Yield Estimation of SRAM and Design of a Dual Functionality Read-Write Driver for SRAM

Student Name: Pulkit Sharma

IIIT-D-MTech-ECE-VLSI-14-16
June 11, 2016

Indraprastha Institute of Information Technology
New Delhi

Advisor
Mohammad S. Hashmi

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Electronics and Communication,
with specialization in VLSI and Embedded Systems

Certificate

This is to certify that the thesis titled **"Yield Estimation of SRAM and Design of a Dual Functionality Read-Write Driver for SRAM"** submitted by **Pulkit Sharma** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *VLSI & Embedded Systems* is a record of the bonafide work carried out by him under my guidance and supervision in the Security and Privacy group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Mohammad S. Hashmi
Indraprastha Institute of Information Technology, New Delhi

Abstract

On chip process parameter variations have become a major challenge to meet the high density demands and the advancement in CMOS technologies. Variations in threshold voltage of transistors due to random dopant fluctuations is one of the most prominent challenges. The performance of memory sub-systems such as Static Random Access Memory (SRAM) is heavily dependent on these variations. Therefore, for sophisticated and high density electronic devices it is critical to keep a check on how tolerant they are towards these parametric fluctuations. In addition, to achieve higher yield (Y), it is desired that the SRAM bit cell qualifies in the order of less than 0.1ppb. Furthermore, the bitline differential voltage of the memory sub-system and the input-offset of the sense amplifier (SA) also deviates from their nominal values because of the parametric fluctuations. This thesis, therefore, proposes an efficient qualitative statistical analysis and Yield estimation method of SRAM sub-system which considers deviations due to variations in the process parameters of both the bit cell and the sense amplifier. The Yield of SRAM is predicted for different capacities of SRAM array by developing a model of memory sub-system in 65nm bulk CMOS technology.

The second part of this thesis tries to address some concerns related to area and power efficient designs of SRAMs and Non-Volatile Memories (NVMs). In this context, a novel, area and power proficient design of a Dual Functionality Read-Write (DFR-W) driver for SRAM sub-system has been proposed. The design is then integrated to a memory sub-system with an operating frequency of 1GHz in CMOS 65nm technology. It is compared with the conventional memory architecture keeping in perspective the power, area, leakage and speed of operation for varied memory capacities.

Acknowledgments

Towards the end of my M.Tech. degree, I would like to pay my gratitude to several individuals who contributed from various perspectives. First of all, I would like to express my profound gratitude to my thesis supervisor, Dr. Mohammad S. Hashmi for his support, direction and motivation all along the way. He has been a great source of inspiration for me. I would like to thank the director Dr. Pankaj Jalote and all my professors at IIITD for providing a positive learning atmosphere to work. I would like to express my gratitude towards my family for their kind co-operation and support which was a tremendous wellspring of motivation. Lastly, I need to thank my companions for all the mindful discourses we had, which incited me to think past the self-evident.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and objective

The scaling of technology nodes into nanometer regime is one of the momentous factors which has led to the rapid advancements in the VLSI industry. This actually substantially helps in meeting the high-density demands of the modern chip designs. The plot in figure 1.1 depicts the trend of technology nodes from the year 1991 onwards and it can be observed that this scaling trend will soon enable extremely high-density chips with tremendous functionality [1].

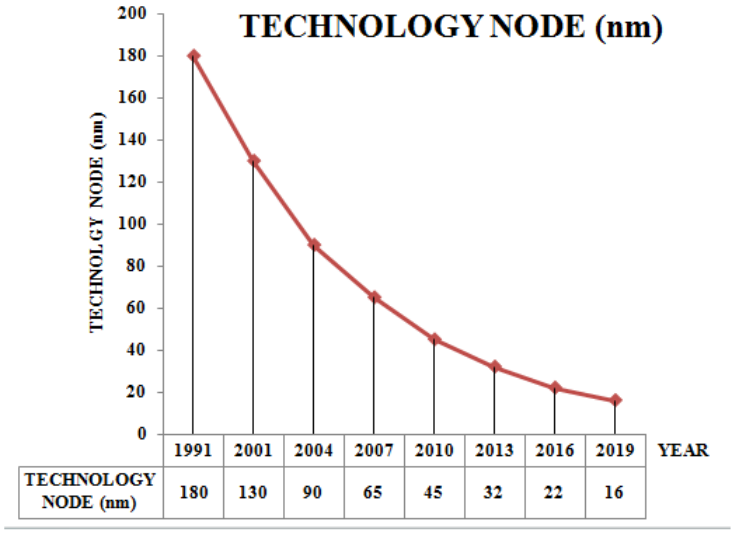

Figure 1.1: A trend showing technology node scaling

However, all of these advancements are accompanied with copious challenges both in design as well as fabrication domain. For example, process parameter variations are such unavoidable repercussion which results in the induction of on chip failure. Whereas the transistor parameters such as channel dimensions, threshold voltage, input offsets etc. can diverge noticeably from

their nominal values. Static Random Access Memory (SRAM) is one electronic design which is intolerant towards such process parameter variation-induced malfunctions. SRAM chip failure can occur even with the failure of one memory cell if redundancy is not considered [2] [3]. Consequently, it becomes evident that yield estimation of SRAM is exceedingly important.

The quality of SRAM cell is typically quantified with the help of figure of merits (FOMs) like cell current, read time, retention voltage, read static noise margin (RSNM), write margin (WM) etc. A cell not meeting the threshold of predefined values of these FOMs is considered as faulty [4]. Prior reports on yield estimations of SRAMs include various methodologies such as highly accurate Monte Carlo (MC) simulations [5] and non-MC methods like importance sampling [6] [7] which intend at eliminating the computation obstacles by compromising the accuracy of estimation. Also, the approach of Design of Experiments (DoE) presented in [8] [9] provide a non-MC yield estimation method. But in all these existing techniques, yield degradation of memory because of variations in process parameters of peripherals like sense amplifier (SA) and write drivers, variation in bitline differential voltage etc. have not been considered.

In this literature, therefore, a new and an extremely effective yield estimation technique of SRAM sub-system of different capacities considering the parametric variations in the memory cell as well as the deviations in process parameters of SA and the bitline differential is presented. This approach helps in predicting the yield of the memory chip with very good accuracy through the modeling of the memory chip while considering all real time on-die failure sources. Contrasting to the technique depicted in [8] this work considers parametric variations in transistors of both the memory bit cells and SA simultaneously.

The present day electronic devices are becoming smarter day by day i.e. they are now capable of taking an intelligent action in response to the reaction in their surrounding. For example, smart phones have become a phenomenon with built in graphics and multimedia. This could not have been possible without high speed and high density embedded memories which store large volumes of information and help in the super fast processing of data. The scaling of technology nodes has led to rapid advancements in the design of high density memory systems like SRAM. These advancements have led to the development of high speed and power efficient memories which have found applications in mobile, cache and Systems on Chip (SoC) [10]. In order to achieve high densities in memories, the design and development of the memory array and its peripherals must be area efficient. Conventionally for faster memory access, SRAMs use sense amplifiers (SA) as reading devices and write drivers as writing circuitry. Over the years, various efforts have been made to design the most sophisticated and high performance SA and write drivers to enhance the performance of the memories. There are many SA and write driver structures available which excel in different figures of merits. For instance, the write driver structure in [11] achieves higher write margin for sub-threshold applications. The main function of a write driver is to synchronously discharge one of the bitlines as per the data input. There are various other write driver circuits mentioned in [12] which are integrated with memory for varied applications

in the fields of communication, computing, artificial intelligence etc.

Similarly, there has been active research in developing different kinds of reading circuitry. Reading circuitry has evolved from simple differential amplifiers to sophisticated and faster sense amplifiers. SA have been an integral block amongst the various other peripherals. They cater to the needs of high speed memory access. They mainly function by amplifying small bitline differential and hence has a direct impact on the read failure of the memory [13].

In this work, therefore, a new read-write circuitry is presented which exhibits dual functionality i.e. a single circuit which can perform both reading and writing operation on a SRAM bit cell. This design eliminates the need for an explicit write driver. Hence, there is a huge decline in the memory chip area because conventionally each column of a memory array requires a SA and a write driver. However, this new design prevents this necessity thereby saving a lot of chip area. This design also helps in reducing the load of bitlines which in turn improves the response time of SRAM.

1.2 Contribution

We have designed the model of complete SRAM sub-system set up and have proposed a novel yield estimation technique of SRAM sub-system for different capacities subjected to parametric variations.

- Pulkit Sharma, Anil K. Gundu, M. S. Hashmi, Modeling and Yield Estimation of SRAM Sub-System for Different Capacities Subjected to Parametric Variations, accepted in 20^{th} *International Symposium on VLSI Design and Test (VDAT-2016).*

In this work, we have proposed a new read-write circuitry for SRAM which exhibits dual functionality i.e. a single circuit which can perform both reading and writing operation on a SRAM bit cell. Results suggest that the new design provides faster memory access and is area and power efficient.

- Pulkit Sharma, M.S. Hashmi, A Novel Design of a Dual Functionality Read-Write Driver for SRAM, submitted to 29^{th} *IEEE System on Chip Conference (SOCC-2016).*

1.3 Terminologies

This section aims at presenting some important terminologies along with their definition.

1.3.1 Figures of Merit (FoM)

Figures of merit depict the parameters which can be used to quantify the performance of a circuit. Some of the FoMs for SRAM are listed below.

- **Hold Static Noise Margin** The Hold SNM is the maximum amount of noise voltage that can be introduced at the output of the two cross coupled inverters of a bit cell, such that the cell retains its data.
- **Read Static Noise Margin** A metric to evaluate the read stability of a SRAM cell is Read Static Noise Margin (RSNM). It is defined as the length of the side of the largest square that can fit into the lobes of the butterfly curve. Butterfly curve is obtained by drawing and mirroring the inverter characteristics while the word line is set high and both bitlines are still pre-charged. Also, it is the maximum amount of noise voltage that can be introduced at the output of the two inverters, such that the cell bit retain its data during the read operation.
- **Write Static Noise Margin** The write noise margin is the measure of the ability to write data into the bit cell. It is defined as the minimum word line voltage or the maximum bit-line voltage needed to perform the write operation in the bit cell.
- **Data Retention Voltage** It is the minimum power supply voltage needed to retain the data in the standby mode. It should be noted that lowering the supply voltage reduces power and leakage.
- **Cell current** The read current flowing through the bit-line, access transistor and pull down transistor is called cell current. Higher cell current denotes faster cell access.

1.3.2 Statistical Yield Analysis

Statistical analysis aims at the calculation of Yield and failure probabilities of systems like SRAMs by the use of mathematical models. There are various kinds of techniques available to predict the yield of SRAMs.

- Monte Carlo Simulations
- Design of Experiments
- Importance Sampling

1.4 Outline

The remaining thesis is organized as below:

Chapter 2 presents the complete design of the SRAM sub-system architecture. The functioning of all the blocks of the architecture is explained in detail. Then the simulation of the complete architecture depicting the timing diagram of the design is depicted.

Chapter 3 deals with proposed yield estimation technique of SRAM sub-system for different capacities. This chapter presents the mathematical model of yield estimation and depicts the parametric variations in the threshold voltage which are the main sources of read failure in the memory. The probability of failure is then estimated for different capacities of the memory array.

Chapter 4 presents a novel and an efficient dual functionality read-write driver (DFR-W driver). The DFR-W driver is integrated with the memory architecture and then compared with the conventional architecture having a sense amplifier as the reading circuitry and a write driver for writing. The comparison is made on the grounds of area, speed and power.

Chapter 5 This chapter concludes the thesis and suggests the possible directions for future work.

Chapter 2

The design of SRAM Sub-System Architecture

2.1 Introduction

In this thesis, SRAM Sub-System is the depiction of one complete column of the memory array along with its corresponding peripheral circuitry. The design of this sub-system is the primary step of this work. Figure 2.1 shows a conventional SRAM sub-system comprising the following components:-

- Pre-Charge block.
- Memory bank comprising of n number of bit cells.
- Circuitry for accessing the memory.
  - A latch type sense amplifier (SA) for reading from the memory.
  - A write driver for writing into the memory.

In this chapter, all the above mentioned components of the sub-system are explained in detail and the design of each circuit is presented. After the integration of all these components into the memory sub-system, a complete simulation of the complete design at an operating frequency of 1GHz in CMOS 65nm technology is also shown. All the simulations are carried out in Mentor graphics' ELDO SPICE simulator.

2.2 Pre-Charge Block

Before reading from any cell in the memory bank, the two bitlines are pre-charged to supply voltage using the pre-charge circuitry shown in figure 2.2. It comprises of two PMOS devices

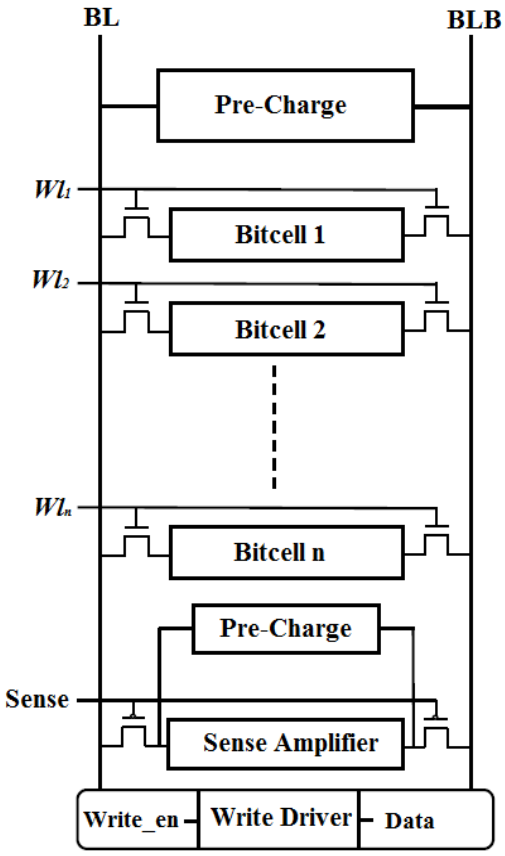

Figure 2.1: Architecture of a conventional SRAM sub-system

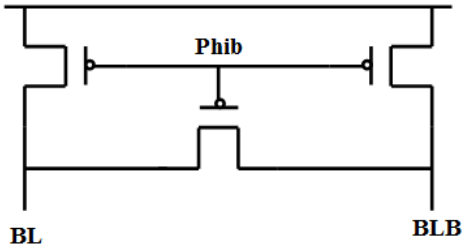

Figure 2.2: Pre-charge Circuitry

connecting the supply to the bitlines. An equalizing PMOS device is used to ensure that both the bitlines are at the same voltage potential. This also helps in reducing the pre-charge time.

2.3 A conventional six Transistor (6T) SRAM bit cell

SRAMs are static storage elements which store a single bit till the power supply is on. As soon as the power supply is turned off all the data stored in the static storage is lost. A 6T SRAM bit cell is used as the fundamental memory unit of the SRAM sub-system. Figure 2.3 shows a conventional 6T SRAM bit cell which acts as the fundamental storage elements. Amongst the six transistors of the bit cell, PDL (Pull Down Left) and PDR (Pull Down Right) are the NMOS devices and PUL (Pull Up Left) and PUR (Pull Up Right) are the PMOS devices. These four transistors form two cross-coupled CMOS inverters which possess a strong positive feedback. The principle of a strong positive feedback enable these cross-coupled devices to store a single binary digit ('1' or '0') at their internal nodes (q or qb). These internal nodes are connected to the peripheral circuitry through two access transistors (AXL and AXR) and the two bitlines (BL and BLB) [2]. There are three modes of operation that are associated with a bit cell.

- Read Mode.
- Hold Mode.
- Write Mode.

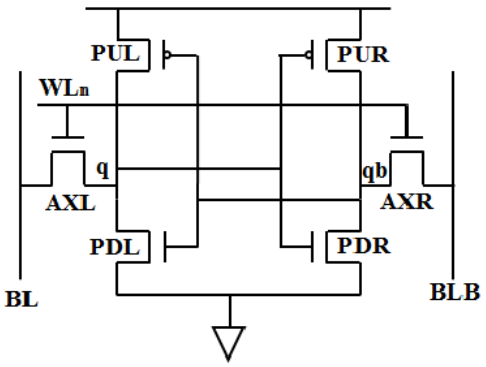

Figure 2.3: 6T SRAM bit cell

2.3.1 Read Mode

In the read mode of operation, the two bitlines are kept pre-charged to the supply voltage. For reading data from the n^{th} bit cell, WL_n (word line corresponding to the n^{th} bit cell) is asserted and the access transistors connect the internal nodes of the bit cell to the bitlines. Assuming

node q stores logic ‘0’ and qb stores logic ‘1’, transistor AXL and the pull down transistor PDL form a discharging potential divider network for BL. As a result, a potential difference develops between BL and BLB which is sensed by the SA for reading.

During the read operation, a surge of electric current flows into the bit cell as BL discharges through AXL and PDL. As a result, a potential difference called the bump voltage develops at the drain of PDL. If this bump voltage exceeds the switching potential of the inverter formed by PUR and PDR, the bits stored in the cell might flip and result in a read failure. To prevent such a situation and minimize the bump voltage, pull down transistor must be sized stronger than the access transistor [4].

2.3.2 Hold Mode

In the hold mode of operation, the bit cells remain detached from the bitlines (WL is not selected) and simply retain the data written into them, while the power supply is on. If the word line is not asserted, the access transistors AXL (access transistor left) and AXR (access transistor right) disconnect the cell from the bit-lines. The two cross coupled inverters formed by PUL (pull-up left)-PDL (pull-down left), PUR (pull-up left)-PDR (pull-down right) will continue to reinforce each other as long as they are connected to the supply. Usually, during hold mode, the power supply voltage is lowered to round about the retention voltage of the cell in order to reduce leakage and power dissipation.

2.3.3 Write Mode

In the write mode of operation, new data is written into the cell. For writing logic ’1’ in the bit cell, when WL_n is asserted, BL remains pre-charged and BLB discharges through the write driver. Consequently, node qb discharges to logic ’0’ through AXR. The positive feedback between the cross coupled inverters force node q to logic ’1’ and writing into the cell is accomplished. The discharging of qb is opposed by PUR therefore, pull up transistors must be sized weaker than the access transistors [4].

2.4 Cicuitry for Memory access

This section aims at depicting various peripheral circuits which are used to access memory for reading and writing purposes. These peripheral devices aim at faster and error free memory access.

2.4.1 Latch type SA

Figure 2.4 shows a conventional latch type SA which is used to read data from the memory bank by simply amplifying a small bitline potential difference.

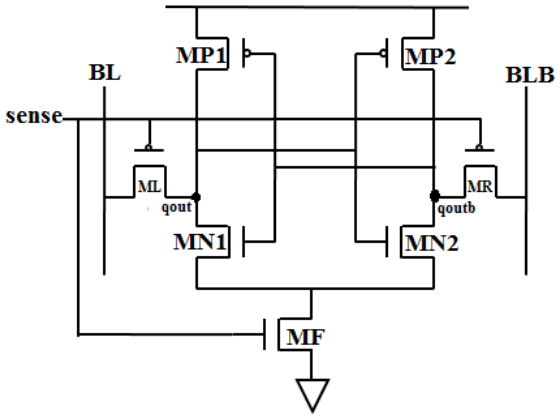

Figure 2.4: A conventional latch type SA

Before reading, the pre-charge circuitry keeps the internal nodes of the SA (qout and qoutb) in a metastable state. During the read mode, one of the bitlines start discharging through the pull down transistor in the bit cell. When the potential difference between the two bitlines becomes greater than the input offset of the SA, the sense signal is made high. The pre-charge circuitry is switched off and the internal nodes of the SA start discharging through MN1, MN2 and the footer MF. The discharging rate of the node corresponding to a higher bitline potential is lower as compared to the discharging rate of the other node. After discharging for some time the latching action begins in the SA and the internal nodes start settling at the appropriate logic levels. Hence, the reading of the data from the bit cell is accomplished [14].

2.4.2 Write driver

Write driver is a device which is mainly used to drain out charge from one of the pre-charged bitlines, to accomplish a successful write operation in a SRAM bit cell. There are various write driver structures that have been designed for various applications. This section presents two such write driver structures.

Figure 2.5 depicts the most basic write driver circuit [4]. It comprises of two stacked NMOS devices connected on either side of the two bitlines. The write enable signal (write_en) is set high in the write mode of operation. Assuming data pin (Data) to be at logic '0', the stacked NMOS devices, connected to BLB are turned on in this state of logic. Consequently, BLB discharges through them and BL stays at logic '1' and as a result logic '1' is written into the selected bit cell. Similarly, logic '0' can be written into the selected cell by making Data pin high

while write_en is selected.

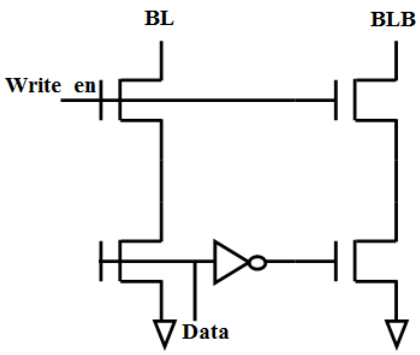

Figure 2.5: Write Driver circuit I

Fig. 2.6 shows another write driver structure [15]. It has two main components the control circuit comprising of two NOR gates and an inverter and the driver. The main function of a write driver is to pull down one of the bitlines to accomplish a successful write operation.

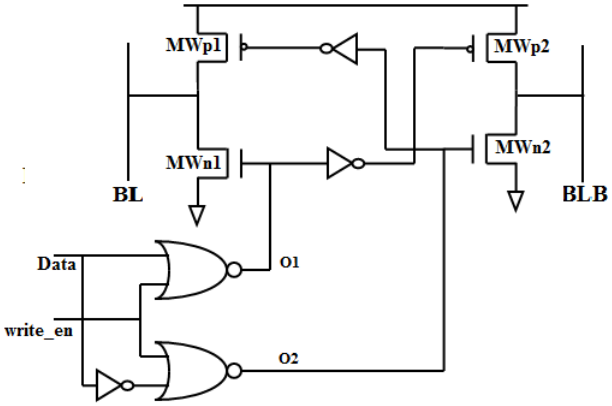

Figure 2.6: Write Driver circuit II

Table 2.1 depicts the truth table of the control circuitry. When the write_en signal is at logic '0' and Data is at logic '1', node O1 will be '0'. As a result, MWn1 and MWp2 will be off. Similarly, node O2 will be '1' and consequently, MWn2 and MWp1 will be on. In this arrangement of logic, BLB will discharge through MWn2 and BL will remain pre-charged. Hence, Data is written in the bit cell when the WL is selected.

Table 2.1: Logic Conditions of the control circuitry of Write Driver

Write_en	Data	O1	O2
0	0	1	0
0	1	0	1
1	0	0	0
1	1	0	0

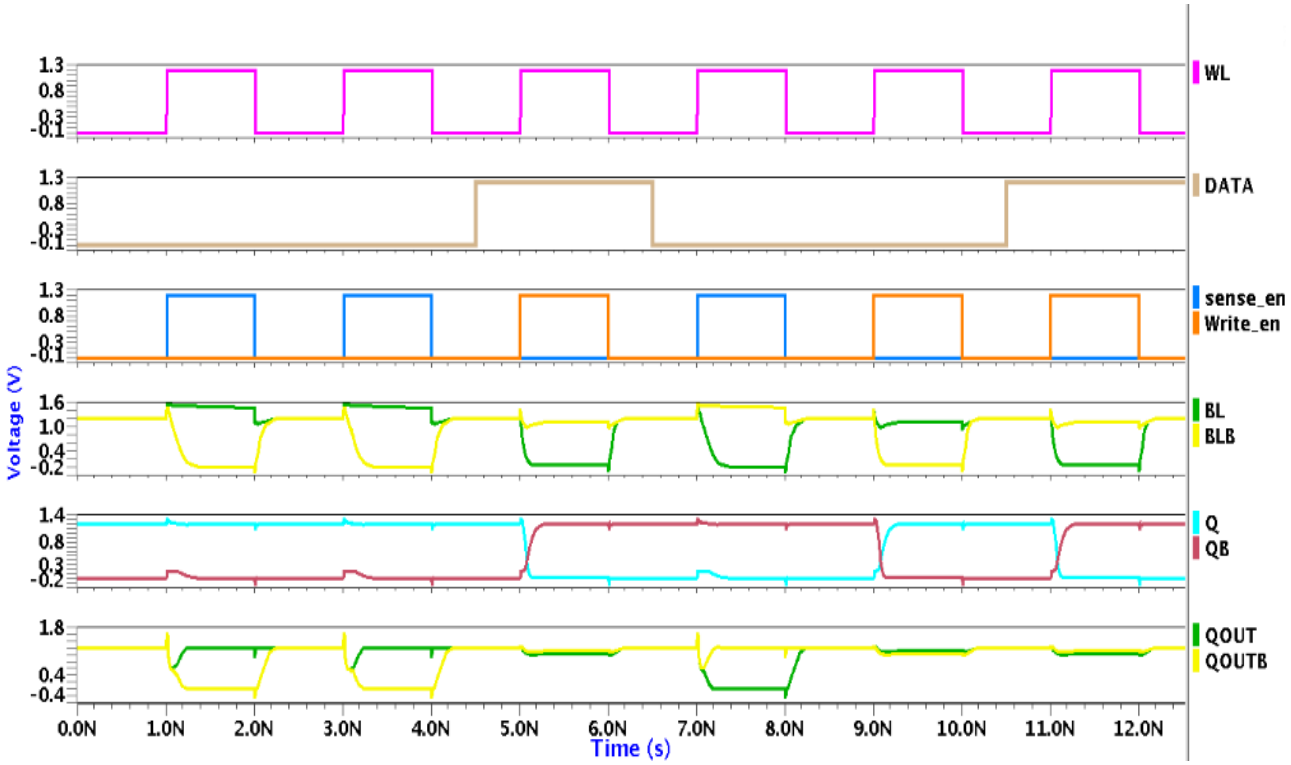

Figure 2.7: Simulation depicting the timing diagram of memory sub system shown in Fig. 2

2.5 Simulation results of the designed SRAM Sub-System Architecture

Figure 2.7 shows the simulation of the designed SRAM sub-system architecture shown in fig. 2.1 with n=128 bit cells. It is depicting the timing diagram of consecutive read and write cycles each having a frequency of 1GHz. Here, nodes Q and QB represent the internal nodes of the first bit cell whereas nodes QOUT and QOUTB depict the internal nodes of SA. This transient analysis depicts six consecutive read (R) and write (W) cycles following RRWRWW order. This simulation was carried out on ELDO SPICE simulator in 65nm CMOS technology.

Table 2.2 depicts the width of all transistors of the bit cell. It can be clearly observed that the pull down devices are sized higher than the access transistors, this ensures high read stability. Also, pull down devices are sized weaker than the access transistors, this ensures high writeablity.

Table 2.2: Sizes of bit cell devices

Device	Width (nm)
PUL/PUR	100
PDL/PDR	230
AXL/AXR	160

Table 2.3: Sizes of SA devices

Device	Width (nm)
MN1/MN2	2000
MP1/MP2	4000
ML/MR	1000
MF	2000

Table 2.3 depicts the width of all transistors of latch type SA. Since each column of the memory array has a single SA, it can be observed that the devices of SA are sized around 10X higher than the devices of bit cell.

The NMOS devices of the stacked NMOS write driver shown in figure 2.5 is sized with a width of 800nm. Table 2.4 depicts the sizes of devices of write driver structure II shown in figure 2.6.

Table 2.4: Sizes of write driver II (figure 2.6) devices

Device	Width (nm)
Mwp1/Mwp2	2000
Mwn1/Mwn2	1000

Chapter 3

Proposed Yield Estimation Method

3.1 Introduction

Process variations have become a major challenge with the advancement in CMOS technologies. The performance of memory sub-systems such as Static Random Access Memory (SRAMs) is heavily dependent on these variations. Also, the VLSI industry requires the SRAM bit cell to qualify in the order of less than 0.1ppb to achieve higher Yield (Y). This chapter presents a novel and an efficient qualitative statistical analysis and Yield estimation method of SRAM sub-system which considers deviations due to variations in process parameters in bitline differential and input offset of sense amplifier (SA) altogether. The Yield of SRAM is predicted for different capacities of SRAM array by developing a statistical model of memory sub-system in 65nm bulk CMOS technology. In this section, memory sub-system failure has been discussed and MC simulation based statistical technique for yield estimation of the entire memory array is formulated mathematically.

A read failure can happen due to one of the following reasons:

- Bit cell failure: The data stored in the bit cells flip during the read operation due to bump voltages greater than the threshold voltage of the corresponding inverter.
- Sense amplifier failure: The SA fails to sense the correct input due to variations in its input offset voltage.
- Both the above mentioned points i.e., the bit cell and SA simultaneously fail during the read operation and wrong data is sampled.

It can be deduced that due to the parametric variations such as the deviation in threshold voltage (V_t) of the transistors, the input offset of the SA and the bitline differential varies randomly. If the input offset voltage becomes greater than the bitline difference during the read operation,

the amplifier will give erroneous output and, therefore, read failure can occur. Hence, predicting the failure probability and yield of the memory chip becomes a necessity.

3.2 Mathematical Model of the proposed Yield Estimation Technique

For the memory sub-system shown in figure 2.1, let $V_1, V_2, ..., V_n$ be the random variables (RV) associated with the change in bitline difference voltage due to process parameter variations in $Bitcell_1, Bitcell_2, ..., Bitcell_n$ respectively.

Now let OV be the random variable associated with the change in input offset voltage of the SA. Then the probability (p) of correct read can be written as:

$$p = P(correct\,read) = P(min(V_1, V_2, ..., V_n) > OV) \tag{3.1}$$

$$p = P(M > OV) \tag{3.2}$$

Where, $M = min(V_1, V_2, ..., V_n)$

Let us assume that $V_1, V_2, ..., V_n$ are independent random variable with cumulative distribution function (CDF) $F_{V_1}(v_1), F_{V_2}(v_2), ..., F_{V_n}(v_n)$ respectively.
If $F_M(m)$ is the CDF of M then we can express it as follows:

$$F_M(m) = 1 - P(V_1 > m)P(V_2 > m)...P(V_n > m) \tag{3.3}$$

$$= 1 - [(1 - F_{V_1}(m))(1 - F_{V_2}(m))...(1 - F_{V_n}(m))] \tag{3.4}$$

Assuming $V_1, V_2, ..., V_n$ and OV as Gaussian random distributions, one can express (3.4) as:

$$(1 - F_{V_n}(m)) = 1 - \frac{1}{2}(1 + erf(\frac{m - \mu}{\sigma\sqrt{(2)}})) \tag{3.5}$$

$$= \frac{1}{2}(1 - erf(\frac{m - \mu}{\sigma\sqrt{(2)}})) \tag{3.6}$$

$$= \frac{1}{2}(erfc(\frac{m - \mu}{\sigma\sqrt{(2)}})) \tag{3.7}$$

Where μ is the mean and σ is the standard deviation of the distribution.

The complementary error function is given by:

$$erfc(x) = 1 - erf(x) \tag{3.8}$$

$$= 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{3.9}$$

$$= \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \tag{3.10}$$

Now assuming, $F_{V_1}(m), F_{V_2}(m), .., F_{V_n}(m)$ are Independent and Identically Distributed (IID) distributions. Then:

$$F_{V_1}(m) = F_{V_2}(m) = ... = F_{V_n}(m) = F_V(m) \tag{3.11}$$

Substituting (3.11) in (3.4). This results in:

$$F_M(m) = 1 - (1 - F_V(m))^n \tag{3.12}$$

$$f_M(m) = \frac{d}{dm}(F_M(m)) \tag{3.13}$$

Where $f_M(m)$ is the probability distribution function (PDF) of M. Now, one can solve (3.12) and (3.13) to get:

$$f_M(m) = n(1 - F_V(m))^{n-1} f_V(m) \tag{3.14}$$

Where $f_V(m)$ is the PDF of V [16].

Now the probability of correct read can be obtained by substituting (3.14) in (3.1).

$$p = P(M > OV) = \iint_R f_{M,OV}(m, ov)\, dm\, dov \tag{3.15}$$

Where region $R = \{-\infty < m < \infty, -\infty < ov < m\}$

Therefore, Yield(Y) of k such memory slices will be given by

$$Yield(Y) = p^k \tag{3.16}$$

3.3 Calculation of failure probability and statistical Yield analysis of SRAM array

3.3.1 Probability Density Function of V and OV

This section aims at formulating the PDFs of random variables corresponding to deviations in bitline differential (V) and input offset voltage of SA (OV). For this purpose, 2500 Monte Carlo simulations were carried out to induce variations in all the bit cells of the memory setup shown in figure 2.1. For these 2500 samples, the mean and standard deviation (STDEV) of V_t variations of bit cell devices have been formulated in table 3.1.

Figure 3.1 depicts the flipping of stored bits during read when bit cell is subjected to threshold voltage variations shown in figure 3.2.

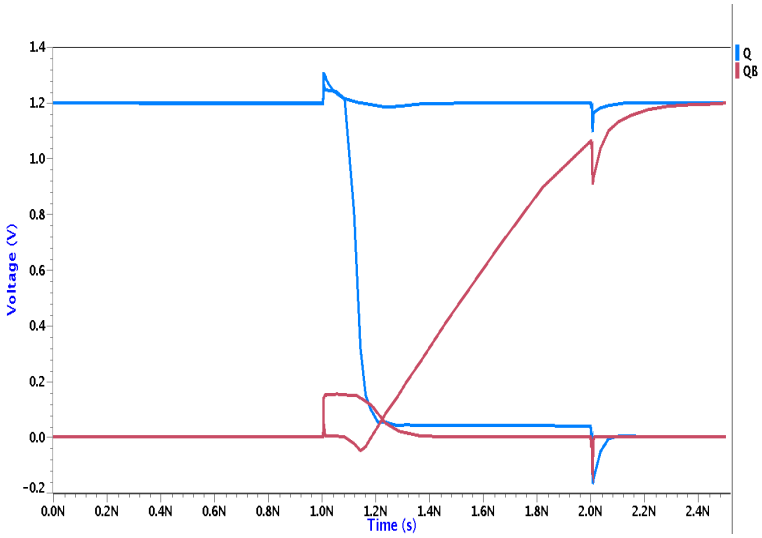

Figure 3.1: Read failure in bit cell subjected to V_t variations

Table 3.2 quantifies the mean and standard deviation of the 2500 samples of V_t variations, shown in figure 3.3, in the transistors of the latch type SA.

Table 3.1: V_t variations in bit cell transistors

DEVICE	MEAN	STDEV
ACCESS, PULL DOWN	478.3mV	26mV
PULL UP	-444mV	26mV

Figure 3.4 depicts the sensing failure during read, as sensing device suffers from on-chip parametric variations. The V_t variations in bit cell produce deviations in bitline differential voltage (V). Similarly, the V_t variations in the sensing device alter its input offset voltage (OV).

Now for correct sensing of data from bit cells, the bitline differential voltage should be greater

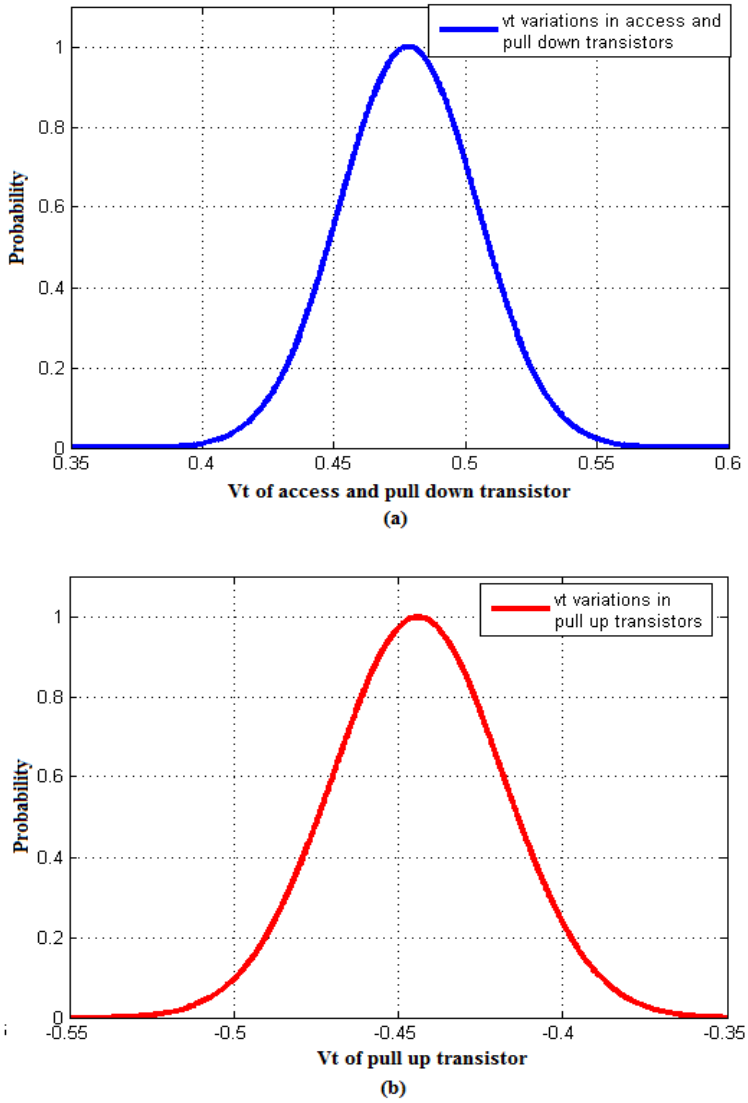

Figure 3.2: V_t variations in bit cell. (a) PDF of the threshold voltage of access and pull down transistors. (b) PDF of the threshold voltage of pull up transistors

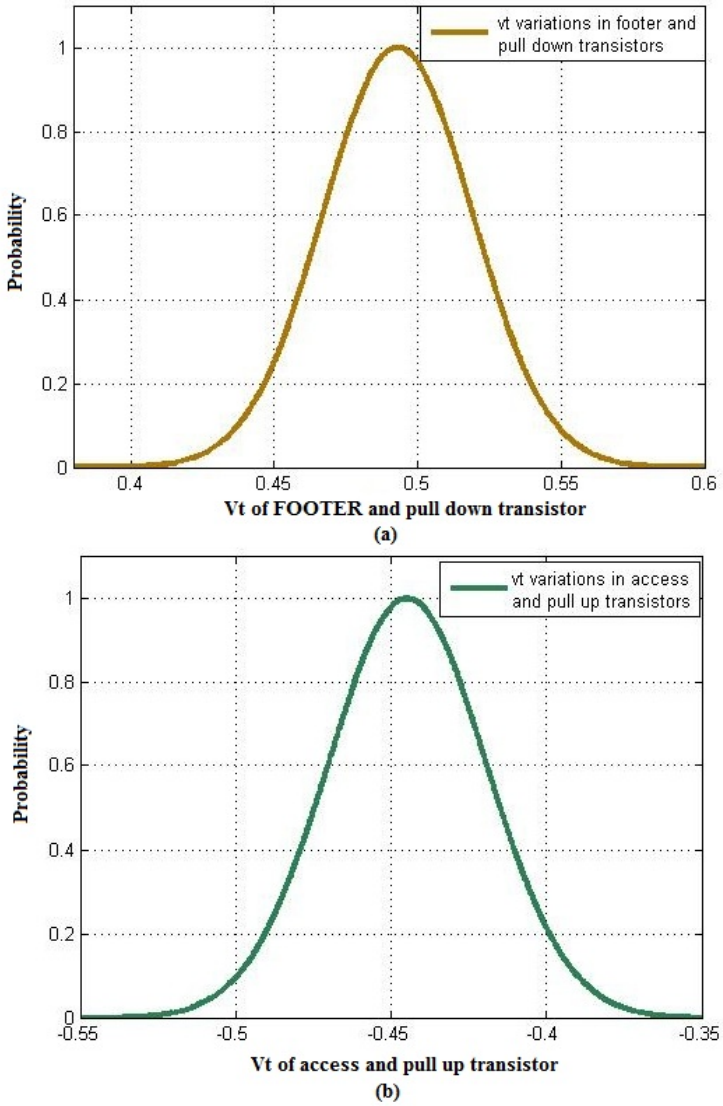

Figure 3.3: V_t variations in the sense amplifier. (a) PDF of the threshold voltage of footer and access transistors. (b) PDF of the threshold voltage of access and pull up transistors

Table 3.2: V_t variations in sense amplifier transistors

DEVICE	MEAN	STDEV
FOOTER, PULL DOWN	493mV	25.9mV
ACCESS, PULL UP	-444.8mV	25.6mV

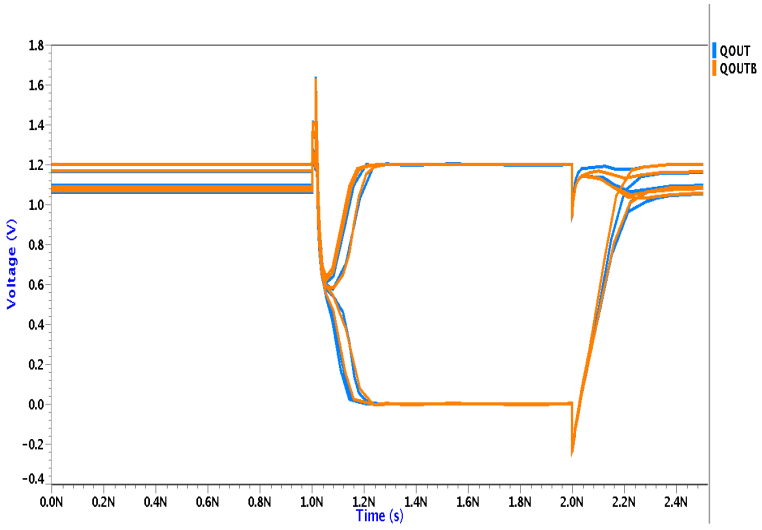

Figure 3.4: Sensing failure in sense amplifier subject to V_t variations

than or equal to the input offset voltage of the sense amplifier. In the absence of such a condition, read failure occurs as shown in figure 3.4.

The data in Table 3.3 quantifies the mean and standard deviation of variations of bitline differential voltages and input offset voltage of SA for a different number of bit cells per memory slice.

The PDFs of random variable (RV) V in (3.14) and OV in (3.15) for n=64 are shown in figure 3.5. Using the PDF of random variable V, the PDF of a minimum of n such random variables (M) is estimated using (3.14).

3.3.2 Statistical Yield analysis using the mathematical model of Yield estimation

From Tables 3.1, 3.2 and 3.3 it can be deduced that if process parameters vary by 6%, the variations in bitline differential is about 4.86% to 5.37% and about 2.5% to 3.1% in the input offset of SA. Yield is estimated for memory arrays modeled from different sizes of memory subsystems. Consequently changing the number of bit cells per SA in one column.

Once the PDF of random variable M is estimated using (3.14), Yield (Y) for any number of bit cells (n) in one slice of memory can be predicted using (3.15).

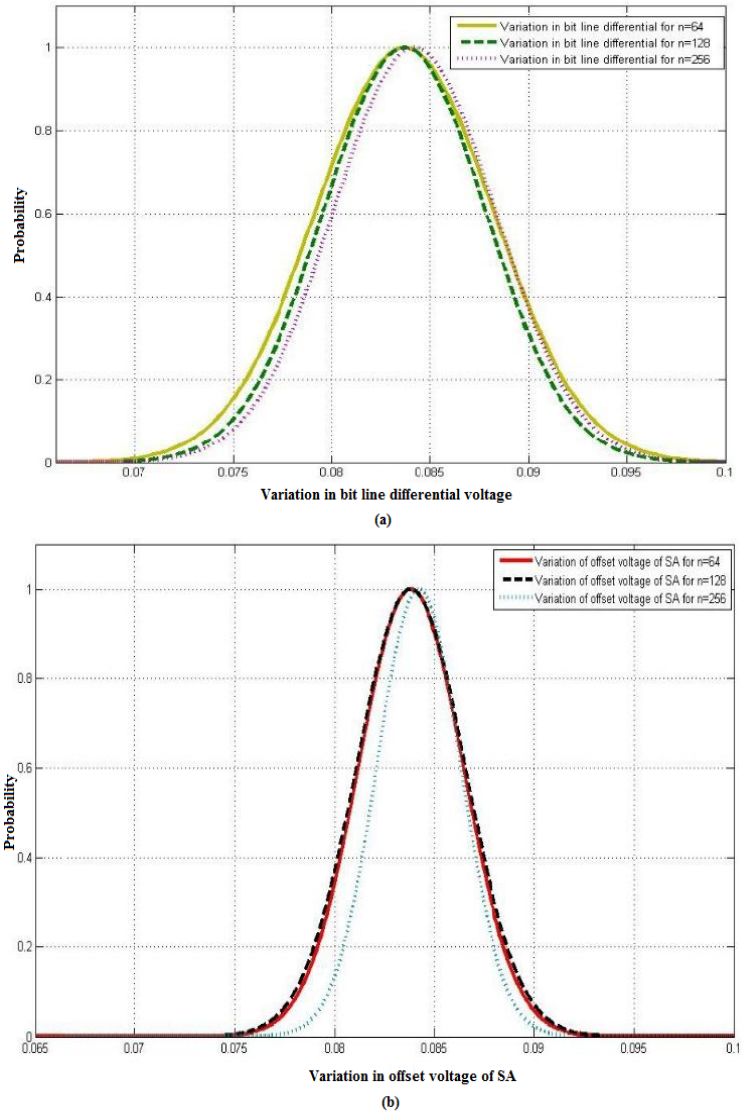

Figure 3.5: (a) PDF of deviation in bitline differential voltage (V). (b) PDF of deviation in input offset voltage (OV) of the sense amplifier

Table 3.3: Variation in the bitlines differential (V) and input offset (OV) of the SA

RV	n=64		n=128		n=256	
	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV
V	83.7mV	4.5mV	83.7mV	4.1mV	84.2mV	4.1mV
OV	83.8mV	2.6mV	83.8mV	2.7mV	84.2mV	2.1mV

For one slice of memory having n cells, if P_f is the probability of failure. Then:

$$probability\ of\ success(P_s) = 1 - P_f \tag{3.17}$$

For k such memory slices, total probability of failure (P_e) is

$$P_e = 1 - (1 - P_f)^k \tag{3.18}$$

Hence, the memory sub-system shown in figure 2.1, is simulated by inducing in the variations in V_t using MC in the read cycle of frequency 1GHz. Table 3.4 provides the failure probability P_f of this sub-system having a different number of bit cells (n=64, 128 and 256) and correspondingly, a single sense amplifier.

Using (3.18), the total probability of failure for different capacities of SRAMs has been calculated as shown in Table 3.5. As a result yield (Y) can be calculated using the following

$$Y = 1 - P_e \tag{3.19}$$

It is apparent that the probability of failure of SRAM chip increases with the increase in memory capacity. It is on expected line. Now this obtained probability of failure can be utilized to estimate the yield of SRAM array using (3.15).

The proposed statistical technique is used to estimate the Yield(Y) of the SRAM chip of different capacities and by varying the size of the memory sub-system. Table 3.5 and (3.19) suggests more than 99.9999% yield for the memory array modeled using the memory sub-system.

Table 3.4: Probability of failure of one memory slice

No. of cells / memory slice(n)	Probability of failure(P_f)
64	$4.802 * 10^{-13}$
128	$1.230 * 10^{-12}$
256	$4.874 * 10^{-12}$

Figure 3.6 depicts the trend showing that as the number of cells in one column per SA are increased, the probability of failure increases on account of increasing bitline capacitance. Consequently, as the number of cells per SA are doubled the failure probability increases by 70%.

3.4 Results and Conclusion

A novel, simple and an extremely accurate method for statistical analysis and Yield (Y) estimation of SRAM sub-system is presented in this chapter. It considers not only the variations in the bit cell but also takes into account the on-chip variations in sense amplifier due to which

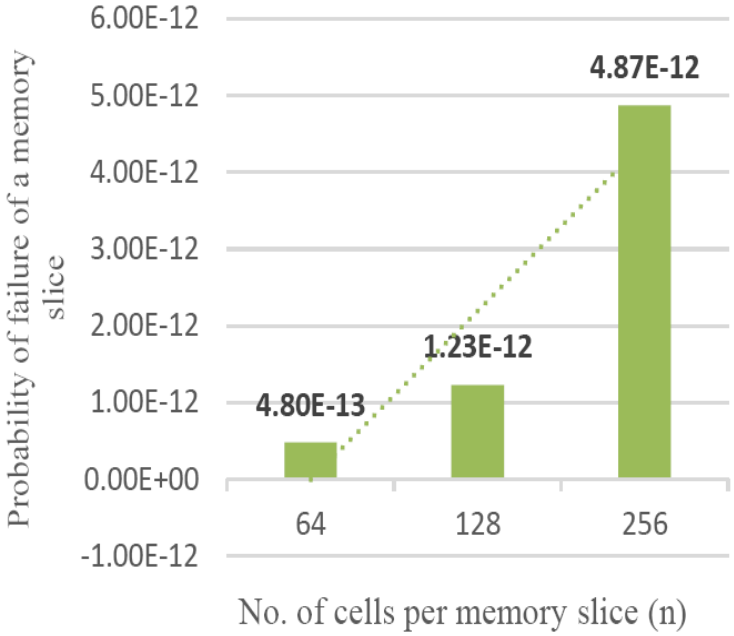

Figure 3.6: Trend showing the variation in P_f with change in number of bit cells (n) per memory slice

Table 3.5: Total Probability of failure (P_e) for different SRAM capacities

Capacity	P_e (Probability of failure)		
	n=64	n=128	n=256
8MB	$5.035 * 10^{-7}$	$6.450 * 10^{-7}$	$1.277 * 10^{-6}$
16MB	$1.007 * 10^{-6}$	$1.290 * 10^{-6}$	$2.555 * 10^{-6}$
32MB	$2.014 * 10^{-6}$	$2.580 * 10^{-6}$	$5.111 * 10^{-6}$
64MB	$4.028 * 10^{-6}$	$5.160 * 10^{-6}$	$1.022 * 10^{-5}$
128MB	$8.057 * 10^{-6}$	$1.032 * 10^{-5}$	$2.044 * 10^{-5}$
256MB	$1.611 * 10^{-5}$	$2.064 * 10^{-5}$	$4.088 * 10^{-5}$
512MB	$3.223 * 10^{-5}$	$4.128 * 10^{-5}$	$8.176 * 10^{-5}$
1GB	$6.446 * 10^{-5}$	$8.256 * 10^{-5}$	$1.635 * 10^{-4}$
2GB	$1.289 * 10^{-4}$	$1.651 * 10^{-4}$	$3.270 * 10^{-4}$

deviations in bitline differential voltage and input offset voltage of SA occur and consequently result in read failure.

The obtained results convey that for 6% variations in threshold voltages, the variation in bitline differential voltages and input offset voltages of SA are about 5.37% and 3.1% respectively.
For 64 bit cells and a single SA SRAM sub-system used at 1GHz read cycle for modeling the SRAM array, the probability of failure comes out to be $4.802 * 10^{-13}$ while as expected increases from $5.035 * 10^{-7}$ for 8MB SRAM capacity to $1.289 * 10^{-5}$ for 2GB SRAM capacity. The memory sub-system follows the same trend for n=128 and 256 as well. This analysis also suggests that for achieving very high yield, the number of cells per SA must not be too large. The yield of the SRAM chip for various capacities is estimated to be more than 99.9999%.

Chapter 4

A Novel Design of a Dual Functionality Read-Write Driver for SRAM

4.1 Introduction

This section presents the complete design of Dual Functionality Read-Write (DFR-W) driver.

4.2 Proposed design of Dual Functionality Read-Write Driver (DFR-W)

Figure 4.1 shows the proposed design of a dual functionality driver which performs both read and write operations in a SRAM architecture.

Transistors M1-M4 and Footer form the functioning block of DFR-W driver whereas M5-M10 form a control circuitry which aligns the circuit either in read or write mode. The sizing of these transistors is done in such a way that $(w/l)_{5-10}$ is roughly one-twentieth times $(w/l)_{1-4}$.

Table 4.1 shows the status of various signals during the corresponding operation in DFR-W driver.

Table 4.1: Logic Conditions of various signals in DFR-W Driver

Operation	RW_b	Sense	Phibs	Data
Before Read	1	0	0	X
During Read	1	1	1	X
Before Write	0	0	0	X
Writing 0	0	0	1	1
Writing 1	0	0	1	0

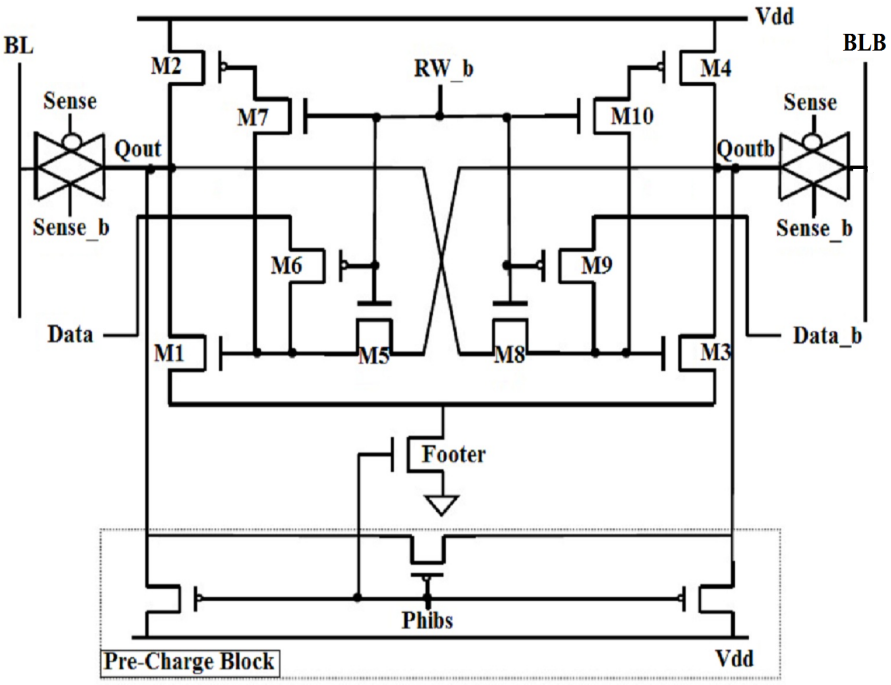

Figure 4.1: Proposed design of DFR-W driver

4.2.1 Reading Operation (RW_b='1')

For reading when RW_b is made '1', as a result, NMOS transistors M5, M7, M8 and M10 are turned on whereas PMOS transistors M6 and M9 are turned off. In this state of logic, transistors M1-M4 (cross coupled) and Footer behave like a latch type SA.

When sufficient potential difference develops between the two bitlines, the pre-charge signal Phibs and sense are made high. The internal nodes of DFR-W driver (Qout and Qoutb) start discharging, but at different rates through the pull down transistors M1 and M3 and the Footer which turns on with the change in the state of Phibs. After the discharging phase, the strong positive feedback formed by M5, M7, M8 and M10 between M1-M4 enforces latching of the two outputs to logic '0' and '1' depending on the data stored in the cell.

4.2.2 Writing Operation (RW_b='0')

For the writing operation in DFR-W driver the RW_b signal is pulled low. As a result, PMOS transistors M6 and M9 are turned on. This links Data and Data_b signals to the gates of pull down transistors M1 and M3 respectively. During write, the pre-charge circuitry is turned off by making Phibs high. Consequently, the NMOS Footer turns on. Sense is kept at logic '0' throughout the write operation so that the bitlines remain linked to the internal nodes of

DFR-W driver.

For writing logic '0' in the bit cell, Data pin is pulled high. As a result, M1 turns on and a discharging path to ground for BL is formed through the transmission gate, M1 and Footer. BL discharges completely and writing of data is accomplished by selection of WL.

For writing logic '1' in the bit cell, Data pin is pulled low (Data_b is high) . As a result, M3 turns on and a discharging path to ground for BLB is formed through the transmission gate, M3 and Footer. BLB discharges completely and writing of data is accomplished by selection of WL.

Figure 4.2 shows the simulation of the proposed DFR-W driver connected to the memory architecture shown in fig. 2.1 with n=128 bit cells. It is depicting the timing diagram of consecutive read and write cycles each having a frequency of 1GHz. Here, nodes Q1 and QB1 represent the internal nodes of the first bit cell whereas nodes QOUT and QOUTB depict the internal nodes of DFR-W driver. PHIB is the pre-charge enable signal for the two bitlines and PHIBS is the pre-charge enable signal of the internal nodes of DFR-W driver. This transient analysis depicts six consecutive read (R) and write (W) cycles following RWWRWR order. This simulation was carried out on ELDO SPICE simulator in 65nm CMOS technology.

4.3 Results

This section provides a comparative study on parameters like speed, power and area between the conventional memory architecture with a latch type SA and a write driver structure II shown in figure 2.6 and the memory architecture integrated with the new DFR-W driver.

4.3.1 Latching Delay

The latching delay is the measure of the speed of memory access. It is defined as the time interval between the instance when the sense signal reaches 50% its maximum value and the differential output of the sensing device reaches 50% of the supply voltage [17].

Figure 4.3 depicts the latching delay of a conventional latch type SA and the proposed DFR-W driver. The comparison has been made for a different number of bit cells per memory slice. The DFR-W driver reduces the load on the bitlines significantly as compared to the conventional memory architecture shown in figure 2.1. As a result, there is a significant improvement in the speed of memory access.

It can be clearly seen that DFR-W driver provides an improvement of 35.58% in latching delay in case of n=64 cells as compared to the latch type SA and an improvement of about 30.31% in memory access speed for n=256 cells per memory slice.

Figure 4.4 shows one complete read cycle of DFR-W driver for n=128. It can be clearly deduced that when RW_b='1' and the sense signal is set, the internal nodes Qout and Qoutb start

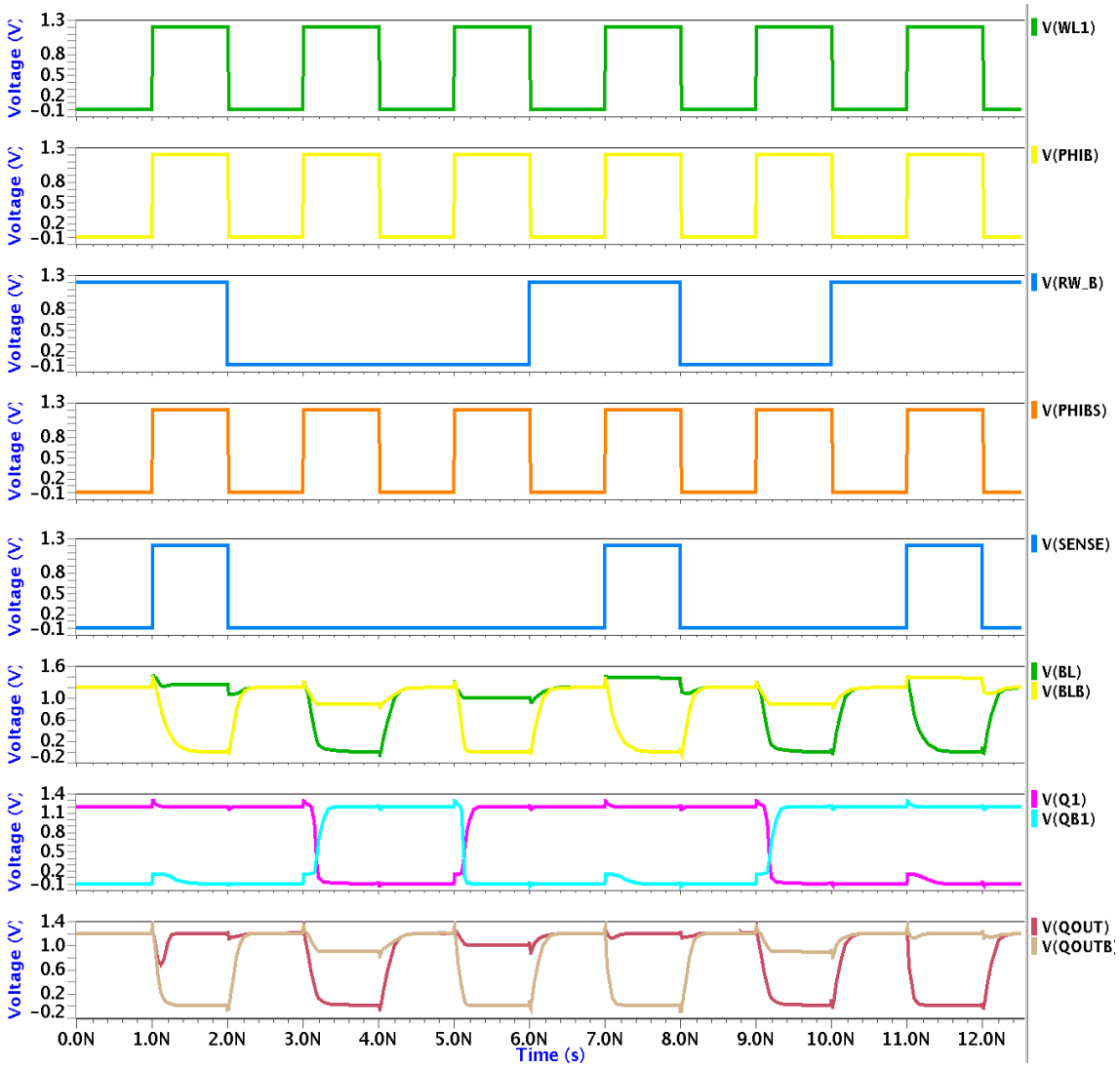

Figure 4.2: Timing Diagram of Read and write cycle of DFR-W driver

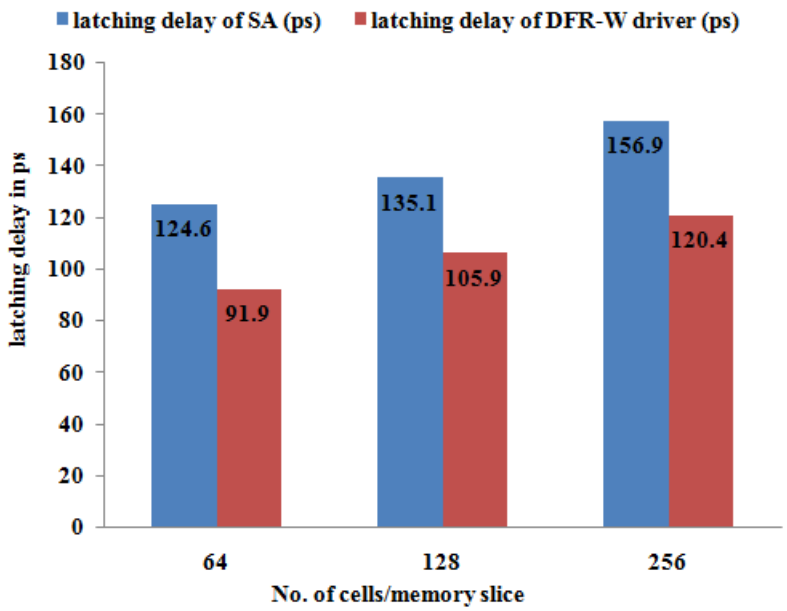

Figure 4.3: Latching Delay of a SA and DFR-W Driver working in read mode

discharging and consequently they settle at the logic levels corresponding to the bit stored in the first bit cell of the memory slice.

4.3.2 Write Time

Write time is the measure of the speed at which data can be written in the memory cell. It is defined as the time taken by the bitline to discharge so that writing operation can be accomplished in one of the cells in the memory array [18].

Figure 4.5 demonstrates the assessment of write time of the write driver and DFR-W driver for different values of n per memory slice. It can be observed that the proposed design is much faster in writing into the cell in comparison with the conventional write driver structure shown in figure 2.6.

For n=64 bit cells in the memory sub-system, it can be clearly observed that there is a decline of 3.4% in writing time of DFR-W driver as compared to the conventional write driver and a decline of 5.7% and 14% in writing time of DFR-W for n=128 and n=256 bit cells respectively.

4.3.3 Leakage Current

In a SRAM architecture, the memory array comprises of symmetric bit cells. At any instance, only one bit cell is selected whereas all others remain in a hold mode or a data retention mode. It is evident that all except one bit cell are off all the time. Ideally in the hold mode of operation,

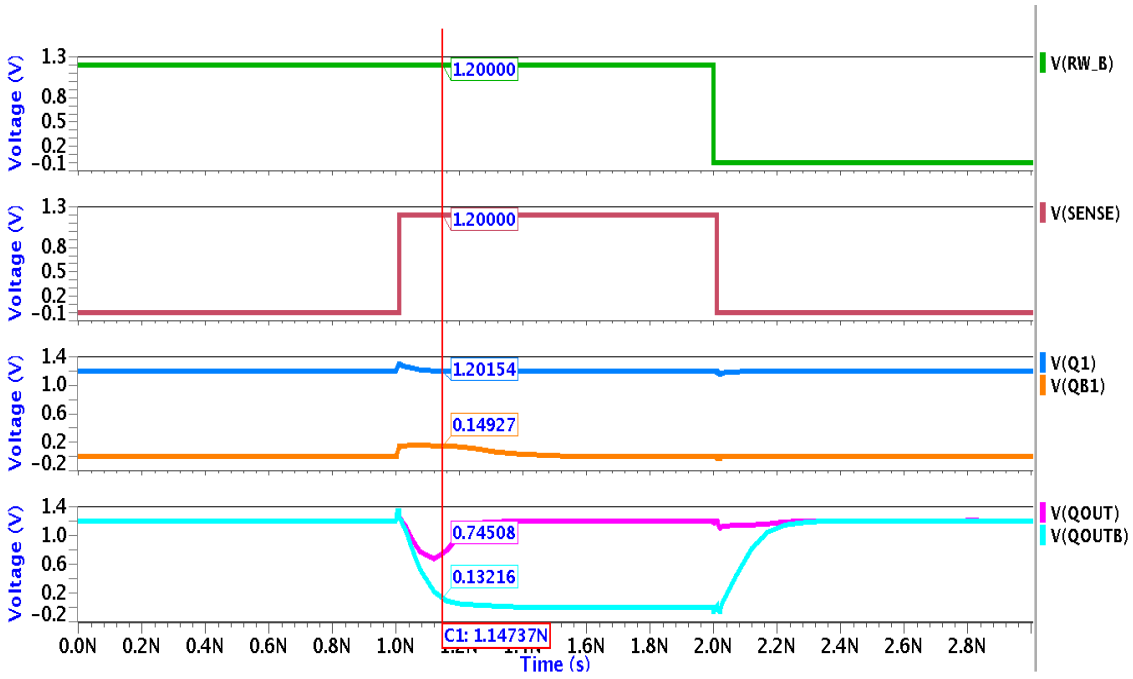

Figure 4.4: Timing Diagram depicting the latching of internal nodes of DFR-W driver during read mode

a cell is not suppose to conduct any current. On the contrary, these cells leak a small amount of current of the order of a few nA. This current is called as sub-threshold leakage current or simply leakage current [19]. Similarly the peripheral circuitry in a SRAM like SA, Write Driver, Address Decoders, Multiplexers etc. are not always on but are selected as per the need of the operation. Hence, leakage in peripherals is as significant as that in the memory array.

Mathematically, equation (1) gives the expression of leakage current [4]. Here V_T is the thermal voltage, V_{gs} and V_{ds} are the gate to source and drain to source voltages respectively of a MOS transistor. V_{th} is the threshold voltage and η is a process dependent term.

$$I_{ds} = I_o e^{\frac{V_{gs}-V_{th}}{\eta V_T}} \left(1 - e^{\frac{V_{ds}}{V_T}}\right) \tag{4.1}$$

Fig. 4.6 depicts the comparison of leakage currents in the peripheral circuitry of the conventional memory architecture and in DFR-W driver when integrated with the memory sub-system.

It can be clearly observed that there is a drastic reduction in the leakage current in DFR-W. SA and write driver together have a leakage current of around 22.8nA whereas the leakage in DFR-W driver which can function both as the SA and the write driver produces a leakage of 8nA.

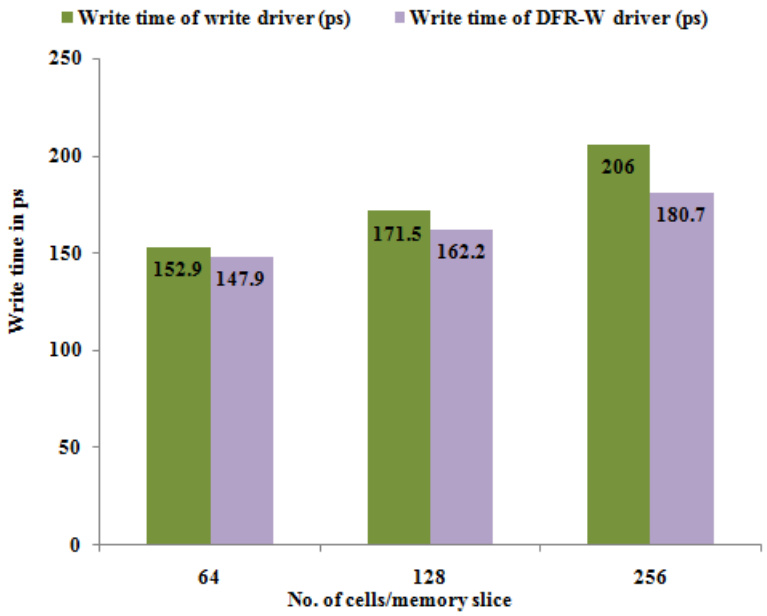

Figure 4.5: Write Time of Write Driver and DFR-W Driver working in write mode

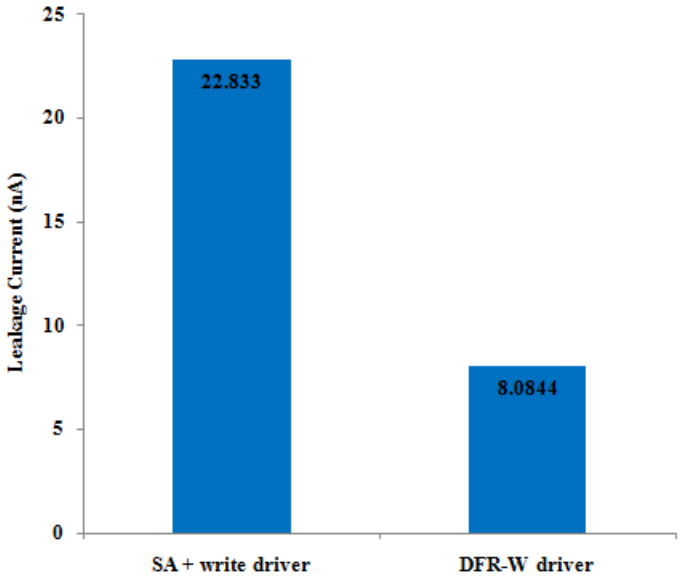

Figure 4.6: Leakage current of a SA and Write Driver and DFR-W Driver

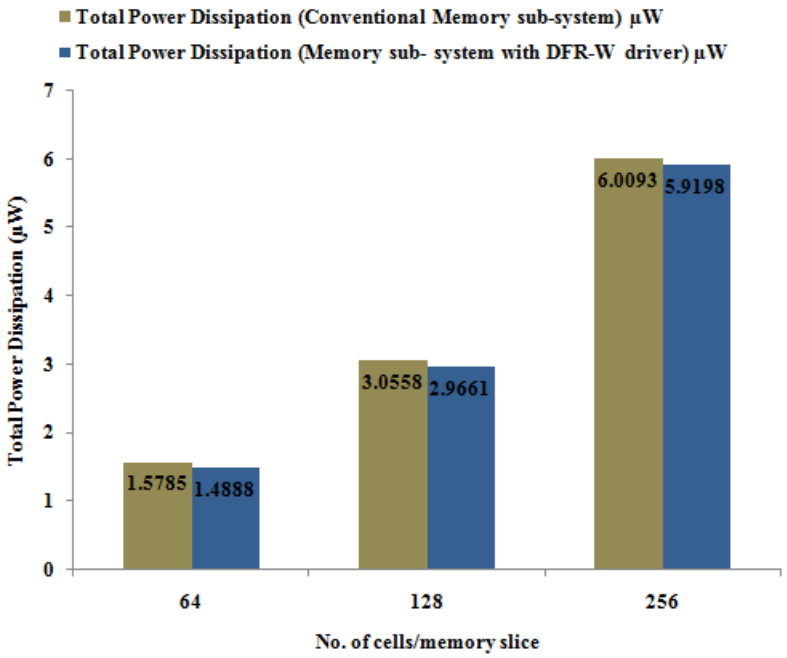

Figure 4.7: power Dissipation in the conventional memory architecture and the memory architecture having DFR-W Driver

4.3.4 Power Dissipation

In CMOS circuits, power dissipation is categorized mainly as Dynamic power and Static power dissipation. Dynamic power is further of two types. The power dissipated due to charging and discharging of the capacitance during dynamic transition of logic states is called switching power, whereas the power which is dissipated due to the shorting of power supply to ground due to the switching of both PMOS and NMOS momentarily, during the transition of logic states is termed as short circuit power. Switching and Short circuit power form a major share of dynamic power.

On the other hand, the main sources of static power dissipation are due to sub threshold leakage in off transistors, due to gate leakage through the gate dielectric and dissipation due to junction leakage in MOS transistors.

Therefore, the total power dissipation of any design is defined as the sum of static and dynamic power [4]. Fig. 4.7 depicts the comparison of power dissipated in a conventional memory architecture and the architecture with DFR-W driver for different capacities. It can be observed that there is a 6% decline in power dissipation in the DFR-W architecture with n=64 cells as compared to the conventional architecture. For n=128 and n=256 bit cells, there is a decline of 3% and 1.5% respectively in power dissipation in the DFR-W architecture cells as compared to the conventional architecture.

4.4 Conclusion and Results

A novel, area and power efficient Dual Functionality Read-Write driver for performing both read and write operations on a SRAM bit cell is proposed in this chapter. It eliminates the need of bulky SA and write driver thereby making the design area efficient. Consequently, the new DFR-W driver reduces the load on bitlines thereby improving the response and access time of memory.

DFR-W memory architecture is simulated at a frequency of 1GHz for varied memory capacities. This chapter also quantifies and compares the latching delay, write time, leakage current and power dissipation of DFR-W memory architecture with the conventional memory architecture comprising of a SA and write driver. The proposed architecture depicts a maximum reduction of 35.58% in the latching delay and a reduction of up to 14% in the writing time. Leakage current in the new design improves greatly to about 8.0844nA in contrast to 22.833nA in the conventional design. The power dissipation reduces up to 6% for the proposed design which makes the design suitable for power efficient application.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Statistical yield estimation of SRAM is a promising technique to estimate the effects of on chip parametric variations like variations in the threshold voltages of transistors due to random dopant fluctuations. These variations, in turn, result in read/write failure.
In chapter three, a new yield estimation technique is presented which considers parametric variations in the bit cell as well as SA. For the sub-system with 64 bit cells, it is estimated that the probability of failure is $4.802 * 10^{-13}$ in a read cycle of frequency 1GHz. Furthermore, the probability of failure for 8MB capacity is $5.035 * 10^{-7}$ while for 2GB capacity it increases to $1.289 * 10^{-5}$. It is also observed that as the load on one SA per column is doubled, the probability of failure of memory slice increases by 70%. The proposed technique estimates the Yield(Y) for SRAM array to be more than 99.9999%.

As a consequence of technology node scaling, memory systems like SRAMs and NVM are becoming denser day by day. Hence, there is an urgent need of area and power efficient designs in memory systems.
In chapter four, a novel dual functionality read-write driver is designed for SRAMs sub-system. The new driver performs both reading and writing operation on a SRAM bit cell. It, therefore, eliminates the need of an explicit write driver for writing. A single circuit performs the function of both SA and write driver. This new design is integrated with a memory sub-system and a comparative analysis is done with respect to a conventional memory sub-system on grounds of leakage, power, speed and area.
DFR-W depicts a reduction of up to 35.58% in latching delay and of about 14% in writing time as compared to the conventional memory architecture. For the new design, there is a drastic decline in leakage current when the device is in hold mode. In DFR-W driver, leakage reduces to about 8.0844nA as compared to 22.833nA in the conventional design. DFR-W driver shows a reduction of up to 6% in power dissipation as compared to the conventional design. The proposed

design performance is found way superior and efficient in terms of speed, area and power.

5.2 Future work

- The work on the statistical yield analysis of SRAM sub-system (in chapter 3) can be used for the yield estimation of complete SRAM chip considering parametric variations in other peripherals like address decoders, multiplexers and write driver.
- In the proposed work (chapter 4), a new DFR-W driver is integrated with the SRAM sub-system. The robustness of the new driver and its tolerance towards on chip parametric variations can be analyzed by the yield estimation of the design using the mathematical model proposed in chapter 3.

Bibliography

[1] C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf *et al.*, "45nm high-k+ metal gate strain-enhanced transistors," in *VLSI Technology, 2008 Symposium on*. IEEE, 2008, pp. 128–129.

[2] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 12, pp. 1859–1880, Dec 2005.

[3] F. Gong, Y. Shi, H. Yu, and L. He, "Parametric yield estimation for sram cells: Concepts, algorithms and challenges," in *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, 2010.

[4] W. N. HE *et al.*, *Cmos Vlsi Design: A Circuits And Systems Perspective, 3/E*. Pearson Education India, 2006.

[5] N. S. Agarwal, Kanak, "Statistical analysis of sram cell stability," in *Proceedings of the 43rd annual Design Automation Conference*. ACM, 2006, pp. 57–62.

[6] J. Yao, Z. Ye, and Y. Wang, "Importance boundary sampling for sram yield analysis with multiple failure regions," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 33, no. 3, pp. 384–396, 2014.

[7] Y. W. J. Yao, Z. Ye, "An efficient sram yield analysis and optimization method with adaptive online surrogate modeling," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 7, pp. 1245–1253, July 2015.

[8] A. K. Gundu, M. S. Hashmi, R. Sharma, and N. Ansari, "Statistical analysis and parametric yield estimation of standard 6t sram cell for different capacities," in *System-on-Chip Conference (SOCC), 2015 28th IEEE International*. IEEE, 2015, pp. 316–321.

[9] H. Park and C.-K. K. Yang, "Stability estimation of a 6t-sram cell using a nonlinear regression," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 22, no. 1, pp. 27–38, 2014.

[10] P. Athe and S. Dasgupta, "A comparative study of 6t, 8t and 9t decanano sram cell," in *Industrial Electronics Applications, 2009. ISIEA 2009. IEEE Symposium on*, vol. 2, Oct 2009, pp. 889–894.

[11] F. Moradi, D. T. Wisland, H. Mahmoodi, and T. V. Cao, "Improved write margin 6t-sram for low supply voltage applications," in *2009 IEEE International SOC Conference (SOCC)*, Sept 2009, pp. 223–226.

[12] A. Pavlov and M. Sachdev, *CMOS SRAM circuit design and parametric test in nano-scaled technologies: process-aware SRAM design and test.* Springer Science & Business Media, 2008, vol. 40.

[13] C. H. Hong, Y. W. Chiu, J. K. Zhao, S. J. Jou, W. T. Wang, and R. Lee, "A low-power charge sharing hierarchical bitline and voltage-latched sense amplifier for sram macro in 28 nm cmos technology," in *2014 27th IEEE International System-on-Chip Conference (SOCC)*, Sept 2014, pp. 160–164.

[14] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, July 2004.

[15] A. Ney, P. Girard, S. Pravossoudovitch, A. Virazel, M. Bastian, and V. Gouin, "A design-for-diagnosis technique for sram write drivers," in *2008 Design, Automation and Test in Europe*, March 2008, pp. 1480–1485.

[16] R. D. Yates and D. J. Goodman, "Probability and stochastic processes," *John Willey & Sons*, 1999.

[17] A. Hemaprabha, K. Vivek, M. Vaijayanthi, and S. Sabeetha, "Implementation of latch type sense amplifier."

[18] E. Karl, Y. Wang, Y. G. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr, "A 4.6ghz 162mb sram design in 22nm tri-gate cmos technology with integrated active vmin-enhancing assist circuitry," in *2012 IEEE International Solid-State Circuits Conference*, Feb 2012, pp. 230–232.

[19] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, Feb 2003.