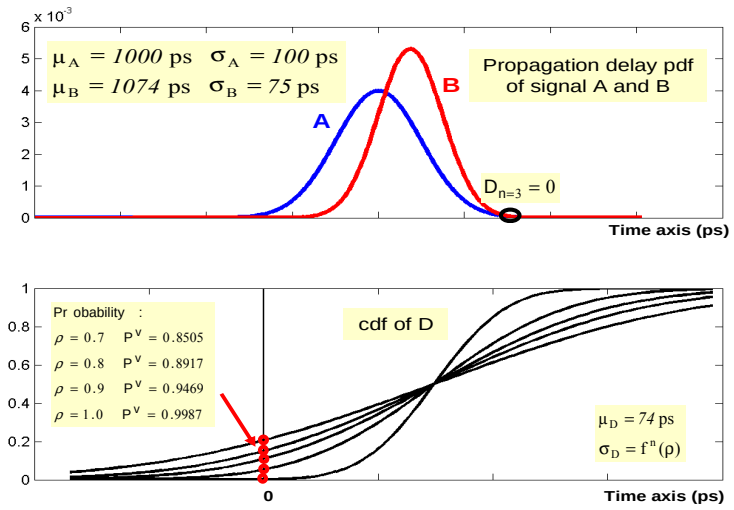

Fig. III.7 Pdf of path delays A and B and cdf of D for different values of correlation coefficients

To overcome the commonly perceived weakness of corner analysis, the most simple and pragmatic solution consists in adopting an empirical design timing margin M_T to account for the lower correlation coefficient value ρ due to local variations i.e. ensuring by proper design solutions that $D_n > M_T > 0$ during worst case analysis performances. Nevertheless, if this

solution seems practical, a question arises: What is the timing margin required in order not to be overly pessimistic or optimistic? To handle this problem, we propose a modelling approach.

III.5 Modelling Approach

In the modelling approach, we will provide a mathematical way of computing:

(i) The probability of meeting a read timing constraint.

(ii) The required read timing margin in the memory, without being too pessimistic in our estimation.

III.5.1 Probability of meeting a read timing constraint

Let us first evaluate the probability of fulfilling a read timing constraint. Assuming that all distributions are normal, even though they are not strictly random variables and so as to remain simple and fully analytical, the distribution D defined as the difference between path delays A and B has mean and standard deviation values given by:

$$\mu_D = \mu_B - \mu_A \tag{III.8}$$

$$\sigma_D = \sqrt{\sigma_A^2 + \sigma_B^2 - 2.\sigma_A.\sigma_B.\rho} \tag{III.9}$$

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

$$P^V = \frac{1}{2}\left[1 + \sqrt{1 - \exp\left(-\frac{2.\mu_D^2}{\pi.\sigma_D^2}\right)}\right] \tag{III.10}$$

The Galton approximation is an analytical approximation of the cdf. In fact, using the properties of the cumulative distribution function (cdf) of the normal distribution, the cdf of a random variable X can be expressed as follows:

$$F(x) = P(X \le x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{x'^2}{2}}\,dx' \quad (x \text{ is }]\,-\infty;\, +\infty\,[\,) \tag{III.11}$$

Hence, the approximation of Galton $\Phi(x)$ is given by:

$$\Phi(x) = \frac{1}{2}\left[1 \pm \sqrt{1 - \exp\left(-\frac{2x^2}{\pi}\right)}\right] \quad (+ \text{ if } x \ge 0 \text{ and } - \text{ if } x \le 0) \tag{III.12}$$

From equation (III.10), the probability P^v of meeting the timing constraint has been evaluated with respect to read timing margin (M_T), normalized with respect to path delay A (μ_A) for different values of correlation coefficients. The results are displayed figure III.8.

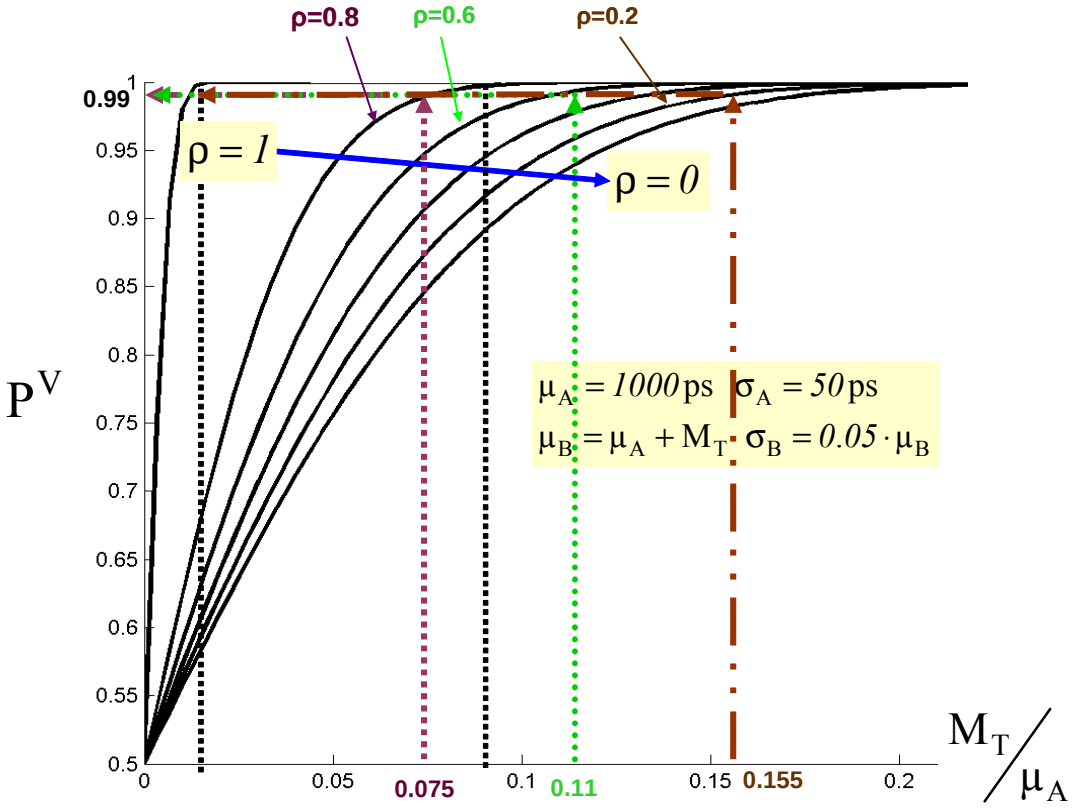

Fig. III.8 P^V variation with respect to read timing margin for different values of ρ

As shown, for ρ=0.8, the required timing margin is 7.5% of μ_A in order to obtain a timing yield of 99%, i.e. a probability of 0.99 of fulfilling the timing constraint. To guarantee a same probability of 99%, the timing margin increases to 11% of μ_A if ρ=0.6 and to 15.5% if ρ=0.2. Thus, the relative read timing margin needed to ensure the same timing yield increases to compensate for the decrease in the correlation coefficient ρ as intra die variations increase.

III.5.1.1 Impact of variability on timing constraint violation and read timing margin

Paths A and B in figure III.4 run across a series of different blocks, having their own timing performances and sensitivities to PVT. In an ideal case, the correlation between those two paths would be equal to 1. However, due to the variations of the device characteristics across

those blocks, the correlation coefficients could be <1. This condition impacts on the path delays and the read timing margin, such that if signal B triggers the sense amplifier before that a specific bit line is being discharged, a read constraint violation is said to occur.

III.5.1.1.1 Impact of path timing and correlation on timing constraint violation

Figures III.9 and III.10 depict the probability of the read timing violations for path delay A, when the correlations of path A are varied. This is performed at two different variability conditions for the path timings in the memory.

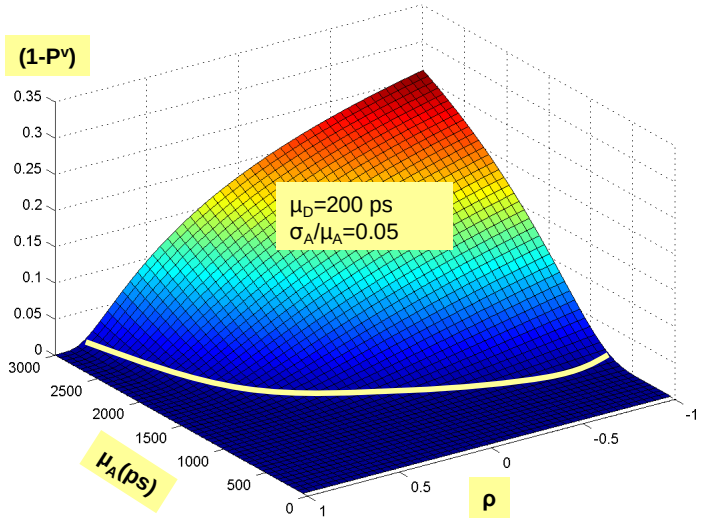

Fig. III.9 Timing constraint violation for a delay variance of 0.05

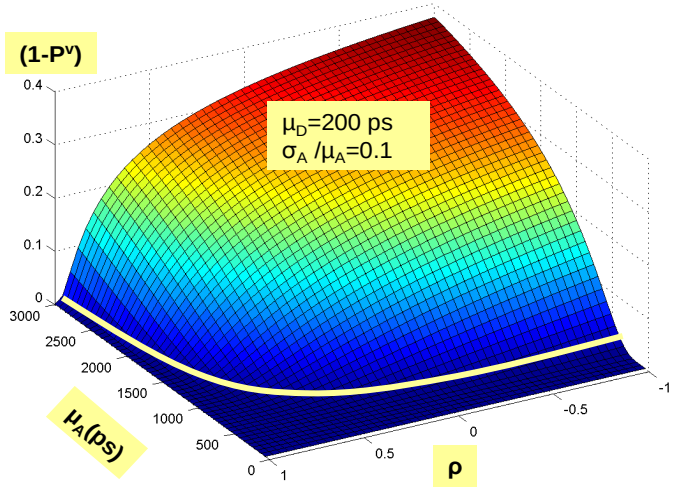

Fig. III.10 Timing constraint violation for a delay variance of 0.1

From the two graphs, it can be clearly seen that, when the correlation decreases due to an increase in local variations, shorter path delays in the memory will be tolerated. Moreover, an increase in the variability in the memory significantly impacts on the probability of a timing constraint violation. When the variability of path A (σ_A/μ_A) increases from 0.05 to 0.1, the contour which specifies the region where the probability of the timing constraint violation is weak, covers a smaller region of the 3D curve.

III.5.1.1.2 Impact of read timing margin and correlation on timing constraint violation

Figures III.11 and III.12 represent the behaviour of the timing constraint violation, at two different variability conditions of path A (n=3.σ), when the correlation coefficients and the read margins are being varied. Suppose for a given technology process (90 nm), a read margin of 200 ps is required for a 256Kb memory when the correlation coefficient equals to 0.8

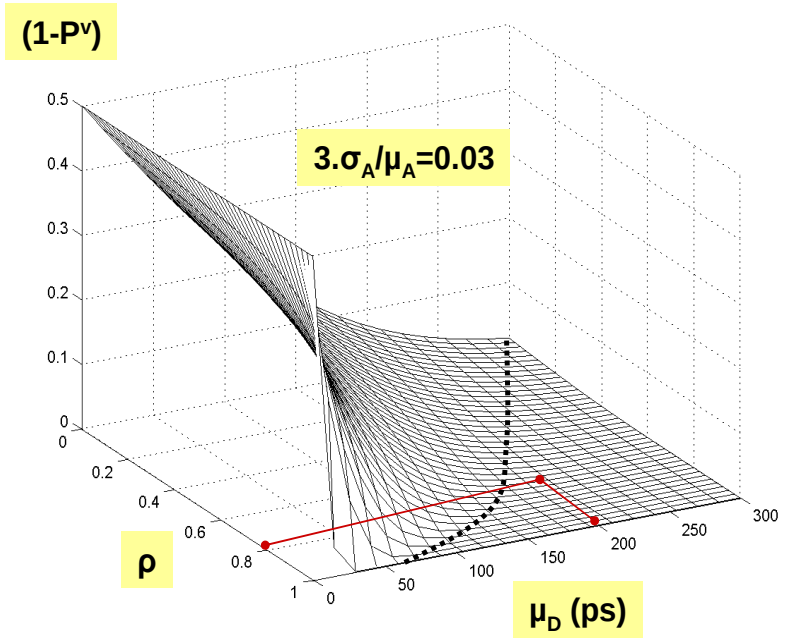

Fig. III.11 Timing constraint violation for a delay variance of 0.03

In order to guarantee the same read margin, with an identical correlation coefficient in 65 nm technology node, the probability of the timing constraint violation drastically increases. For instance, when the variability increases by 3.3 folds in the figures below, the failure probability when ρ=0.8 and μ_D=200 ps increases from 0% to 20%.

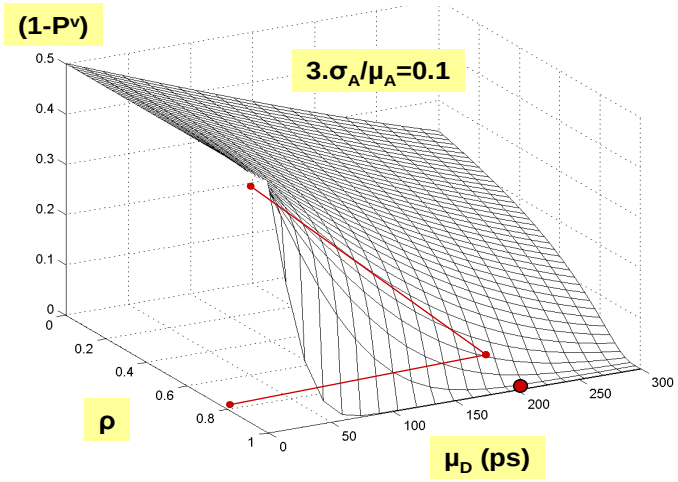

Fig. III.12 Timing constraint violation for a delay variance of 0.1

III.5.2 Computation of the read timing margin

The second step in the modelling approach consists in analysing the first step of the modelling approach the other way round i.e. given that we have a predefined read timing yield, what is the required read timing margin to fulfill such constraint?

We start by doing a simple hypothesis by supposing that we want to have a read timing margin μ_D at $n.\sigma_D$ such that:

$$\mu_D - n.\sigma_D = 0 \tag{III.13}$$

By substituting (III.8) and (III.9) in (III.13), equation (III.13) becomes:

$$(\mu_B - \mu_A) - n.\sqrt{\sigma_A^2 + \sigma_B^2 - 2.\sigma_A.\sigma_B.\rho} = 0 \tag{III.14}$$

Let the sensitivities of path delays A and B to process variations be:

$$V_A = \frac{\sigma_A}{\mu_A} \text{ and } V_B = \frac{\sigma_B}{\mu_B} \tag{III.15}$$

The sensitivities of path delays A and B to process variations are known and found to be roughly constant over a wide range (-20% to +20%) of μ_A and μ_B values. By replacing the following expressions in (III.14):

$$\sigma_A = V_A.\mu_A \text{ and } \sigma_B = V_B.\mu_B \tag{III.16}$$

The path delay μ_B of B can be expressed as:

$$\mu_B = -\frac{a}{b}\left[\sqrt{1 - \frac{b.c}{a^2}} \pm 1\right] \tag{III.17}$$

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

Since the delay of path B should be greater than that of A, μ_B can be formulated as follows:

$$\mu_B = -\frac{a}{b}\left[\sqrt{1-\frac{b.c}{a^2}}+1\right] \qquad \text{(III.18)}$$

Once the delay of path B is computed, the required read timing margin μ_D^{Yield} in (III.8) can be computed to meet a timing yield value defined at $n.\sigma_D$, i.e. such that $\mu_D \pm n.\sigma_D$ possesses a delay value guaranteeing a proper read operation. The subsequent read timing margin is given by:

$$\mu_D^{Yield} = -\frac{a}{b}\left[\sqrt{1-\frac{b.c}{a^2}}+1\right] - \mu_A \qquad \text{(III.19)}$$

III.5.2.1 Impact of variability on read timing margin

Equation (III.19) has been used to evaluate the timing margin M_T required to get a read timing constraint of 99.87% ($3\cdot\sigma$) for different variance values of path delay B, with the delay variance of path A being fixed. The results are shown in figure III.13.

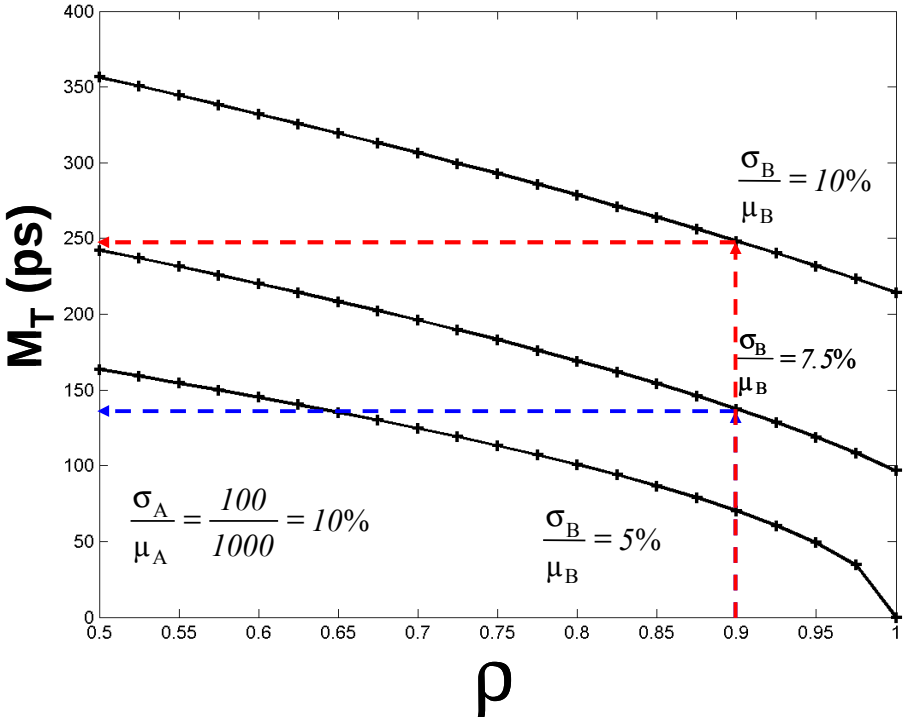

Fig. III.13 Evolution of read timing margin with correlation coefficients for different variance values of path B

A decrease of 2.5% of the standard deviation value of the path delay, when ρ equals to 0.9, leads to a significant decrease of around 40% of the timing margin $M_T = \mu_D^{Yield}$, while complying with a timing yield constraint of 99.87%. Note that this reduction corresponds to 10% of the discharge time of the bit line (BL in figure III.4) by an SRAM cell (denoted by cc in figure III.4), during which bit line is discharged by 70mv from its initial voltage in the memory.

The curve representing the computed read timing margin μ_D^{Yield} for $n=3.\sigma$ and $\rho=0.9$ has also been plotted with respect to variability and path delay B as shown in figure III.14. The figure shows that when the memory operates at high voltage, path delay B is 1ns and its variability is around 2.5%. When the operating voltage of the memory is lowered, μ_B increases to 3.8ns along with an increase in variability of path B, which is around 6%. As a result, to ensure proper read operation when variability increases, the required read timing margin needs necessarily to be increased from 100ps to 200ps, demonstrating the need in developing techniques for adjusting the read timing margins with respect to supply voltage.

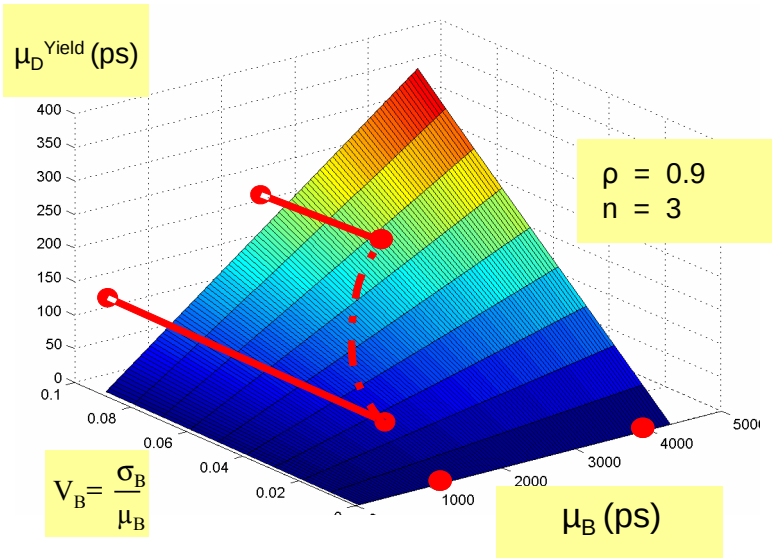

Fig. III.14 Evolution of read timing margin with variability and path delay

Conclusion

In this chapter, we have seen that corner based analysis can no longer be used in circuit characterization owing to its pessimistic nature. In fact, designers continue to introduce increasing safety margin to ensure that the circuit will perform correctly. Furthermore, another limitation of the corner method is its inability in capturing local variations, such that it considers a statistical correlation of 1 between different path delays. This condition is no longer tolerable in complex data path, for e.g. in eSRAM, displaying racing conditions. Considering a correlation of 1 between path delays, involved in signal races, can lead to optimistic conclusions. In order to handle the weakness of corner analysis, we have proposed a simple and pragmatic way of (i) computing the read timing margin required in the memory without being overly pessimistic in our approach, and (ii) of calculating the probability of meeting this constraint. Some applications of this modelling approach will be detailed in the next chapter.

Chapter 4

Applications of the modelling approach

An overestimation of the variability effect renders the design of ICs harder, along with an increased in design effort and exhaustive timing verification. On the other hand, its underestimation leads to manufacturability problems and eventual yield loss [Mut07]. This is particularly true as local variations keep on increasing relentlessly. To overcome the optimism and pessimism introduced by best and worst case analysis, we have developed an efficient way of computing the required read timing margin and the probability of meeting this timing constraint. In this chapter, two applications of the modelling approach will be discussed in details. Moreover, we will analyse the potential benefits that can be derived by such applications in terms of read timing performances in the memory.

INTRODUCTION

In the previous chapter, we have shown that the methodology based on corner analysis, in cases involved in signal races, can give rise to the underestimations of the occurrence of a timing constraint violation. Moreover, we have also seen that the limitations associated with the corner analysis methods are simply due to the assumption that the statistical correlation coefficients between the pdfs of path delays are equal to 1. Physically, the limitations of the corner analysis methods arise since all transistors (N or P) are considered to be perfectly identical, i.e. there exists only a unique and global model of transistors. This unique and global notion of transistor models seems less worthwhile in the design of complex system on chips in advanced technologies (90nm, 65nm and 45nm), just like the global notions of temperatures and voltages [Lia05, Las07, Kan98]. Therefore, it appears necessary in developing design methodologies and performance analysis methods which no longer consider silicon quality, supply voltage or even temperature as global and uniform variables of a circuit.

The aim of this chapter is to consider the above stated problems and see some direct applications of the modelling approach introduced in chapter 3. These applications can be categorized as follows:

(i) Using the probability equation (III.10), developed in chapter 3, to provide a failure probability map of the memory, showing the critical areas of the memory which are more prone to timing constraint violations.

(ii) To develop a statistical sizing methodology of the dummy bit line driver structure (chapter 1, section I.5.2) to enable its proper sizing under timing yield constraints i.e. for a given timing yield at n.σ. The dummy bit line driver structure is an essential component in an auto synchronized memory, as it is responsible for triggering the sense amplifier at the appropriate time during a read operation. In this sizing methodology, both equations (III.10) and (III.19) will be used. Simultaneously, a new dummy dummy bit line driver structure which is less sensitive to process and voltage variations compared to the original dummy bit line driver structure [Gou06] will be introduced. The statistical sizing method will be applied to the original and the proposed dummy structures, and their performances will be compared in terms of variability and read margin reductions at constant timing yield.

All the simulations will be carried out on a 256Kb SRAM in a 90nm technology node, using Hspice.

IV.1 Applications

IV.1.1 Failure probability map

The first practical application of equation (III.10) has been used in plotting the failure probability map of the memory, while considering different correlation (ρ) scenarios between the position of a particular memory cell (CC) being read, denoted by d, with respect to the dummy bit line driver. The figure IV.1 illustrates the map convention, representing a memory made up of 512 rows and 512 columns (256 Kb), with the dummy bit line driver representing the origin.

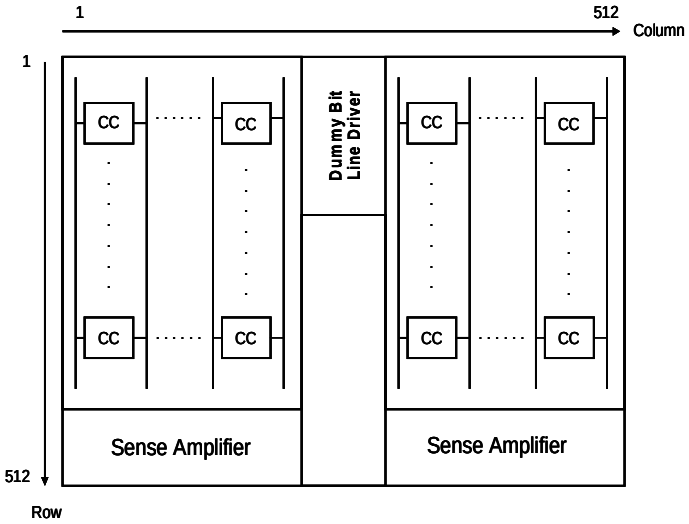

Fig. IV.1 Map convention of the memory

In fact, the correlation between the location of a memory cell and the dummy bit line driver depends on the distance d between them. Generally, the value of this correlation is unknown and is obtained from measurements on silicium. In our study, we could not have access to the values of the correlation coefficients. Thus, we formulated some hypothesis about how this correlation could vary with respect to the distance d.

The different types of variations considered for ρ could be (figure IV.2):

(i) Uniform

(ii) Linear

(iii) Exponential

(iv) Hyperbolic

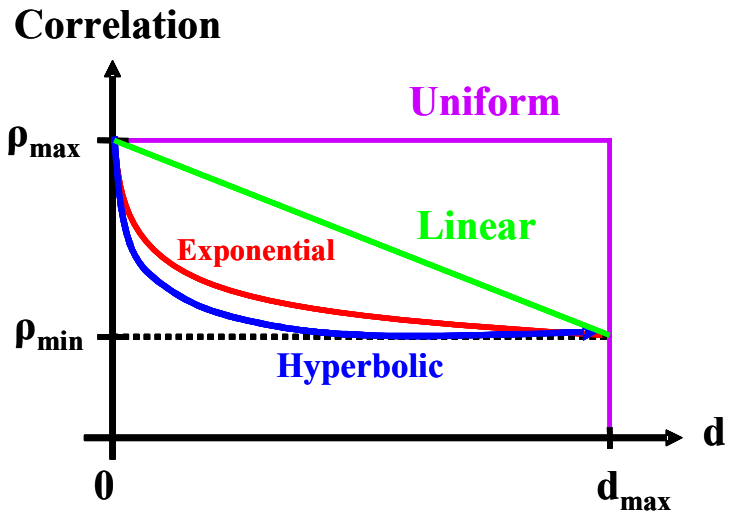

Fig. IV.2 Considered scenarios for behaviour of ρ

Among the four types of variations considered, the one which is more representative of the behaviour of the correlation coefficient as the distance d increases seems to be hyperbolic function. This is confirmed through pelgrom's equation [Pel89], which specifies that the correlation varies inversely with the distance d.

Statistical simulations have been performed for three couples of voltages and temperatures to assess the probability of having a timing constraint violation $P_{failure}$ of the memory cells with respect to their locations in the memory core. To plot the curves, we considered a ρ_{min} of 0.75 and a ρ_{max} of 0.95. The 3 couples considered are:

(i) A nominal voltage Vdd_{nom} and temperature T_{nom} i.e. 1.2V and 27˚C.

(ii) A minimum operating voltage Vdd_{min} and high temperature T_{max} (worst case condition) i.e. 1.0V and 125˚C.

(iii) A maximum operating voltage Vdd_{max} and low temperature T_{min} (best case condition) i.e. 1.32V and -40˚C.

The next subsections report the results obtained for the four types of variations considered.

IV.1.1.1 Calculated failure probability map (Vdd_{nom}, T_{nom})

We can see that when the correlation varies uniformly (figure IV.3 (a)), the critical areas of the memory which are more prone to timing constraint violations are memory cells found at the top of the memory array, as they are far from the dummy bit line driver and the sense amplifier, and owing to the resistance of bit lines and word lines. Memory cells found in any particular row have an identical timing constraint violation probability since the correlation coefficient remains unvaried. Concerning the linear, exponential, and hyperbolic functions, the critical locations of memory cells are those found in the upper part of the memory core as indicated in red. However, those that display the highest probability of timing constraint violations are located at the top extremities of the memory core. This is because SRAM cells found in those areas are far from the dummy bit line driver and they are the furthest from the sense amplifier. Furthermore, the resistivity effects of bit lines and word lines are the greatest during the read operation for memory cells found there.

As we move closer to the dummy bit line driver, the correlation coefficient increases to 0.95. Hence, SRAM cells found close to the dummy bit line driver have a failure probability, which decreases significantly. For instance, a hyperbolic behaviour leads to a failure probability which is nearly halved i.e. from 110 ppm to 60 ppm.

Areas of the memory core having the lowest failure probability rate are found in the lower part of the memory core, as indicated by the dark blue area. In this region, memory cells are close to the sense amplifier such that the discharge rate of bit lines is shorter, implying a greater read margin of the memory.

These failure probability maps could be very helpful in test chip monitoring. For instance, engineers measuring the read access time in those areas of the memory would be more alert when such operations are being carried out.

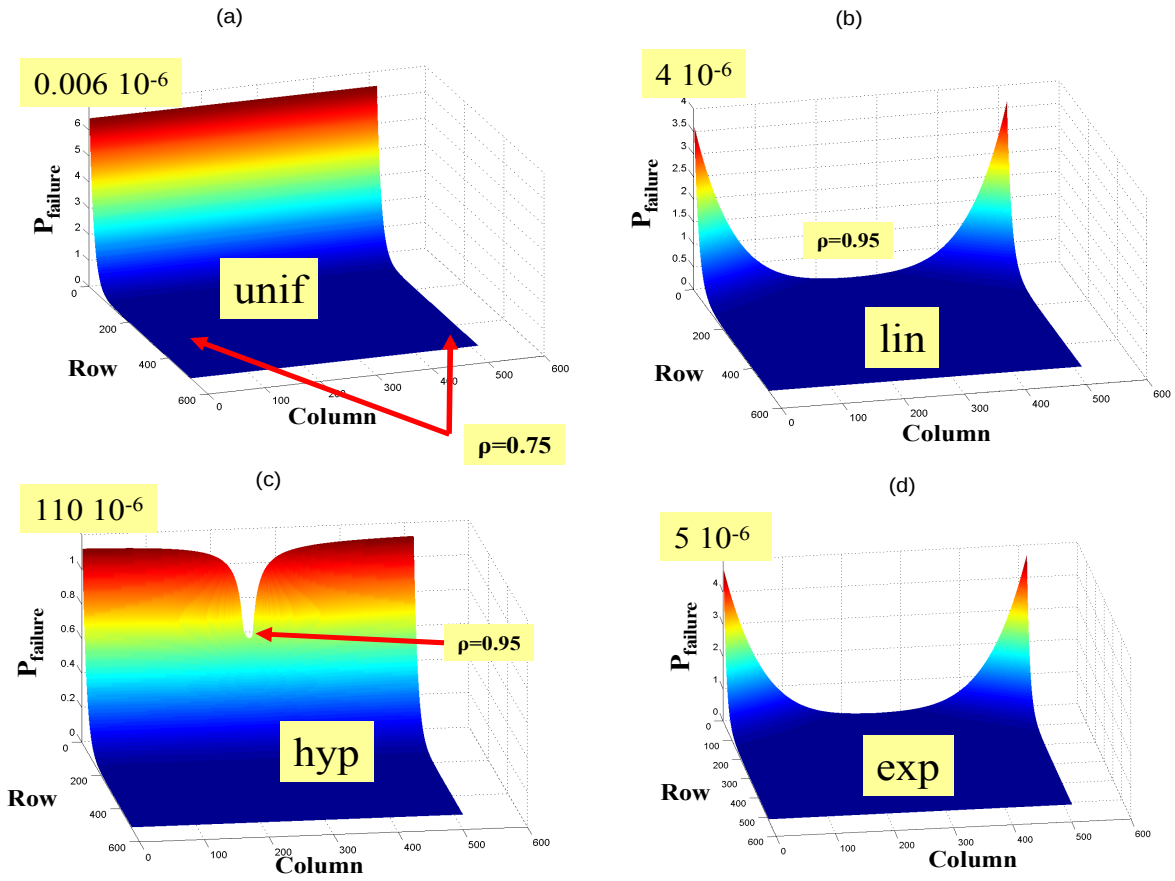

Fig. IV.3 Failure probability map for (a) Uniform (b) Linear (c) Hyperbolic and (d) Exponential variations of ρ at nominal operating conditions

Figure IV.4 summarizes the most critical areas of the memory core, depending on the types of correlations considered.

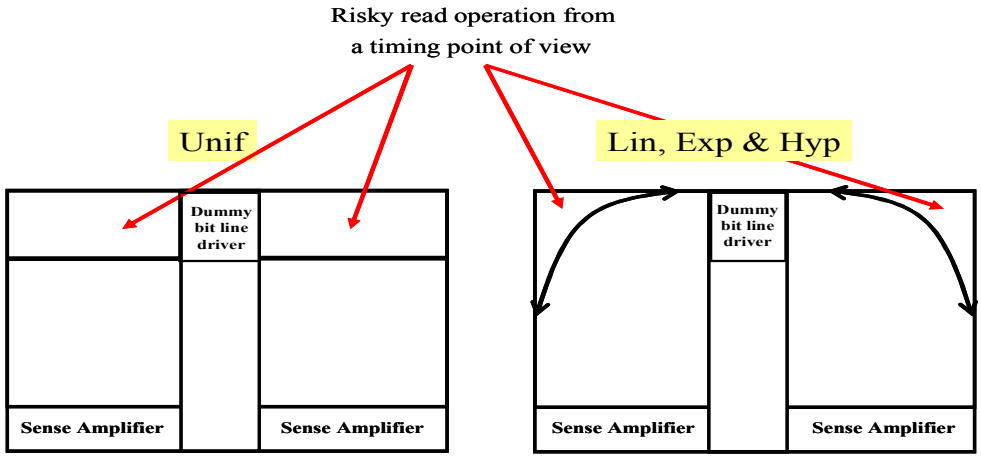

Fig. IV.4 Representation of memory areas most likely to experience read timing constraint violations

IV.1.1.2 Calculated failure probability map (Vdd_{min}, T_{max})

Figures IV.5 (a), (b), (c) and (d) represent the probability of the read timing constraint violations for the uniform, linear, hyperbolic and exponential behaviours of the correlation coefficients under worst case operating conditions of the memory i.e. Vdd=1.0V and T=125°C.

The critical areas of the memory core which are more prone to read failures are the same as in the nominal process. However the main difference results in the occurrence of the probability of the timing constraint violation, which increases drastically compared to the nominal process. In the case of a hyperbolic variation, the failure rate increases by 180 folds for memory cells found at the extremity of the memory core. This is because at higher temperatures, the delay associated with logic gates increases and impacts on the read access time.

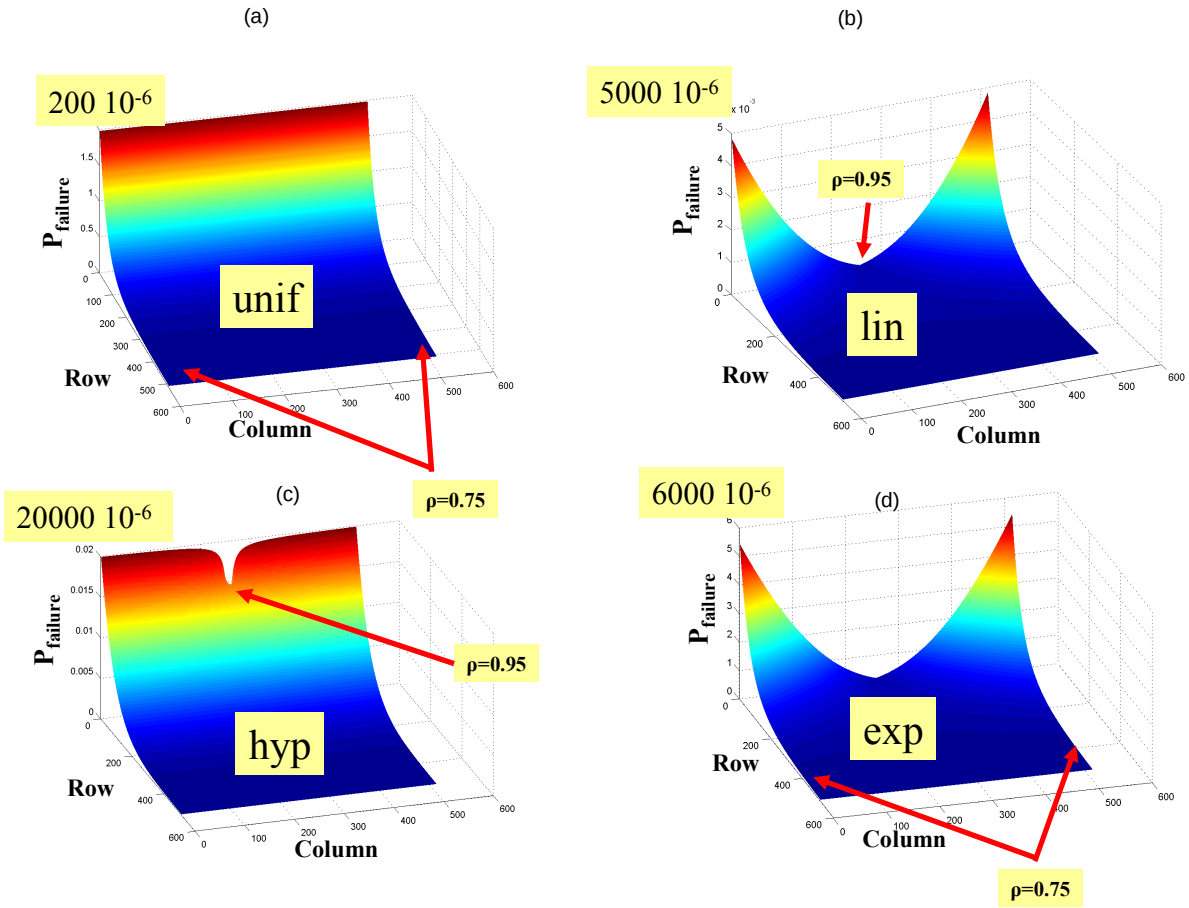

Fig. IV.5 Failure probability map for (a) Uniform (b) Linear (c) Hyperbolic and (d) Exponential variations of ρ at worst operating conditions

IV.1.1.3 Calculated failure probability map (Vdd_{max}, T_{min})

Figures IV.6 (a), (b), (c) and (d) show the probability of the read timing constraint violations for the uniform, linear, hyperbolic and exponential behaviours of the correlation coefficients under best case operating conditions of the memory i.e. when Vdd=1.32V and T=-40˚C. Similarly, as in the worst and nominal cases, the critical areas of the memory cells are found in the same regions of the memory core during a read operation for identical reasons stated in IV.1.1.1. However, in the best case conditions, the probability of the timing constraint violation is significantly reduced. In the hyperbolic variations, the reduction in the probability failure for memory cells found at the extremity of the memory core is around 18 folds, compared to the nominal process.

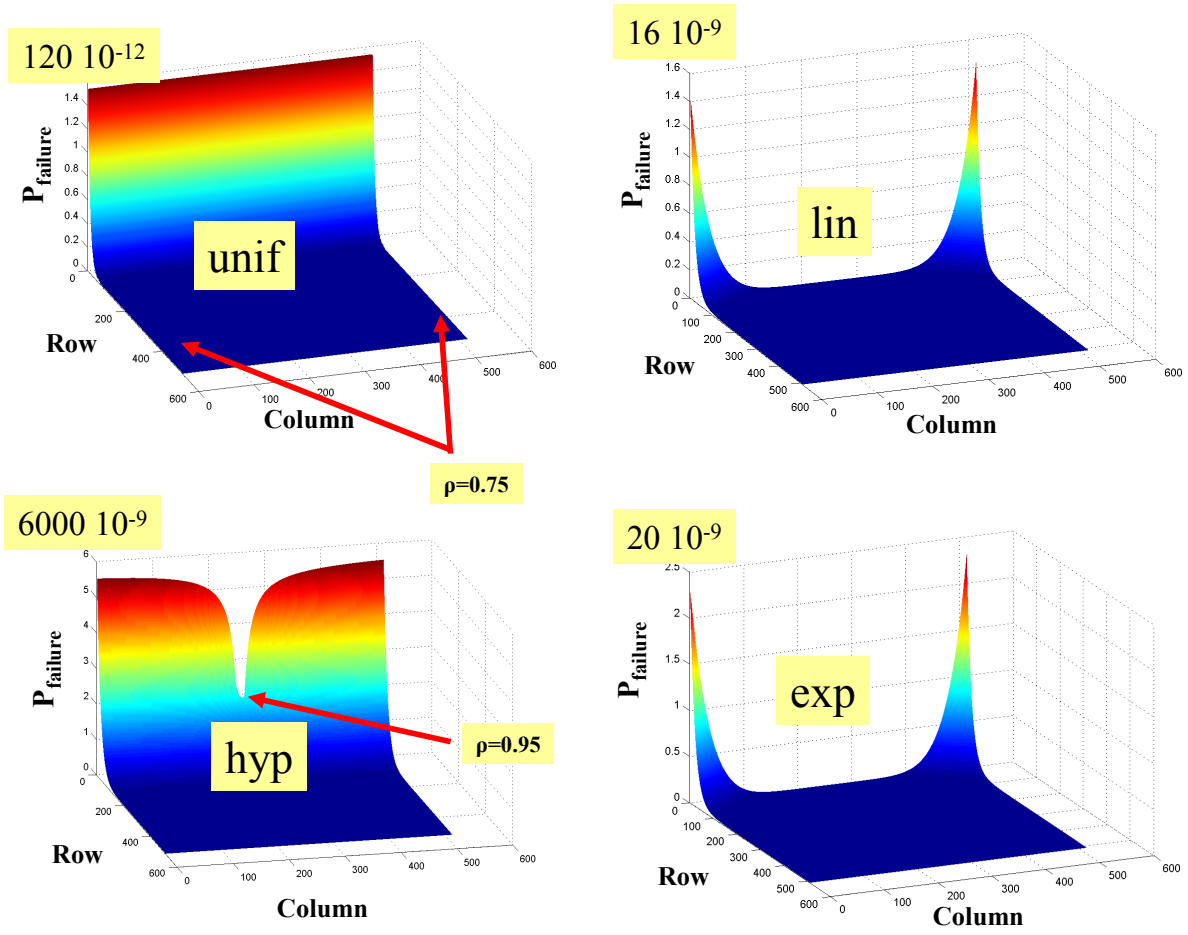

Fig. IV.6 Failure probability map for (a) Uniform (b) Linear (c) Hyperbolic and (d) Exponential variations of ρ at best operating conditions

IV.1.2 Statistical sizing methodology of dummy bit line driver

The second application of the modelling approach is used in the statistical sizing of the memory for optimizing its performances. The sizing method has been applied to the dummy bit line driver since it is the main structure of the memory ensuring the proper triggering of the sense amplifier when bit line or its complementary is being discharged. The statistical sizing of the dummy bit line driver allows sizing of the memory so as to meet a predefined timing yield constraint. The sizing method is composed of three basic steps and two verification procedures. The sizing is carried out at a nominal process and can be described as follows:

Step 1: Identification of the most critical (Vdd, T°) $_{Crit}$ condition

Starting from an initial solution, for instance a read margin obtained in the corner analysis method, the first step involves the identification of voltage and temperature conditions (Vdd, T°) $_{crit}$ having the poorest timing yield. In fact, under signal races conditions, the (Vdd, T°)$_{crit}$ conditions 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 of supply voltage and temperature, 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 of temperatures and voltages. The results of these simulations lead to the (Vdd, T°) $_{Crit}$ condition demonstrating the smallest timing yield, by simply looking for the (Vdd, T°) $_{Crit}$ condition displaying a highest numerical value of the following expression:

$$\frac{\mu_A \mu_B}{(\mu_B - \mu_A)^2} \qquad \text{(IV.1)}$$

An illustration of the utility of this method, in the identification procedure of the (V, T) $_{crit}$, is shown in table IV.1, where expression (IV.1) has been used in identifying the conditions at which sizing should be performed for different temperature and voltage conditions at a nominal process.

It can be seen that the temperature and voltage conditions at which sizing should be performed is at 1.0V and 125˚C, where the value of the expression $(\mu_A . \mu_B)/ (\mu_B-\mu_A)^2$ is the greatest. At this condition, the probability of meeting the timing constraint violation is the greatest.

Table IV.1 Voltage and temperature conditions at which sizing should be performed using expression (IV.1)

Vdd (V)	Temperature (˚C)	$(\mu_A.\mu_B) /(\mu_B-\mu_A)^2$	$(1-P^V)$ %
1.0	-40	54	1.64
	27	126	2.65
	125	1292	7.14
1.2	-40	118	0.22
	27	103	0.13
	125	202	0.34
1.32	-40	97	0.02
	27	69	0.00
	125	98	0.01

The proof can be derived by considering the probability of fulfilling a read timing constraint P^v given by (refer to section III.5.1):

$$P^v = \frac{1}{2}\left[1 + \sqrt{1 - \exp\left(-\frac{2.\mu_D^2}{\pi.\sigma_D^2}\right)}\right] \tag{IV.2}$$

As (Vdd, T°) $_{crit}$ represents the condition showing the highest probability at which a timing constraint violation is said to occur, P^v should be minimum at this condition. Hence P^V will have the minimum value if:

$$\frac{\mu_D^2}{\sigma_D^2} \tag{IV.3}$$

is also a minimum.

$$\text{Let } \alpha = \frac{\sigma_A}{\mu_A} \approx \frac{\sigma_B}{\mu_B} \tag{IV.4}$$

By substituting σ_D by (III.9) and $\sigma_{A,B} = \alpha.\mu_{A,B}$ in (IV.3), expression (IV.3) can be expressed as:

$$\frac{\mu_D^2}{\alpha\sqrt{1 + \frac{2.\mu_A.\mu_B}{(\mu_B - \mu_A)^2}(1-\rho)}} \tag{IV.5}$$

Consequently, expression (IV.5) is a minimum if:

$$\frac{\mu_A.\mu_B}{(\mu_B - \mu_A)^2} \tag{IV.6}$$

has the highest numerical value.

Step 2: Variability estimation and computation of required timing margin

The second step requires the estimation of the variability of paths A and B involved (σ_A/μ_A, σ_B/μ_B) in the signal races, if the value of variability is unknown. To do so, Monte Carlo simulations of the critical path are performed at the critical condition of voltage and temperature $(Vdd, T°)_{Crit}$ 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 expression (III.19). It should be noted that, if measured silicon values of the correlation coefficients ρ are available, these values can be used instead of those obtained through simulations.

Step 3: Sizing for a given timing yield

After the computation of μ_D^{Yield}, the third step consists in sizing the structure to obtain the computed read timing margin obtained in step 2. The sizing procedure is carried out at a typical process and at the voltage and temperature conditions $(Vdd, T°)_{Crit}$ obtained in step 1.

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 $(Vdd,T°)_{Crit}$ to obtain μ_A, μ_B, σ_A, σ_B and ρ values. The timing yield is then evaluated using (III.10). 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 and ρ.

Step 5: Second verification step of the timing yield

Step 5 is optional but strongly recommended. 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 (Vdd, T°) couples are greater than the predefined constraint $(Vdd, T°)_{Crit}$, the verification step is over. However, if the constraint is not satisfied, step 1 should be repeated with the new sizing obtained.

It should be noted that the statistical approach can be applied to the memory if the critical operating conditions of the memory can be quickly identified, otherwise statistical simulations need to be performed over a wide range of temperature and voltage conditions. Nevertheless,

the main difficulty in the identification of the critical conditions is due to the existence of reversal of temperature dependency. Reversal of temperature dependency causes a systematic shift of the best and worst case corners at any temperatures. In other words, the best and worst case corners are not always found at low and high temperatures. The identification procedure introduced in step 1 in the statistical flow renders this method independent 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 due to reversal of temperature dependency. Figure IV.7 resumes the statistical sizing steps of the dummy bit line driver.

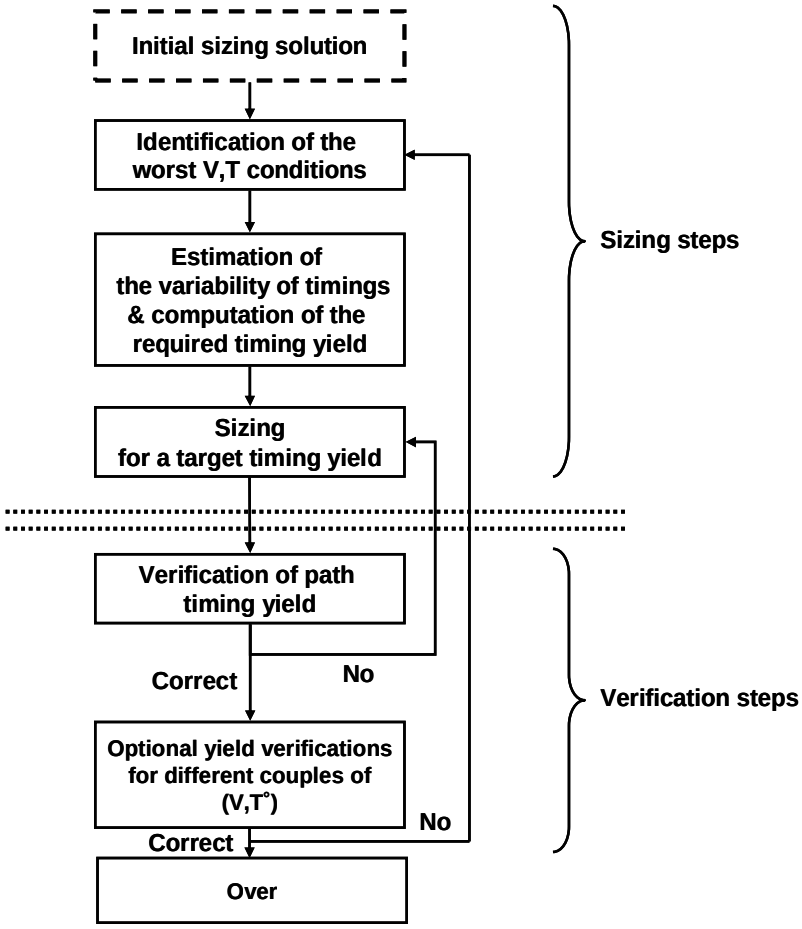

Fig. IV.7 Statistical sizing procedure of dummy bit line driver

It can be seen that the realization of the flow depends mostly on step 5 i.e. on the number of voltage and temperature conditions at which verifications are being made. In our case, we

have been verifying on average around 12 voltage and temperature conditions which took around 18 hours. However, a verification of all those conditions does not need to be performed. Thus, the complete realization of the statistical algorithm easily takes less than one day.

Although, the sizing method has been applied to the original dummy bit line driver structure, it can be generalized to other topologies of dummy bit line drivers. The statistical sizing procedure, introduced herein, will be used afterwards to compare the performances under a given timing yield constraint between the original dummy bit line driver and a new dummy bit line driver which is less sensitive to variability effects.

IV.1.3 Dummy bit line driver with reduced variance

In a more specific context involved in the design of advanced technologies of embedded SRAM memories, the corner methods seem no longer enough to satisfy the timing constraints without the use of an increasing timing margin caused by an increase of intra die variations. This fact brings up a question: Is it possible to maintain, or even reduce the design timing margins through design? More specifically, is it possible to lessen the rate of increase of the timing margins while taking into account during the design process the sensitivity performances with respect to PVT variations?

The solution put forward consisted in designing a new dummy bit line driver (DBD) displaying current performances : (a) less sensitive to manufacturing process variations compared to a classic dummy bit line driver [Gou06] illustrated in figure IV.9 (b) and (b) more similar sensitivities to supply voltage and temperature as a 6T SRAM cell (figure IV.9(c)).

IV.1.3.1 Utility of dummy bit line driver

The dummy bit line driver is an essential component of the embedded SRAM memory. In the absence of an internal clock signal, the dummy bit line driver acts as a metronome during a read cycle operation of the memory. It guarantees as shown in figure IV.8 (b), the proper triggering of the appropriate sense amplifier through signal dummy bit line, when the potential difference between its input signals BL and BLB has reached the required level for a correct read operation (around 10% of Vdd).

As we have already seen, paths A and B (figure IV.8 (a)) run across a series of different structures characterized by their own timing performances showing specific sensitivities to PVT dispersions. Thus, the dummy bit line driver needs ideally to discharge the dummy bit line such that input paths A and B of the sense amplifier are practically synchronized. Nevertheless, on a practical aspect, signal B lags behind signal A; the difference between path delays A and B representing the read timing margin. The proposed dummy bit line driver tracks better path delay A compared to the original dummy bit line driver, as it will be seen in section IV.1.3.3.

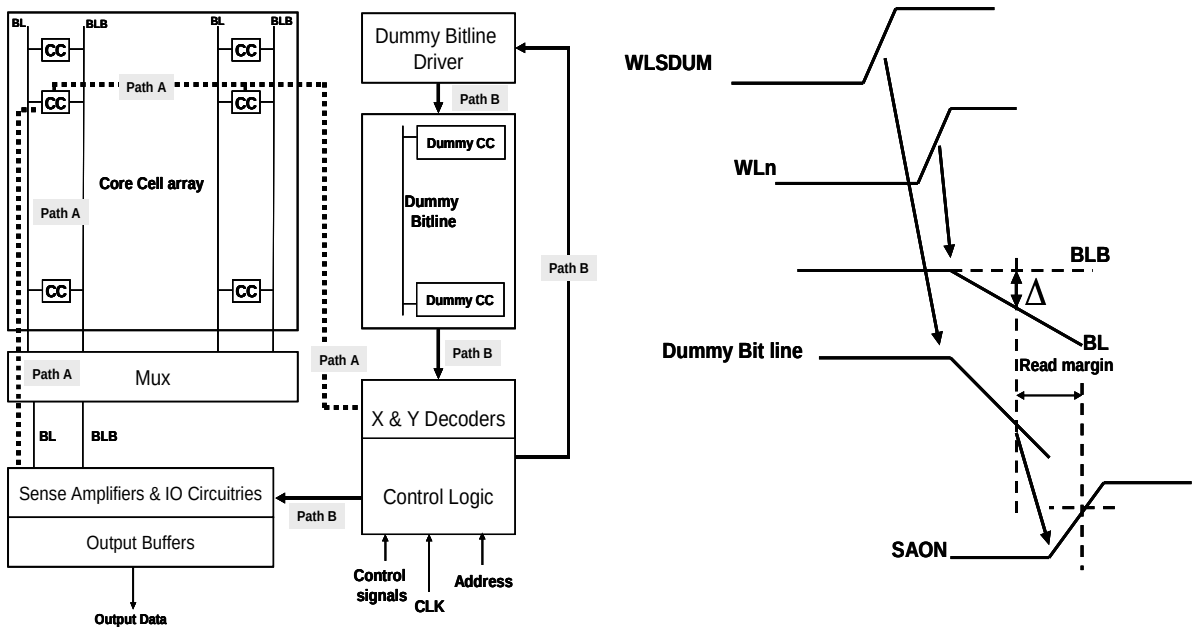

Fig. IV.8 (a) Signal races between paths A and B (b) Timing diagram of read operation

IV.1.3.2 Operating mode

The new dummy bit line driver is represented in figure IV.9 (a). It ensures the discharge of its output signal Out connected to the dummy bitline, which has been previously precharged to a potential of Vdd through transistor Pr. The discharge of Out occurs when signal WLSDUM is at its logic state '1' and when input pins Iadj are polarized at Vdd. Furthermore, this discharge is operated by a current whose value can be controlled with respect to the supply voltage, thanks to the use of a word comprising of 4 bits Iadji (i=1, 2, 3, 4). The word of 4 bits is either provided by the environment of the memory or individually hard coded, according to a predefined Vdd value. The main difference between the two structures is that the original

dummy bit line driver makes use of stacked transistors to represent the pass gate and the pull down of a 6T SRAM cell. Using stacked transistors means that the sensitivity of the total discharge current flowing through PGi and PDi (i=1, 2, 3, 4) is less representative of the discharge current flowing through PGcc1 and PDcc1 when bit line is being discharged (figure IV.9 (c)). In terms of operating mode, the original structure operates similarly as the proposed dummy bit line driver i.e. it discharges the node Out when WLSDUM is at '1' and the input pins Iadji are individually hard coded. It can also be noted that the original dummy bit line driver has been simplified compared to its original form, as signal rwb (figure I.24) has been replaced by WLSDUM, since only read operation is being analysed.

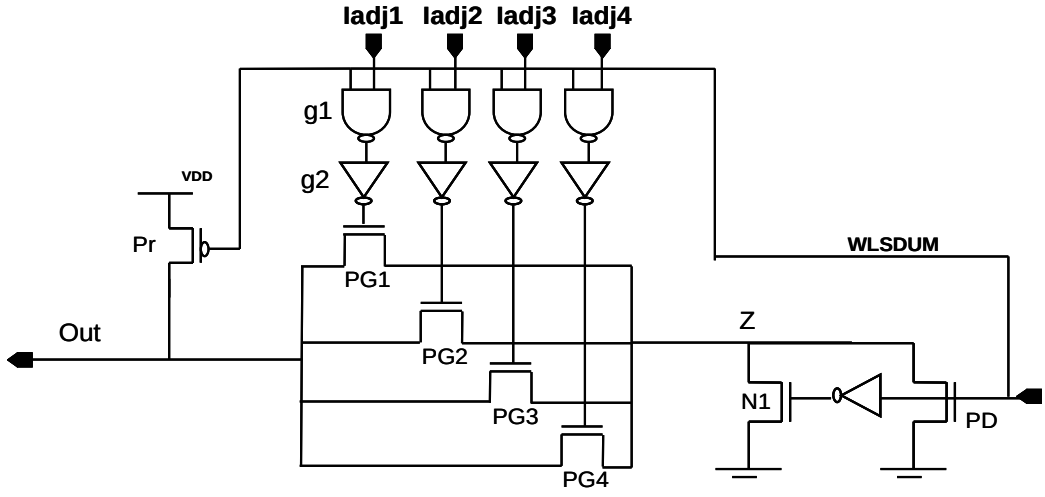

Fig. IV.9 (a) Proposed dummy bit line driver

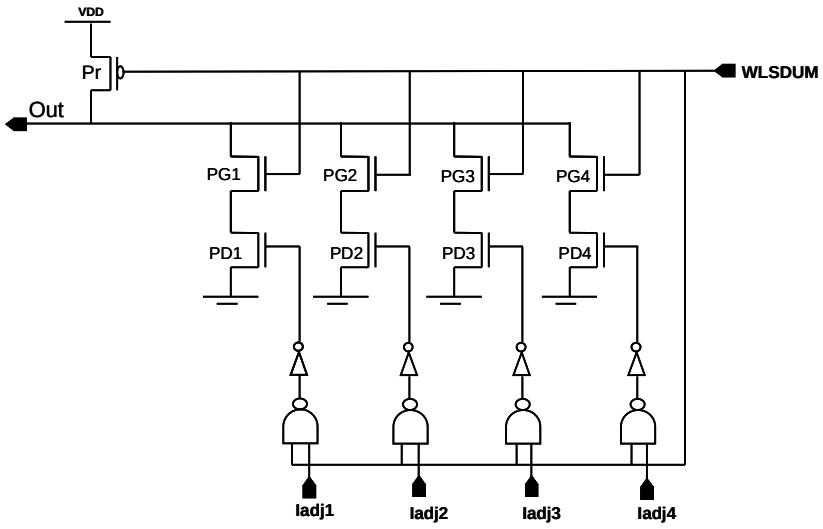

Fig. IV.9 (b) Original dummy bit line driver

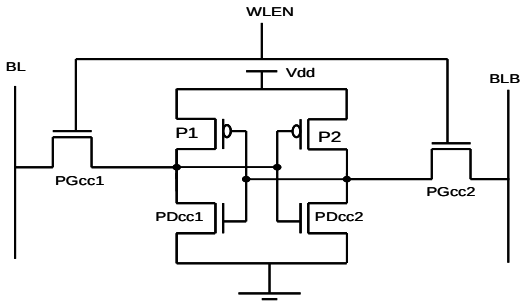

Fig. IV.9 (c) 6T SRAM cell

IV.1.3.3 Specific characteristics

The DBD introduced in figure IV.9 (a) has been designed in view of demonstrating specific sensitivities to supply voltages, temperatures and manufacturing processes. In fact, the topology of the DBD has been realized such that the discharge characteristics of dummy bit line match those of bit line being discharged by an SRAM cell. The aim of this approach is to obtain identical sensitivities, with respect to voltage and temperature, of the timing performances of the couples (dummy bit line driver, dummy bitline bit line) and (SRAM cell, bit line).

As shown in figures IV.9 (a) and IV.9 (c), transistors PD and PGi (i=1, 2, 3, 4) of the proposed dummy bit line driver are akin to transistors PDcc1 (2) and PGcc1 (2) of the SRAM cell. Moreover, logic gates g1 and g2 will mimic the signal WLEN which controls pass gate PGcc1 (2). The transistor Pr is used for precharging dummy bit line while transistor N1, of small size, sets node Z to 0 V at the beginning of a read cycle operation.

The sizing of the dummy bit line driver has been carried out in order to reduce the effect of local variations. This has been made possible through the use of pass gate transistors in the proposed strcture having their widths two to three times greater than the widths of pass gate transistors found in the reference dummy bit line structure. In this way, the variability of current performances in the structure is reduced.

IV.1.3.3.1 Sensitivities to vdd and temperature

To validate the choice of the new dummy bit line driver, the performances of both dummy bit line drivers have been compared. Assuming that an ideal dummy bit line driver displays similar temperature and voltage sensitivities as an SRAM cell, separate evaluations have been

carried out to investigate the dependency of the sensitivities of the propagation delays on temperatures and voltages. The delay measured on the dummy bit line drivers is between the activation of signal WLSDUM at Vdd/2 and the discharged of dummy bit line (connected to node out) at Vdd/2 (figure IV.9 (a) and (b)). For the SRAM cell, the delay measured is between the activation of signal word line WLEN at Vdd/2 and the discharge of bit line at Vdd/2 (figure IV.9 (c)). The following structures have been considered:

(i) A proposed dummy bit line driver (figure IV.9 (a)) controlling a dummy bit line comprised of 512 dummy SRAM cells (denoted as structure S1 afterwards),

(ii) A reference dummy bit line driver (figure IV.9 (b)) controlling a dummy bit line composed of 512 dummy SRAM cells (denoted as structure S2 afterwards),

(iii) An SRAM cell controlling a bit line with 511 SRAM cells (denoted as S3 afterwards).

The evaluation of the sensitivities has been performed at a nominal process. It should be noted that the temperature sensitivity of the metal, in which bit lines have been realized, has been considered. The sizings of structures in figures IV.9 (a) and IV.9 (b) have been done in such a way that the maximum current supplied to dummy bit line, by the dummy bit line drivers, shares absolutely the same value (All potentials Iadji being equal to Vdd).

Figures IV.10 (a) and IV.10 (b) summarize the results obtained. They represent sensitivities evolutions of $D_{S3/S1}$ and $D_{S3/S2}$ with respect to supply voltage and temperature; $D_{S3/S1}$ ($D_{S3/S2}$) corresponding to the difference in delay between structures S3 and S1 (S2) under different operating conditions of (Vdd, T°).

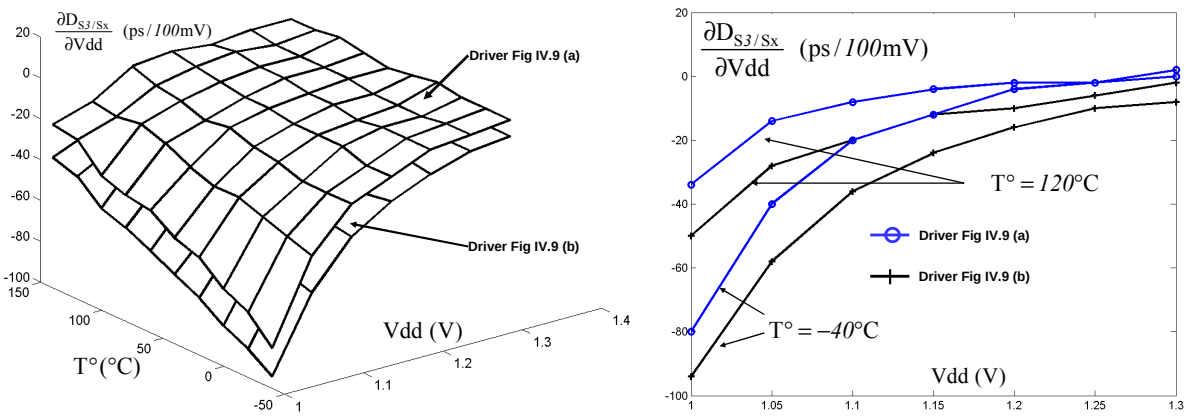

Fig. IV.10 (a) Sensitivities of D with respect to supply voltage

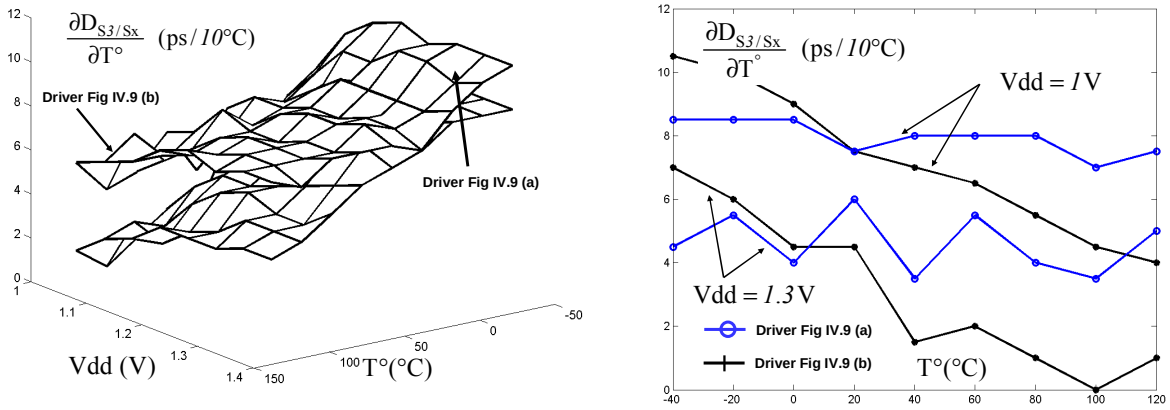

Fig. IV.10 (b) Sensitivities of D with respect to temperature

As illustrated in figure IV.10 (a), the sensitivity of $D_{S3/S1}$ (V, T°) with respect to supply voltage possesses smaller absolute values compared to the sensitivity of $D_{S3/S2}$ (V, T°). More precisely, the absolute value of the sensitivity of $D_{S3/S1}$(V,T°) to supply voltage is smaller respectively by 15 ps/100mV and 5 ps/100mV at Vdd=1.0V and Vdd=1.3V. Hence, we can conclude that the proposed bit line driver shows closer voltage behaviour to the SRAM cell compared to the driver in figure IV.9 (b).

However, as far as performances with respect to temperature variations are considered, both drivers considered demonstrate an almost identical sensitivity as shown in figure IV.10 (b). Indeed, the sensitivities of $D_{S3/S1}$ (V, T°) and $D_{S3/S2}$ (V, T°) to temperature have both values ranging from 1 ps/10°C to 10ps/10°C. Thus, the proposed driver does not exhibit a particular advantage as far as temperature variations are concerned.

IV.1.3.3.2 Sensitivity to process variations

To evaluate the dispersions in timing performances of structures in figures IV.9 (a) and IV.9 (b), we performed Monte Carlo simulations over a wide range of voltages and temperatures. The model card used in Hspice simulations is the bsim4.3.0 model which takes into account local and global variations. Moreover, 2000 runs were performed during each simulation.

IV.1.3.3.2.1 Dispersion values of discharge current

Those simulations provided us with the relative dispersion values σ_I/μ_I of the discharge current, summarized in table IV.2.

The relative dispersion value of the discharge current of the proposed dummy bit line driver is

globally weaker, except at low voltage and low temperature. The decrease in the variability value of σ_I/μ_I can reach 33% under certain temperature and voltage conditions, which gives rise to a subsequent decrease of variability in timing performances. This will be seen in further details later.

Table IV.2. Temperature conditions at which sizing should be performed for different operating voltages

Vdd (V)	Temperature (˚C)	σ_I/μ_I (%) (Fig. IV.9 (b))	σ_I/μ_I (%) (Fig. IV.9 (a))
1.0	-40	9.8	11.9
	27	9.4	9.6
	125	9.0	7.6
1.2	-40	6.8	6.3
	27	6.9	5.5
	125	7.0	4.9
1.32	-40	6.0	5.0
	27	6.1	4.6
	125	6.3	4.2

IV.1.3.3.2.2 Correlation values of propagation delays

The correlation values of the propagation delays between S1 (propagation delay of proposed dummy bit line driver) and S3 (propagation delay of SRAM cell) and S2 (propagation delay of reference dummy bit line driver) and S3 (propagation delay of SRAM cell) have also been measured as shown in table IV.3.

Results demonstrate that the proposed dummy bit line driver generally displays higher correlation value with respect to SRAM cell compared to reference dummy bit line driver. Thus, it can be inferred that proposed dummy bit line driver is less sensitive to process variations.

Table IV.3. Correlation values of propagation delays

Vdd (V)	Temperature (˚C)	ρ (S1, S3)	ρ (S1, S2)
1.0	-40	48	41
	27	49	40
	125	50	40
1.2	-40	45	43
	27	47	42
	125	48	42
1.32	-40	45	45
	27	46	45
	125	48	45

IV.1.3.3.3 Comparison of average power consumption

The last specific characteristic analysed between the reference and proposed dummy bit line drivers was in terms of average current consumption during the read cycle time of the memory. Although the proposed dummy bit line driver has been sized with pass gate transistors having twice or thrice the size of pass gate transistors in the reference dummy bit line driver, results show a slight increase in the relative current consumption (ΔI) between proposed driver ($<I_{prop}>$) and reference driver ($<I_{ref}>$), except at low voltage as shown in table IV.4. More specifically, we can see that at 1V, the relative increase lies between 4.2% and 8.8%

Table IV.4: Current consumption comparison of both DBDS

Vdd (V)	T° (°C)	$<I_{ref}>$/µA	$<I_{Prop}>$/µA	$\Delta I/<I_{ref}>$ (%)
1	-40	116.5	126.7	8.8
	27	109.4	117.9	7.8
	125	102.8	107.1	4.2
1.2	-40	180.0	181.3	0.7
	27	173.4	176.9	2.0
	125	163.0	166.4	2.1
1.32	-40	200.8	201.0	0.1
	27	197.5	198.9	0.7
	125	190.2	192.0	0.9

IV.1.3.4 Performance comparisons

In section IV.1.3.3, both dummy bit line drivers have been sized while imposing a similar discharge current. The principal idea lying behind this was to perform an initial comparison between both drivers, nevertheless this does not seem quite satisfactory. In fact, only the timing yield, the timing performance or even the design timing margin M_T represent pertinent comparisons if studied under different PVT conditions.

In order to perform comparisons under constant timing yield, we have sized both dummy bit line drivers using the statistical method introduced in IV.1.2.

IV.1.3.4.1 Performance comparisons without adjusting read timing margin to supply voltage

A first series of validation has been performed. It consisted in (a) substantiating the sizing methodology and (b) estimating the gain in a read cycle operation or in the timing yield through the use of the dummy bit line driver introduced in the previous section. The reference and proposed dummy bit line drivers have been alternatively used in the critical path of an SRAM, with the potential Iadji (i=1, 2, 3 and 4) fixed at Vdd.

IV.1.3.4.1.1 Impact on variability and timing yield

Both structures were sized, with the above stated polarization, to obtain a difference of μ_B-μ_A=200 ps at a typical process with a supply voltage of 1.2V and at a temperature of 27°C. The mean values (μ_A and μ_B) and the standard deviations (σ_A and σ_B) of the characteristic delays of signal races shown in figure IV.8 (a) 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.

Tables IV.5 and IV.6 report the simulation results obtained. In the results presented in table IV.5, the variability reduction ΔV_B lies between 1.2% and 2.6% and has a mean value of 1.8%. This corresponds to a 22% decrease of the relative variance $\Delta V_B/V_{B\,|Ref}$.

In fact, the decrease of variability in the timing performances of path B impacts on the probability of satisfying the timing constraint. To assess the effect of variability reduction, we have computed with the use of equation (III.10) the probability ($1-P^V$) that the timing constraint is not satisfied for different correlation coefficient values for both structures. Table IV.6 resumes the results calculated. A decrease in variability ΔV_B of mean value 1.8% yields a significant reduction in the probability of occurrence of a timing constraint violation as illustrated in table IV.6. For instance at 1V, 125˚C and ρ=0.75, the reduction in the probability of the timing constraint violation is reduced by a factor of 8.8 times for a relative variance decrease of 30%.

Table IV.5 Variability reductions (REF: Reference DBD, Prop: Proposed DBD)

Vdd (V)	T (°C)	μ_A (ps)	V_A (%)	μ_B^{Ref} (ps)	V_B^{Ref} (%)	μ_B^{Prop} (ps)	V_B^{Prop} (%)	ΔV_B (%)
1	-40	1919	9	2421	12	2368	10	1.7
1.08		1534	7	1863	9	1834	7	1.9
1.18		1228	6	1464	8	1456	6	1.7
1.26		1061	6	1264	7	1266	5	1.4
1.32		963	5	1153	6	1160	5	1.3
1	0	1994	9	2456	11	2048	9	2.1
1.08		1609	7	1943	9	1916	7	2.1
1.18		1236	6	1558	7	1551	6	1.7
1.26		1123	6	1357	7	1360	5	1.5
1.32		1021	5	1245	6	1252	5	1.3
1	40	2065	8	2487	10	2450	8	2.4
1.08		1680	7	2007	9	1994	7	2.1
1.18		1361	6	1637	7	1639	6	1.7
1.26		1183	5	1438	7	1447	5	1.5
1.32		1078	5	1324	6	1336	5	1.3
1	80	2134	8	2511	10	2495	7	2.5
1.08		1750	7	2060	8	2066	6	2.2
1.18		1427	6	1700	7	1721	5	1.7
1.26		1244	5	1504	6	1529	5	1.4
1.32		1136	5	1389	6	1417	5	1.3
1	125	2209	8	2532	10	2546	7	2.6
1.08		1828	7	2106	8	2142	6	2.1
1.18		1500	6	1757	7	1804	5	1.7
1.26		1313	5	1563	6	1613	5	1.4
1.32		1201	5	1448	6	1500	5	1.2

Table IV.6 Probability of a timing constraint violation

Vdd (V)	T (°C)	$(1-P^V)$ for $\rho=0.95$ ($\times 10^{-9}$)		$(1-P^V)$ for $\rho=0.75$ ($\times 10^{-6}$)	
		Ref	Prop	Ref	Prop
1	-40	10358	10	2666	1123
1.18		543	0	344	11
1.32		0	0	7	0
1	0	8077	0	3347	925
1.18		0	0	3	0
1.32		0	0	1	0
1	125	187034	0	19671	2239
1.18		8	0	513	0
1.32		0	0	3	0

IV.1.3.4.1.2 Reduction of read cycle time at constant timing yield

Relying upon previous results, it seems possible to reduce the read cycle time of the memory by decreasing the read timing margin. This is achieved by using the proposed dummy bit line driver (DBD), while ensuring a similar timing yield as the reference DBD.

In order to evaluate the potential gain in terms of read cycle time, the proposed DBD and the reference DBD have been sized so as to satisfy a probability of 99.87% of meeting the timing constraint ($n=3.\sigma$). The sizing has been performed with the proposed sizing methodology, introduced in section IV.1.2. Then, the resulting values of μ_D, σ_D and P^V have been computed over a whole range of temperatures and voltages. Table IV.7 summarizes the overall results. The calculated probabilities of fulfilling the constraint are very close to the expected 99.87% value, thus validating the statistical sizing method. Simultaneously, we observe a reduction in the read timing margin $\Delta\mu_D$ (normalized with respect to path delay A, μ_A) lying between 2.4% and 6.4% of μ_A. This decrease corresponds to an average reduction of 4 % in the read cycle time of the memory.

Table IV.7. Reduction of the read timing margin (REF: Reference DBD, Prop: Proposed DBD)

Vdd (V)	T (°C)	μ_A (ps)	μ_D^{Ref} (ps)	P^V_{Ref} (%)	μ_D^{Prop} (ps)	P^V_{prop} (%)	$\Delta\mu_D/\mu_A$ (%)
1.0	-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

IV.1.3.4.2 Performance comparisons with adaptation of read timing margin to supply voltage

Just like the first series of validation, the second series of validation consisted in (a) corroborating the sizing methodology and (b) estimating the gain in the read cycle time brought by the use of the dummy bit line driver, while adjusting the current supplied by the DBD to the applied voltage. An adjustment of the supply current of the DBD also implies adapting the read margin to the actual applied supply voltage. To perform the comparison, both reference and proposed DBDs have been introduced in the critical path.

IV.1.3.4.2.1 Sizing of DBDs

The supply voltage range considered, varying from 1.0V to 1.47V, has been arbitrarily split over four ranges of voltages: GV1= [1, 1.08[, GV2= [1.08, 1.2[, GV3= [1.2, 1.32[and GV4= [1.32, 1.47[. After this dichotomy procedure, pass gate transistors PG1 to PG4 of the proposed DBD have been successively sized with the method developed in section IV.1.2. This methodology has been sequentially applied to each voltage range GVi (i=1 to 4), starting with GV1. The timing yield had been set at 99.87% i.e. n=3 and the correlation value ρ considered was equal to 0.9.

IV.1.3.4.2.2 Reduction of the read margin at constant timing yield

Once the statistical sizing method has been performed, we have run Monte Carlo simulations in order to obtain the mean values (μ_A et μ_B) and standard deviation values (σ_A et σ_B) of the characteristic delays of the signal racing conditions between paths A and B. The statistical simulation has been carried out over the whole voltage and temperature ranges [-40°C, 125°C] considered. The results obtained were used to compute μ_D and σ_D (III.8, III.9) and the probability P^V (III.10) of meeting the timing constraint.

Figure IV.11 shows the behaviour of the read timing margin μ_D for temperatures ranging from -40°C to 125°C, and for supply voltages varying between 1.0V to 1.47V. The surface in the figure represents the evolution of μ_D, with respect to both temperature and supply voltage variations, when the current supplied by the proposed DBD is adjusted to the supply voltage of the memory. Similarly, the other curve illustrates the variation of μ_D when the current supplied by the proposed DBD is not adjusted to supply voltage. In other words, it represents

the case when we polarize the potentials Iadji (i= 1, 2, 3, 4) such that the constraint is fulfilled over the whole range of applied voltages [1V, 1.47V] and temperatures [-40°C, 125°C].

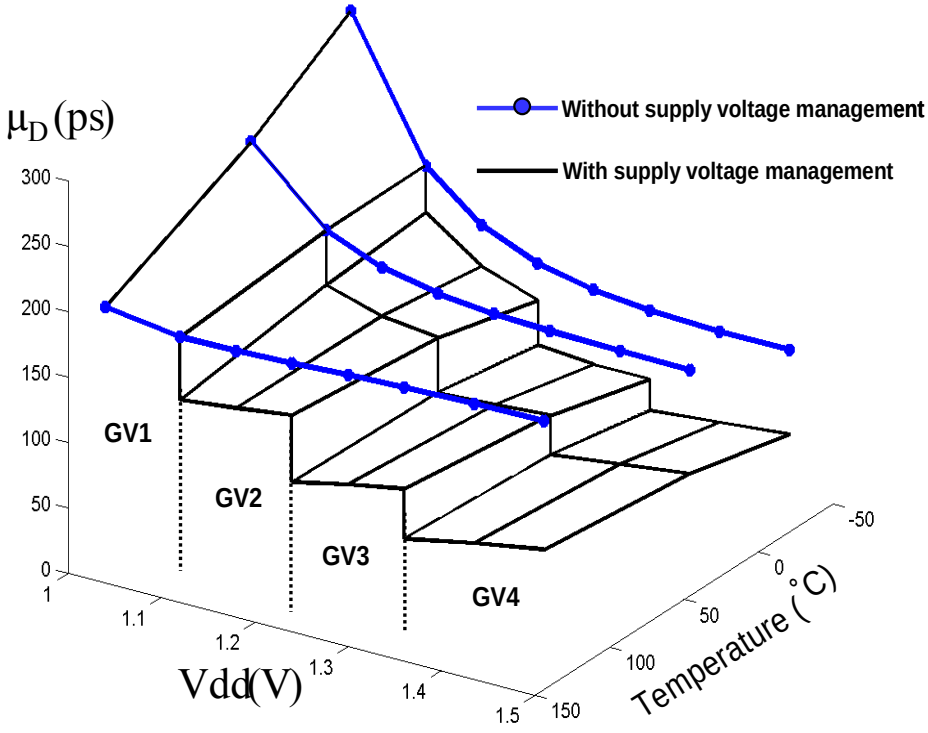

Fig. IV.11 Evolution of read timing margin at constant timing yield

As expected, the adjustment of the supply current to the supply voltage results in a significant decrease in the read timing margin value, over the voltage range GV2 to GV4. The reduction of μ_D ($\Delta U/U$) in figure IV.12 fluctuates between 20% for GV2 and 63% for GV4 with voltage and temperature fluctuations. In fact, the maximum reduction of 63% in μ_D at 1.32V and 125˚C corresponds to 10% of the time taken by an SRAM cell to discharge its respective bit line by 70mv under the same temperature and voltage conditions.

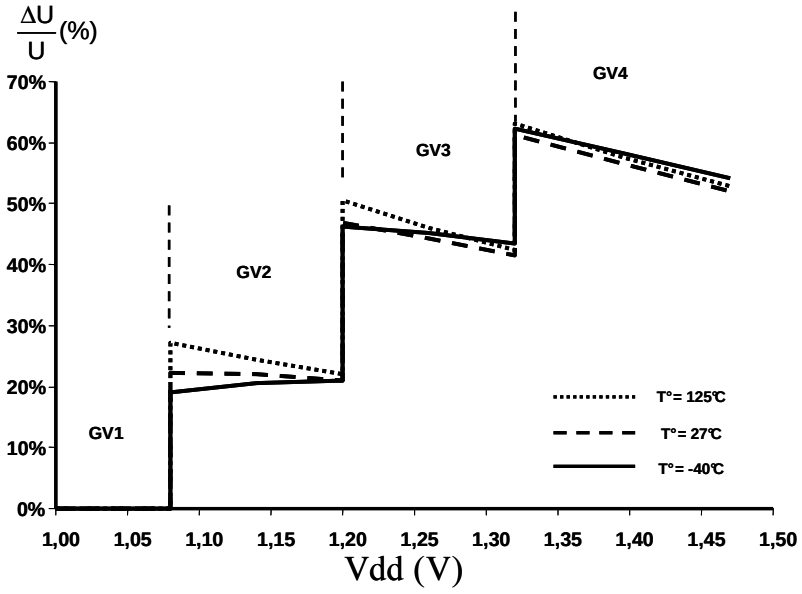

Fig. IV.12 Reduction in read timing margin with adjustment of supply current to supply voltage

This significant decrease in μ_D value entails a proper controlled reduction of the read timing yield. The results in figure IV.13 represent clearly the tendency of the probability of satisfying the timing constraint P^V with respect to Vdd and temperature, as we adapt or not input potentials Iadji to supply voltage.

More specifically, the surface corresponds to the evolution of P^V with Vdd and temperature variations when potentials Iadji are adjusted with respect to the supply voltage of the memory. On the other hand, the other curve reflects the behaviour of P^V when the potentials Iadji are chosen such that the timing constraint is fulfilled over a wide range of voltages [1V,1.47] and temperatures [-40°C, 125°C]. As it can be seen, the probability of meeting the timing constraint is close to the values (99.87%) imposed by the timing yield constraint, thereby confirming the sizing methodology introduced in section IV.1.2 i.e. the sizing methodology works well at constant and varying read timing margins. Besides this aspect, we notice that a change in polarization of bits Iadji is accompanied by the appropriate probability of fulfilling the timing constraint.

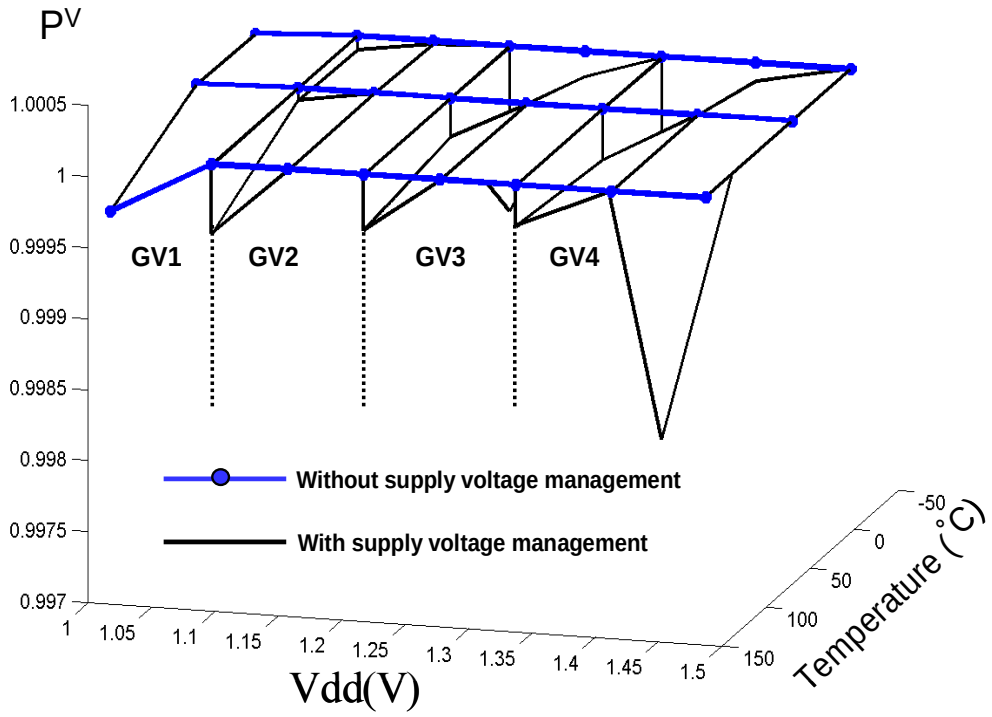

Fig. IV.13 Evolution of P^V with respect to temperature and voltage variations

IV.1.3.4.2.3 Comparisons with the reference DBD

The overall sizing procedures and validations undergone by the proposed DBD were also applied to the reference DBD. This has been performed in order to analyse the gain in reduction of μ_D brought by a maximum decrease of 30% (section IV.1.3.3.2) of the relative variability σ_I/μ_I of the current supplied by DBD in the discharge process of dummy bit line. The first series of comparisons are in terms of the variability reductions in the read timing margin (ΔV_B), when the discharge current of both DBDs are varied with supply voltage variations. Table IV.8 reports the results. The first column Iadji corresponds to the respective branches of transistors selected with respect to supply voltage. For instance Iadji=1, 2 (figure IV.9 (a)) means that branches Iadj1 and Iadj2 are selected at Vdd=1.08V and Vdd=1.14V. The relative reduction in variability is quite important, lying between 5.8% and 24.7%. This reduction has been achieved by using pass gate transistors in the proposed DBD which is 2 to 3 times the sizes of pass gates used in the reference DBD.

Table IV.8. Reduction of the read timing margin between reference (REF) and proposed (Prop) DBDs with voltage adaptations

Iadji	Vdd (V)	T (°C)	μ_A (ps)	V_A (%)	μ_B^{Ref} (ps)	V_B^{Ref} (%)	μ_B^{Prop} (ps)	V_B^{Prop} (%)	$\Delta V_B/V_B^{Ref}$ (%)
1	1.00	-40	1919	9.0	2412	10.8	2303	9.8	9.3
		27	2041	8.5	2456	9.9	2358	8.1	18.4
		125	2209	7.7	2497	9.1	2453	6.8	24.7
1, 2	1.08	-40	1534	7.5	1793	8.5	1734	7.5	11.7
		27	1657	7.2	1914	8.1	1849	6.7	18.0
		125	1827	6.7	2029	7.7	1998	6.0	22.6
	1.14	-40	1334	6.6	1540	7.5	1494	6.4	14.4
		27	1451	6.5	1678	7.3	1626	5.9	18.9
		125	1616	6.2	1813	7.0	1793	5.5	22.0
1, 2, 3	1.20	-40	1181	6.0	1306	6.9	1277	5.9	13.7
		27	1291	5.9	1445	6.7	1404	5.5	17.6
		125	1448	5.7	1585	6.5	1565	5.2	19.8
	1.26	-40	1060	5.6	1176	6.2	1151	5.3	14.4
		27	1164	5.5	1315	6.2	1279	5.1	17.5
		125	1312	5.3	1457	6.1	1441	4.9	18.8
1, 2, 3, 4	1.32	-40	963	5.2	1043	5.4	1023	5.1	5.8
		27	1060	5.1	1177	5.5	1139	4.9	10.1
		125	1200	5.0	1315	5.5	1289	4.8	13.3
	1.47	-40	825	4.6	910	4.8	894	4.5	6.1
		27	911	4.6	1035	4.9	1004	4.5	9.7
		125	1037	4.5	1168	5.1	1147	4.5	12.0

The other performance comparisons realized is in terms of reduction in the read timing margins. Table IV.9 resumes the results obtained. It represents the relative decrease of the value of μ_D, $(\mu_D^{Ref}-\mu_D^{Prop})/\ \mu_D^{Ref}$. It also reports the values of $(\mu_D^{Ref}-\mu_D^{Prop})/\mu_A$ so as to quantify the impact of the reduction of the design timing margin. Actually, μ_A represents the mean delay required by an SRAM cell to drive the difference in potential of the input signals BL and BLB of the sense amplifiers at the required voltage for a proper read operation.

Table IV.9 Impact of the decrease in the relative variability of the current of DBD

Iadji	Vdd (V)	T (°C)	$(\mu_D^{Ref}-\mu_D^{Prop})/\ \mu_D^{Ref}$ (%)	$(\mu_D^{Ref}-\mu_D^{Prop})/\ \mu_A$ (%)
1	1.00	-40	22	5.7
		27	24	4.8
		125	15	2.0
1, 2	1.08	-40	23	5.7
		27	25	4.8
		125	16	1.6
	1.14	-40	22	4.4
		27	23	4.5
		125	10	1.9
1, 2, 3	1.20	-40	23	3.9
		27	27	4.1
		125	15	1.4
	1.26	-40	21	3.4
		27	24	3.7
		125	11	0.9
1, 2, 3, 4	1.32	-40	24	3.0
		27	32	3.9
		125	23	1.7
	1.47	-40	16	2.7
		27	24	3.8
		125	15	1.4

We observe that a decrease of σ_I/μ_I results in a minimum and maximum relative reductions of μ_D ranging from 10% to 32%. These relative reductions are significant as far memory performances are concerned, since they correspond to a decrease going from 1% to 3% of the read cycle time of a memory. The results obtained illustrate well that considering variability performances can ensue in a substantial decrease of the design margins even at low supply voltage. However, these results are only obtainable with statistical design techniques.

IV.1.3.4.2.4 Statistical method vs. corner method

In the previous subsection, we have validated the optimization method of the dummy bit line driver under a constant timing yield constraint. After developing the design method, which statistically sizes the DBD at a typical process, we wanted equally to evaluate the resulting timing margins for the best and worst case corners of our structures.

We have compared the values of μ_D obtained during a performance analysis with a typical process and values of μ_D^{Corner} obtained for a worst case analysis (slow process) and a best case

analysis (fast process) at different temperatures and voltages. Figure IV.14 shows the evolution of the ratio μ_D^{Corner} / μ_D with voltage variations for three values of temperatures T° (-40°C, 27°C and 125°C).

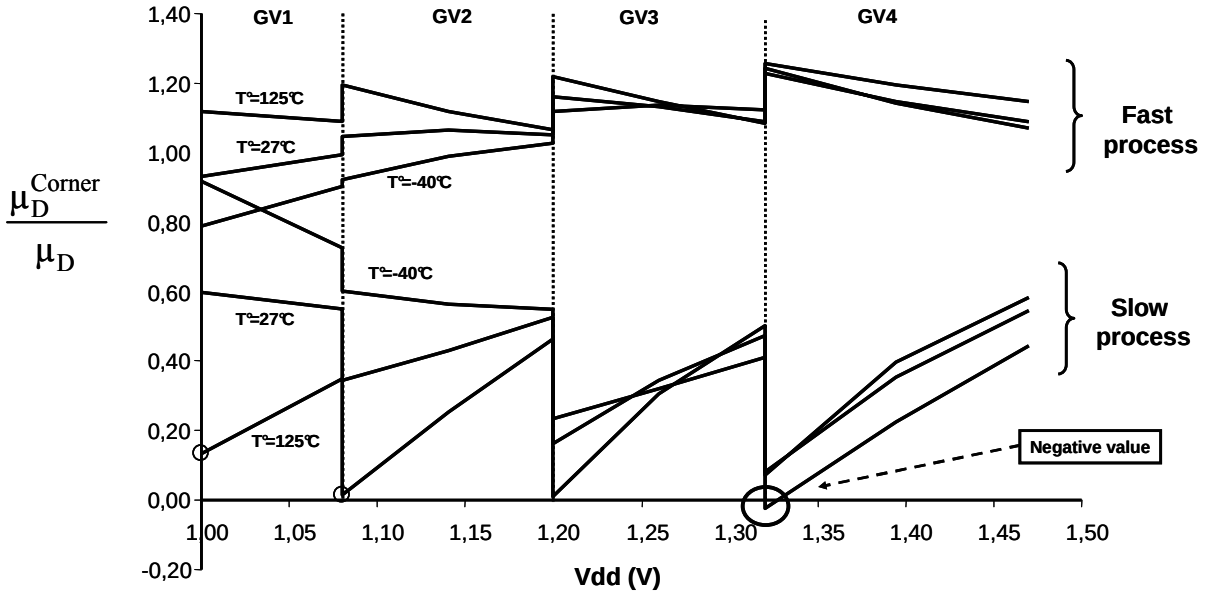

Fig. IV.14 Evolution of μ_D^{corner}/μ_D with respect to supply voltage variations at 3 different temperatures

It can be observed that values of μ_D^{Corner} obtained with a slow and fast process both diverge from the typical process. However, the values of μ_D^{Corner} obtained at a slow process are clearly smaller than μ_D. It can be seen in the above figure that the ratio of $\mu_D^{Corner}/\ \mu_D$ is negative at Vdd=1.32V and T°=125°C. The negative value suggests the pessimism associated with process corners. According to the corner analysis method, a timing constraint violation is said to occur at Vdd=1.32V when the process is of bad quality and at high temperature. This result is in total contradiction with Monte Carlo results, even when considering a statistical correlation coefficient of 0.8 between both path delays.

Conclusion

In this chapter, we have seen two examples of the applications of the modelling approach. The first application is in fact an interesting one, since it allows us to detect critical areas of the memory which are most likely to undergo timing constraint violations in presence of process variations. The second application, which consists in proposing a statistical sizing methodology of the dummy bit line driver, gives us an appropriate manner of sizing the

dummy bit line driver in the presence of local variations. The statistical characteristic of this method enables the exploration of design methodologies which are out of reach through the use of process corners. Simultaneously, we have proposed a novel dummy bit line driver showing less sensitivity with respect to process and voltages variations. Simulation results have demonstrated that the use of the statistical sizing method and the proposed driver can drastically improve the performance of the memory in terms of reduction in the read timing margins, while ensuring a high timing yield.

Conclusion

Conclusion

Performances of SoCs, especially in terms of manufacturing yield and reliability, are mainly dependent on embedded memories like SRAMs, which will keep on occupying a larger proportion of the chip. Understanding functionality of SRAMs, in view of improving their robustness in presence of variability aspects, has therefore become a necessary evil for designers particularly in the presence of variability conditions.

As transistors geometries continue to scale unrelentingly, the emergence of variability phenomena owing to process, voltage and temperature variations seems more than ever to limit the performances of ICs. Undoubtedly, those yield detractors constitute a severe hindrance and a major challenge for designers who need to design their circuits in a shorter time to market and a shorter time to volume, owing to market pressure continuing to amount.

To handle the impact of manufacturing process variations along with the operating conditions in circuit design, the corner based methodology has often been the mainstay for designers who would adopt conservative margins for guard banding their designs against variability conditions. However, if corner analysis method has proven to be a reliable method in the past, it now appears as being obsolete in ensuring a feasible design.

The exaggerated pessimism of corner analysis entails unnecessary large chip area, excessive power dissipation and increases design effort. It can be inferred that corner analysis will virtually negate the benefit of transistor scaling as the overall performance of the design is reduced with the corner method. Moreover, these sources of variations (process, voltage and temperature) are too numerous and complex to be captured within a small set of process corners, thus making corner analysis inadequate to account for the impact of local variations on the timing performances of signal races. Indeed, in the presence of signal racing conditions described in chapter three, the corner method considers a correlation of one between the two main path delays involved in the read cycle operation. This assumption gives rise to an overestimation of the probability of carrying out successfully a read operation.

Statistical methodologies and SSTA tools have come forth as an alternative to corner analysis for timing analysis. The statistical approach allows the possibility of breaking the barriers of corner method and to model holistically factors affecting process variations in a single

analysis run [Pra06]. Consequently, we have tailored a statistical optimization methodology of the signal races ensuring a constant timing yield. This method is based on the traditional methods, since the optimization is carried out while considering a typical process rather than a slow or a fast process. The optimization method proposed has been applied to the principal signal races of a 256Kb embedded SRAM, in a 90nm technology node. This approach has been particularly introduced to optimize the critical path of the SRAM, in which the reference dummy bit line driver has been replaced by a more robust structure to manufacturing process variations and to supply voltages.

Results have demonstrated that the use of the optimization method and the proposed dummy bit line driver improves significantly the reduction in the design timing margins, for instance the reduction in the relative timing margins lies between 10% to 32% when compared to the original dummy bit line driver. This represents on average a 3% decrease in the read cycle time of the memory.

Another main advantage of the statistical sizing methodology, seen in chapter four, resides in its statistical characteristic that allows the exploration of design methodologies which are out of reach through the use of process corners. After statistically sizing our dummy bit line driver, we performed a corner analysis of the memory at a slow process, high temperature (125˚C) and high voltage (1.32V). Corner results have shown that at these conditions, a timing constraint violation has occurred, which is in fact in total contradiction with Monte Carlo results that indicate a positive read timing margin at the same temperature and voltage conditions.

As some future prospects, it would be interesting to compare statistical results derived from the modelling approach with results obtained from SSTA tools like Solido stat applied to the memory to optimize its performances in presence of variability conditions. Furthermore similar performance comparisons, as described in the second application, could also be performed in other SRAM configurations like tall and normal configurations and in more complex technologies (65nm and 45nm) in order to see the efficiency of the modelling approach.

References

References

[Adi01] B. Adibi et al., "Transistor performance: The impact of implant doping accuracy," Solid State Tech., Vol. 44, No. 1, pp. 68-69, Jan 2001.

[Aga06] k. Agarwal, S. Nassif, "Statistical analysis of SRAM cell stability," Design Automation Conference, 2006 43rd ACM/IEEE, pp. 57-62, July 2006.

[Amr98] B. S. Amrutur and M. A. Horowitz "A Replica Technique for Wordline and Sense Control in Low-Power SRAM's," IEEE Journal of Solid-State Circuits, Vol. 33, No. 8, pp. 1208-1219, Aug 1998.

[Aro95] N. D. Arora, R. Rios, and C. L. Huang, "Modeling the Polysilicon Depletion Effect and Its impact on Submicrometer CMOS Circuit Performance," IEEE Transactions on Electron Devices, Vol. 45, No. 5, pp. 935-943, May 1995.

[Ase98] A. Asenov, "Random Dopant Induced Threshold Voltage Lowering and Fluctuations in Sub-0.1µm MOSFET's: A 3-D "Atomistic" Simulation Study," IEEE Transactions on Electron Devices, Vol. 45, No. 12, pp. 2505-2513, Dec 1998.

[Ase03] A. Asenov, S. Kaya, and A. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," IEEE Transactions on Electron Devices, Vol. 50, No. 5, pp. 1254-1260, May 2003.

[Bhu07] S. Bhunia, S. Mukhopadhyay, and K. Roy, "Process Variations and Process-Tolerant Design," 20th International Conference on VLSI Design, pp. 699-704, Jan 2007.

[Bor03] S. Borkar et al., "Parameter Variations and Impact on Circuits and Microarchitecture," Proc. 40th Design Automation Conference, pp. 338-342, 2003.

[Bor05] S. Borkar, "Designing Reliable Systems from Unreliable Components: The Challenges of Transistor Variability and Degradation," IEEE Micro, Vol. 25, No. 6, pp.10-16, Dec. 2005.

[Bos80] D. C. Bossen, and M. Y. Hsiao, "A System Solution to the Memory Soft Error Problem," IBM J. Res. Develop, Vo. 24, No. 3, pp. 390-397, May 1980.

[Bow02] K. Bowman, S. Duvall, and J. Meindl, "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration," IEEE J. Solid-State Circuits, Vol. 37, No. 2, pp. 183-190, Feb 2002.

[Bru97] T. A. Brunner, "Impact of lens aberrations on optical lithography," IBM J. Res. Develop, Vol. 41, No. 12, pp. 57-67, Jan/March 1997.

[Cha03] H. Chang, and S. S. Sapatnekar, "Statistical Timing Analysis Considering Spatial Correlations Using a Single PERT-Like Traversal," Proceedings of the International Conference on Computer Aided Design, pp. 621-625, 2003.

[Che06] X. Chen, and D. Velenis, "Effects of Parameter Variations on Low-Power SRAM Decoder," International Review of Electrical Engineering, Vol. 1, No. 2, pp. 247-253, May-June 2006.

[Cho07] M. Choi, "Modelling of Deterministic Within-die Variation in Timing Analysis, Leakage Current Analysis, and Delay Fault Diagnosis," Ph.D Thesis, Georgia Institute of Technology, May 2007.

[Cro05] J. A. Croon, W. Sansen, and H. E. Maes, "Matching properties of Deep Sub Micron MOS transistors," Springer edition, 2005.

[Dup02] E. Dupont, M. Nicolaidis, and P. Rohr, "Embedded Robustness IPs for Transient-Error-Free ICs," IEEE Design and Test Computers, Vol. 19, No. 3, pp. 54-68, May 2002.

[Eis97] 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, Dec 1997.

[Ele05] http://electronicdesign.com (Timing analysis rounds the corner to statistics)

[Gat01] A. Gattiker et al., "Timing Yield Estimation from Static Timing Analysis," Proceedings of the IEEE International Symposium on Quality Electronic Design, pp. 437-442, 2001.

[Gou06] V. Gouin et al., "Memory Circuit with Supply Voltage Flexibility and Supply Voltage Adapted Performance," US patent US 2006/0050572 A1.

[Gray00] K. Gray, "Adding error correcting circuitry to ASIC memory," IEEE Spectrum, Vol. 37, No. 4, pp. 55-60, April 2000.

[Har01] T. P. Haraszti, "CMOS Memory Circuits," Kluwer Academics Publishers (Second edition), 2001.

[Hum06] E. Humenay, D. Tarjan, and K. Skadron, "Impact of Parameter Variations on Multi-Core Chips," Proc. Workshop on Architectural Support for Gigascale Integration, June 2006.

[Ito01] K. Itoh, "VLSI Memory Chip Design," Springer edition, 2001.

[Kan98] Yi-Kan Cheng, P. Raha, C. C. Teng, E. Rosenbaum, and S. M. Kang, "ILLIADS-T: an Electrothermal Timing Simulator for Temperature Sensitive Reliability Diagnosis of

CMOS VLSI chips" IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 17, No. 8, pp. 668-681, Aug 1998.

[Kay01] S. Kaya, A. Brown, A. Asenov, D. Magot, and T. Linton, "Analysis of statistical fluctuations due to line edge roughness in sub-0.1 μ MOSFETs," Proc. 2001 International Conference on Simulation of Semiconductor Processes and Device, pp. 78-81, 2001.

[Kha06] A. B. Khang et al., "Lens Aberration Aware Timing-Driven Placement," Design, Automation and Test in Europe, Vol. 1, pp. 1-6, March 2006.

[Khe06] M. Khellah et al., "Wordline and Bitline Pulsing Schemes for Improving SRAM Cell Stability in Low-Vcc 65nm CMOS Designs," IEEE Symposium on VLSI Circuits Digest of Technical Papers, pp. 9-10, 2006.

[Kim98] I. Kim, Y. Zorian, G. Komoriya, H. Pham, F. P. Higgins, and j. Lewandowski, "Built in Self Repair for Embedded high Density SRAM," Proc. IEEE International Test Conference (ITC), pp. 1112-1119, 1998.

[Kim03] N. S. Kim et al., "Leakage Current: Moore's Law Meets Static Power," IEEE Computer Society, Vol. 36, No. 12, pp. 68-75, Dec. 2003.

[Ick] http://www.icknowledge.com/misc_technology/CMP.pdf

[Koc02] M. Kocher, and G. Rappitsch, "Statistical Methods for the Determination of Process Corners," Proceedings of the IEEE International Symposium on Quality Electronic Design, pp. 133-137, 2002.

[Kum06] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Impact of NBTI on SRAM Read Stability and Design for Reliability," Proceedings of the 7th International Symposium on Quality Electronic Design, pp. 210-218, March 2006.

[Las07] B. Lasbouygues, R. Wilson, N. Azémard, and P. Maurine, "Temperature and Voltage Aware Timing Analysis," IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, Vol. 26, No. 3, pp. 801-815, March 2007.

[Liao05] W. Liao, L. He, and K. M. Lepak, " Temperature and Supply voltage Aware Performance and Power Modelling at Microarchitectural level", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 24, No. 7, pp. 1042-1053, July 2005.

[Luo07] H. Luo et al., "Modeling of PMOS NBTI Effect Considering Temperature Variation," Proceedings of the 8th International Symposium on Quality Electronic Design, pp. 139-144, 2007.

[Mag] http://www.magma-da.com/c/@0R9zP3c8uXvBE/Pages/QuartzSSTA.html

[Mal87] W. Maly, "Realistic Fault Modelling for VLSI Testing," Proceedings of ACM/IEEE Design Automation Conference, pp. 173-180, 1987.

[Mal96] W. Maly, H. Heineken, J. Khare and P. K. Nag, "Design for manufacturability in submicron domain," Proc. of ICCAD 96, pp. 690-697, Nov. 1996.

[Mak07] T.M. Mak, "Infant Mortality-The lesser Known Reliability issue**,"** Proc. of the 13th IEEE International On-Line Testing Symposium, pp. 122, July 2007.

[Mar99] M. Margala, "Low-Power SRAM Circuit Design," Proc. IEEE MTDT Workshop, pp. 115-122, Sept 1999.

[Mcp06] J. W. Mcpherson, "Reliability Challenges for 45nm and Beyond," Proceedings of the 43rd ACM/IEEE DAC, pp. 176-181, July 2006.

[Miz94] T. Mizuno, J. I. Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant numbers in MOSFET's," IEEE Trans. Electron Devices, Vol. 41, No. 11, pp. 2216-2221, Nov 1994.

[Muk05] S. Mukhopadhyay and K. Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS," IEEE Trans. Comput-Aided Des. Integr. Circuits Syst., Vol. 24, No. 12, pp. 1859-1880, Dec 2005.

[Mut07] A. Mutlu, et al. "An Exploratory Study on Statistical Timing Analysis and Parametric Yield Optimization," Proceedings of the IEEE International Symposium on Quality Electronic Design, pp. 677-684, 2007.

[Nag93] P. K. Nag, and W.Maly, "Yield Learning Simulation," Proc. of TECHCON, pp. 280-282, Oct 1993.

[Nas84] S. R. Nassif, A. J. Strojwas, and S. W. Director, "FABRICS-II: A statistical based IC fabrication process simulator," IEEE Transactions on Computer Aided Design, Vol. 3, No. 1, pp. 40-47, Jan 1984.

[Old00] P. Oldgies et al., "Modeling line edge roughness effects in sub 100nm gate length devices," Proc. of Simulation of Semiconductor Processes and Devices (SISPAD), pp. 131-134, Sept 2000.

[Ouy00] C. Ouyang, K. Ryu, L. Milor, W. Maly, G. Hill, and Y. K. Peng, "An Analytical Model of Multiple ILD Thickness Variation Induced by interaction of Layout Pattern and CMP Process," IEEE Transactions on Semiconductor Manufacturing, Vol. 13, No. 3, pp. 286-292, Aug 2000.

[Pap07] A. Papanikolaou et al, "Reliability issues in deep deep sub-micron technologies: time-dependent variability and its impact on embedded system design," Proc. of the 13th IEEE International On-Line Testing Symposium, pp. 121, July 2007.

[Pra06] EETIMES 2006, "Practical Applications of Statistical Timing Analysis," http://www.eetimes.com/showArticle.jhtml?articleID=196700482

[Pau07] B. C. Paul et al., "Negative Bias Temperature Instability: Estimation and Design for Improved Reliability of Nanoscale Circuits, " IEEE Transactions on Computer Aided Design of integrated Circuits and Systems, Vol. 26, No. 4, pp. 743-751, April 2007.

[Pel89] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors." IEEE Journal of Solid State Circuits, Vol. 24, No. 5, pp. 1433-1439, 1989.

[Ram97] N. H. Ramadan, "Redundancy Yield Model for SRAMS," Intel Technology Journal, Q4, 1997.

[Rao06] R. R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Analytical Yield Prediction Considering Leakage/Performance Correlation," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 25, no. 9, pp 1685-1695, Sept 2006.

[Rod02] E. Rodney, Y. Tellier, and S. Borri, "A silicon-Based Yield Gain Evaluation Methodology for Embedded SRAMs with Different Redundancy Scenarios," Proc. IEEE MTDT Workshop, pp. 57-61, July 2002.

[Roy03] K. Roy, S. Mukhopadhyay, and H. Mahmoodi, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," Proceedings of the IEEE, Vol. 91, No. 2, pp. 305-327, February 2003.

[San03] R. S. Santiesteban et al., "Effect of Tilt Angle Variations in a Halo Implant on V_{th} Values for a 0.14µm CMOS Devices," IEEE Transactions on Semiconductor Manufacturing, Vol. 16, No. 4, pp. 653-655, Nov. 2003.

[Sap04] S. S. Sapatnekar, "Timing," Springer edition, 2004.

[Sch01] V. Schober, S. Paul and O. Picot, "Memory Built in Self Repair using redundant words," Proceedings International Test Conference, pp. 995-2001, 2001.

[Sch03] D. K. Schroder, and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," Journal of Applied Physics, Vol. 94, No. 1, pp. 1-18, July 2003.

[See87] E. Seevinck, "Static Noise Margin Analysis of MOS SRAM Cells," IEEE Journal of Solid State Circuits, Vol. 22, No. 5, pp. 748-754, Oct 1987.

[Seg04] J. Segura, and C. F. Hawkins, "How it works, how it fails," Wiley-IEEE Press, April 2004.

[Shy84] 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, Dec 1984.

[Sin99] K. Singhal, and V. Visvanathan, "Statistical Device Models from Worst Case Files and Electrical Test Data," IEEE Transactions on Semiconductor Manufacturing, Vol. 12, No. 4, pp. 470-484, Nov 1999.

[Sin04] R. Singh, and N. Bhat "An Offset Compensation Technique for Latch Type Sense Amplifiers in High-speed Low-power SRAMs," IEEE Transactions on VLSI Systems, Vol. 12, No. 6, pp. 652-657, June 2004.

[Soc07] Extreme DA 2007, "Statistical Timing Analysis: Sign-off for a New Generation," http://www.soccentral.com/results.asp?CatID=488&EntryID=23270

[Stat07] EETIMES 2007, "Statistical Static Timing Analysis: A View from the future," http://www.eetimes.com/showArticle.jhtml?articleID=201201507

[Sto98] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling Statistical Dopant Fluctuations in MOS transistors," IEEE Transactions on Electron devices, Vol. 45, No. 9, pp. 1960-1971, Sep 1998.

[Ter01] C. Terswiesch, R. E. Bohn, "Learning and Process Improvement during Production Ramp-up," International Journal of Production Economics, Vol. 70, No. 1, pp. 1-19, 1998.

[Tsa93] R. S. Tsay, "An Exact Zero Skew Clock Routing Algorithm," IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, Vol. 12, No. 2, pp. 242-249, Feb 1993.

[Tsi99] Y. Tsividis, "Operation and modelling of the MOS transistors," Second edition, Oxford University Press, 1999.

[Wan06] H. Wang, M. Miranda, F. Catthoor, and W. Dehaene, "On the Combined Impact of Soft and Medium Gate Oxide Breakdown and Process Variability on the Parametric Figures of SRAM components," 14th IEEE International Workshop on Memory Technology, Design, and Testing, pp. 71-76, Aug 2006.

[Won93] H. S. Wong and Y. Taur, "Three dimensional atomistic simulation of discrete random dopant distribution effects in sub 0.1 mm MOSFET's," in International Electron Device Meeting Tech. Dig., pp. 705-708, 1993.

[Yam05] M. Yamaoka et al., "Low-Power Embedded SRAM Modules with Expanded Margins for Writing," IEEE International Solid-State Circuits Conference Digest of Technical Papers, Vol. 1, pp. 480-481, Feb 2005.

[Zha95] J. C. Zhang and M. A. Styblinski, "Yield and Variability Optimization of Integrated Circuits," Kluwer Academic Publishers, 1995.

[Zha06] K. Zhang et al., "A 3-GHZ 70-Mb SRAM in 65nm CMOS Technology with integrated Column-Based Dynamic Power Supply," IEEE Journal of Solid State Circuits, Vol. 41, No. 1, pp. 146-151, January 2006.

[Zho96] Q. Zhou, and K. Monharam, "Elmore model for energy estimation in RC trees," Proceedings of the 43rd annual Conference on DAC, pp. 965-970, July 2006.

[Zor02] Y. Zorian, "Embedded Memory Test & Repair: Infrastructure IP for SOC Yield," International Test Conference 2002 (ITC'02), pp. 340-349.

Publications

Publications

M. Yap San Min, P. Maurine, M. Bastian, M. Robert, "Variabilité de process et de performances des mémoires SRAM embarquées ", 6ème Colloque FTFC (Faible Tension Faible Consommation) de l'ISEP, Paris, 2007.

M. Yap San Min, P. Maurine, M. Bastian, M. Robert, "Process variabilities and performances in a 90nm embedded SRAM" IEEE International Integrated Reliability Workshop (IIRW), 15-18 Oct 2007, Fallen Leaf Lake, Californie (To be published).

M. Yap San Min, P. Maurine, M. Bastian, Michel Robert, "Process variability considerations in the design of an eSRAM**",** Proc. IEEE MTDT Workshop, pp. 23-26, Dec 2007, Taipei, Taiwan.

M. Yap San Min, P. Maurine, M. Bastian, M. Robert, "A novel dummy bitline driver for read margin improvement in the design of an eSRAM"**,** 4th IEEE International Symposium on Electronic Design, Test and Applications (DELTA), 23-25 Jan 2008, Hong Kong (To be published)**.**

RESUME

Parallèlement à l'accroissement de la part dévolue à la mémoire au sein des circuits, l'évolution technologique s'accompagne d'une augmentation de la variabilité des performances, notamment dues aux variations de procédés de fabrication, de la tension d'alimentation et de la température. En terme de conception, ces variations de procédés de fabrication et de conditions de fonctionnement sont généralement prises en compte en adoptant une approche pire et meilleur cas (méthode des corners). Cependant, l'accroissement de la variabilité des procédés de fabrication conduit à l'accroissement relatif de la fourchette d'estimation des performances, comme la marge temporelle de lecture dans une mémoire embarquée de type SRAM. Ainsi la seule alternative, permettant de s'affranchir de la méthode des corners, réside dans l'adoption de techniques statistiques et notamment l'analyse statistique des performances temporelles. Cette analyse statistique des performances des SRAMs constitue le cœur de cette thèse. Dans un premier temps, nous avons démontré les limites de la méthode des corners sur la marge temporelle de lecture de la SRAM. Puis, nous avons développé une approche de modélisation permettant de s'affranchir de la méthode des corners. Cette approche a ensuite été utilisée dans le dimensionnement statistique de la mémoire afin d'optimiser ses performances temporelles, réduisant ainsi la marge excessive de lecture introduite par l'approche traditionnelle.

TITRE

Statistical analysis of the impact of within die variations on eSRAM internal signal races

Abstract

Aggressive technology scaling has led to the progressive degradation of transistor performances due to variability conditions such as process, voltage and temperature fluctuations. To handle the impact of manufacturing process variations along with the operating conditions in circuit design, corner based methodology is performed by characterizing the circuit under best case and worst case conditions (corner analysis method). However, the increase of variability in manufacturing processes results in an overestimation of performances, for instance like the read timing margin of an embedded SRAM. Hence, the only alternative to overcome the hurdle linked to corner analysis method is in the use of statistical design techniques, more specifically through statistical timing analysis. The statistical timing analysis of SRAM performances constitutes the essence of this thesis. First of all, we have shown the serious limitations associated with the corner analysis method as far as the read timing of the SRAM is concerned. Based on these results, we have then proposed a modelling approach as an alternative to the corner based method. This approach has then been used in the statistical sizing of the memory so as to optimize its timing performances, thereby mitigating the excessive read timing margin introduced by the traditional method.

DISCIPLINE : Microélectronique

MOTS-CLES : Analyse statistique, Course de signaux, Marge de lecture, SRAM, Variabilité

Université de Montpellier II : Sciences et Techniques du Languedoc
LIRMM : Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier
161 Rue Ada- 34392 Montpellier Cedex 5