Variability-Aware Design of Static Random Access Memory Bit-Cell

by

Vasudha Gupta

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2008

© Vasudha Gupta, 2008

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The increasing integration of functional blocks in today's integrated circuit designs necessitates a large embedded memory for data manipulation and storage. The most often used embedded memory is the Static Random Access Memory (SRAM), with a six transistor memory bit-cell. Currently, memories occupy more than 50% of the chip area and this percentage is only expected to increase in future. Therefore, for the silicon vendors, it is critical that the memory units yield well, to enable an overall high yield of the chip. The increasing memory density is accompanied by aggressive scaling of the transistor dimensions in the SRAM. Together, these two developments make SRAMs increasingly susceptible to process-parameter variations. As a result, in the current nanometer regime, statistical methods for the design of the SRAM array are pivotal to achieve satisfactory levels of silicon predictability.

In this work, a method for the statistical design of the SRAM bit-cell is proposed. Not only does it provide a high yield, but also meets the specifications for the design constraints of stability, successful write, performance, leakage and area. The method consists of an optimization framework, which derives the optimal design parameters; i.e., the widths and lengths of the bit-cell transistors, which provide maximum immunity to the variations in the transistor's geometry and intrinsic threshold voltage fluctuations. The method is employed to obtain optimal designs in the 65nm, 45nm and 32nm technologies for different set of specifications. The optimality of the resultant designs is verified. The resultant optimal bit-cell designs in the 65nm, 45nm and 32nm technologies are analyzed to study the SRAM area and yield trade-offs associated with technology scaling. In order to achieve 50% scaling of the bit-cell area, at every technology node, two ways are proposed. The resultant designs are further investigated to understand, which mode of failure in the bit-cell, becomes more dominant with technology scaling. In addition, the impact of voltage scaling on the bit-cell designs is also studied.

Acknowledgements

I would like to acknowledge my supervisor, Professor Mohab Anis, without whose invaluable guidance and encouragement, this thesis would not have been possible. I thank my co-supervisor, Professor M. Elmasry, for his support. I would also like to thank the student members of the VLSI group, especially Javed Jaffari, Akhilesh Kumar and Mohamed Abu-Rahma, with whom I had several discussions on the issues related to this work.

I am grateful to the readers of my thesis, Professor Manoj Sachdev and Professor John Hamel. I also thank the technical and administrative staff of the Department of Electrical and Computer Engineering, University of Waterloo.

I am deeply grateful to my husband, Gaurav, whose love and encouragament helped me stay motivated and focussed. He has been a consistent source of strength for me. I also take this opportunity to thank my parents and my sister, whose thoughts, concern and ever-encouraging words rejuvenated me and made my days brighter. I also thank my parents-in-law and all my friends.

To my husband and my parents

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

1.1.1 Evolution of Embedded Memory

The rapid development of CMOS technology over the last three decades, has been fuelled by technology scaling and consistent improvement in the MOSFET manufacturing processes. The concept of MOSFET memory was perfected and commercialized in the seventies [1]. Robert Dennard of IBM conceived the dynamic memory cell (a memory cell is a circuit capable of storing single bit of information - “1” or 0“”) using a single MOSFET, and a capacitor in 1968 [2]. With several process improvements to control the leakage, the first single MOSFET dynamic random access memory (DRAM) chip with 2k-bits was developed in 1971. Over the next several years, DRAMs were employed in a widespread manner as the main computer memory.

However, DRAM performance has not kept pace with the improving processor performance, as depicted in Fig. 1.1 [3]-[4]. The growing gap between the processor and the DRAM performance has necessitated the introduction of several levels of

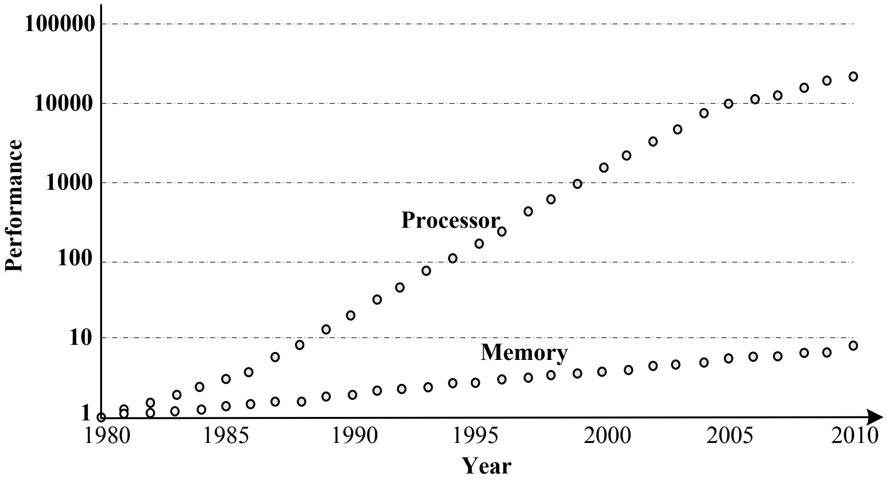


Figure 1.1: Processor and memory performance over time. Baseline for memory performance is 64kB DRAM in 1980 [4].

memory hierarchy [4], ranging from high-performance, small sized but more costly on-chip memories to slower, large sized but affordable off-chip DRAM, magnetic or optical memories. To improve the system performance, the processor tries to keep the frequently used data and instructions closer to itself, that is, in the faster on-chip memory, which is called the “cache”. For example, in personal computers, on-chip cache levels are often called L1 and L2 memories. The memory hierarchy is depicted in Fig. 1.2. Addresses from a slower, larger memory are mapped onto a faster, smaller memory in the next level, which is closer to the processor. The speed and the cost per bit increase as one moves from the secondary storage to the registers.

On-chip cache memories provide faster access times mainly by eliminating the delay across the chip interface, and by employing smaller capacity memory blocks. To realize an on-chip cache, the use of high-density, single transistor, embedded DRAM may seem plausible. However, if the standard logic process is used to fabricate embedded DRAMs, the memory exhibits high leakage. This is because

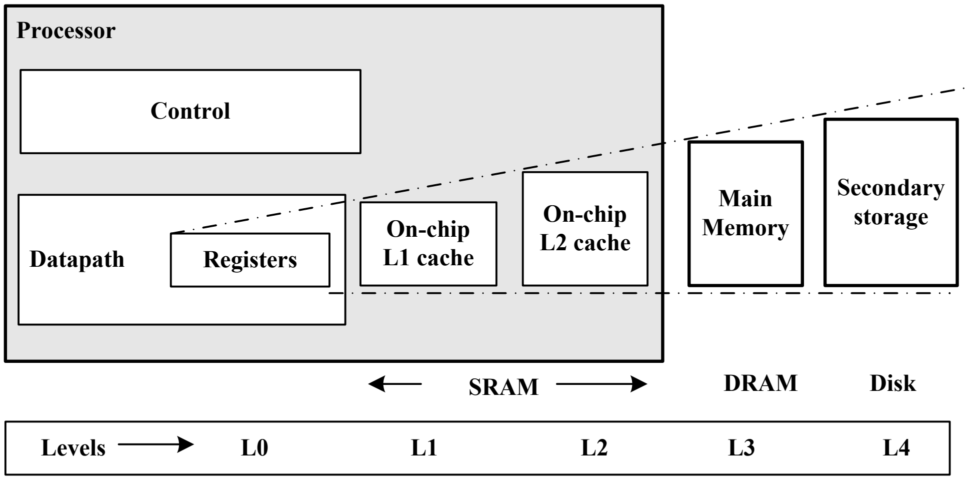


Figure 1.2: Memory hierarchy.

the transistor threshold voltage in the standard logic process is relatively lower than that in the standard DRAM process. The leakage can be controlled if the embedded DRAM cell is designed with more than one transistor. But the associated area penalty undermines the area advantage that DRAMs have over six-transistor static random access memories (SRAM). Alternatively, one can use the standard DRAM process to achieve a high density (1T) on-chip cache. But since this process involves a high threshold voltage to limit leakage, it also limits the performance of the system, and the cache may not serve its purpose.

The DRAM cell stores charge on a capacitor to realize memory, as depicted in Fig. 1.3(a). Compared to this, the six transistor SRAM cell has a feedback latching mechanism to retain data. An SRAM memory cell has a flip-flop like circuit, which enables storage of data indefinitely- as long as the power supply remains available. Because SRAMs do not store data on capacitors, they do not require “refreshing” as DRAM does [1]. Therefore, the primary advantage of the SRAMs stems from the fact that the processor can fetch data from SRAM at a faster rate than it can from the DRAM, because a significant part of the DRAM’s

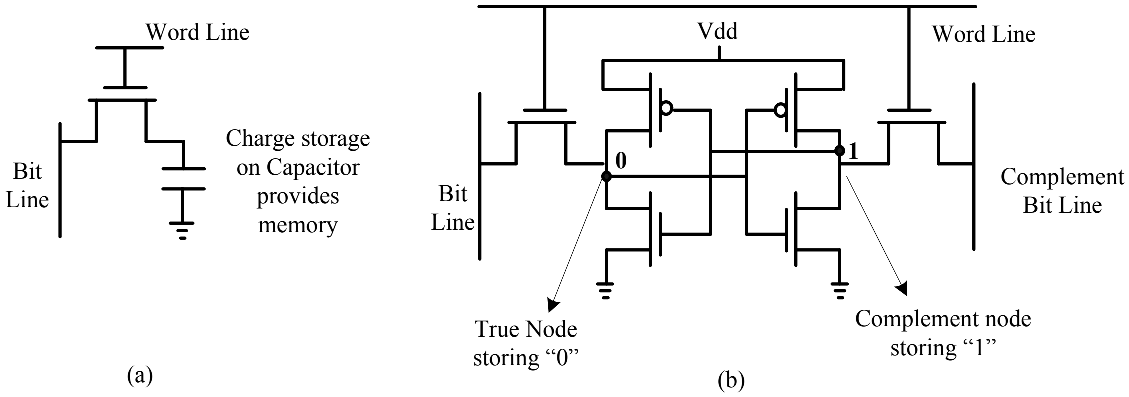


Figure 1.3: (a) DRAM cell with single transistor (1T) and capacitor (b) SRAM cell with six transistors (6T). Data is read out from the cell or written into the cell, when the word line turns on the access transistor. Bit line holds the read out data or the data that is to be written into the cell.

cycle time is consumed by the "refresh" operation. Another advantage with the embedded SRAMs is that these can be fabricated with the standard logic process and do not require any additional steps (which are needed for DRAMs, e.g. to fabricate the storage capacitor). Therefore, today, on-chip memory is most often realized with the embedded SRAMs.

The disadvantage with the SRAMs is the associated cost. An SRAM cell employs several transistors (instead of one transistor and capacitor in DRAM) to store a single bit of data, and occupies more area than the DRAM cell. Therefore, for the same chip area, a DRAM chip would enable storage of more bits (more memory capacity) than the SRAM chip. Assuming that the cost to manufacture the two chips is similar, the cost per memory bit is higher for the SRAM (less memory for the same cost). This explains the incessant demand to design SRAM cells within the smallest possible area. SRAM approaches other than the 6T cell such as the 4T and the 5T versions or cells using resistor loads may be used. The SRAM cell size may reduce significantly with these other approaches, but at the cost of the

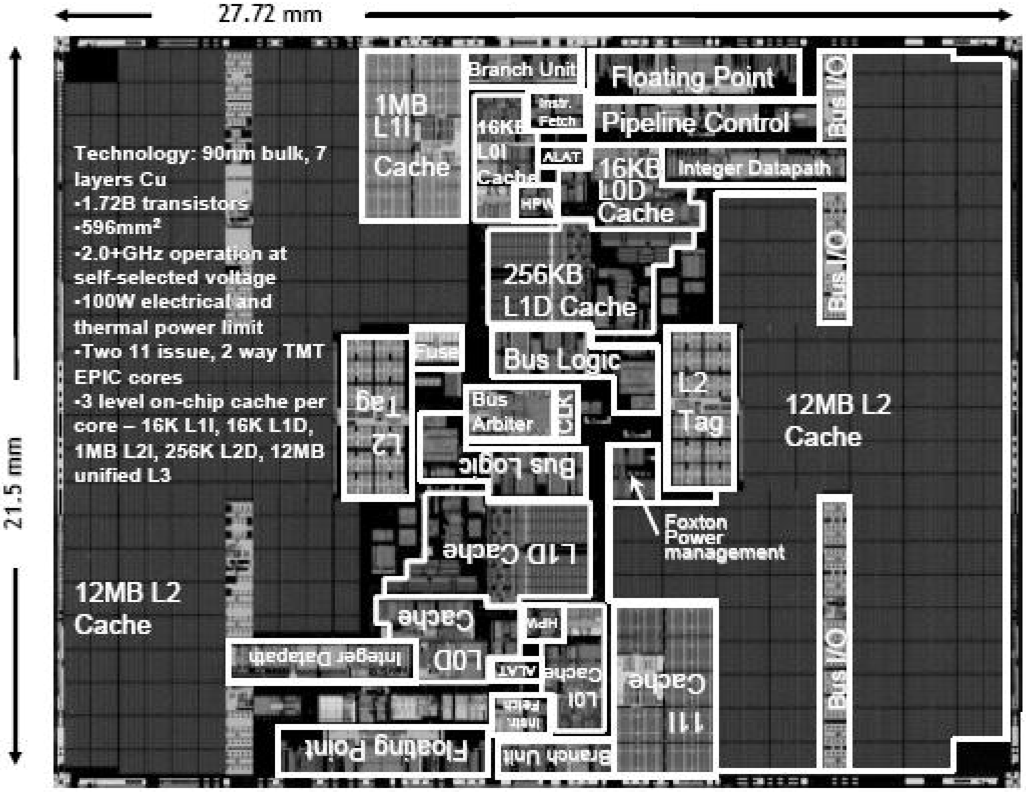


Figure 1.4: Micrograph of the dual-core Itanium-2 processor, Source: Intel 2005 ISSCC papers [5].

additional technology steps, which are required to develop the stacked loads. The 4T and 5T versions also suffer from degraded noise margins, especially at low voltages. Therefore, the 6T version of the SRAM cell remains the most cost-effective choice to be deployed as the embedded memory. Even though the 6T SRAM cell does not require any additional processing steps, some modest technology enhancements, such as the shared gate and diffusion contacts (explained later) and tighter layout design rules, can greatly improve the SRAM density.

Large quantities of on-chip memory enhance the data storage and manipulation capacity of the chip, resulting in higher speeds and enabling increasing integration of more and more functionality on the same die. Fig. 1.4 shows the micrograph of

Intel's Dual-Core Itanium-2 processor code named 'Montecito', which was discussed in 2005 [5], and released in 2006. It consists of 1.72 billion transistors in all. Out of these, only 57 million form the core logic. As many as 1.55 billion transistors lie in the 12MB caches on the left and right flanks of the chip There are other cache and tag memories and all in all, more than 90% of the chip area is occupied by embedded memory. Even for processors such as ARM which are used in mobile phones and do not perform extensive number crunching, and for ASICs used in cameras, etc., memory occupies more than 50 percent of the die area. According to the ITRS (International Technology Roadmap for Semiconductors) [6], the on-chip memory density is only going to increase in future. Moreover, unlike the logic gates, where the impact of variability on the circuit metrics such as delay, gets averaged out; in memories, every single cell must function reliably. Therefore, a high-yielding embedded SRAM is absolutely critical to ensure an overall high chip yield.

1.1.2 Technology Scaling

A spectacular increase in the integration density and computational complexity in digital integrated circuits has been witnessed in the last few decades. Fig. 1.5 shows the total number of transistors in the Intel microprocessors, starting with the first microprocessor 4004 to the recent Pentium 4 microprocessor [7]. The graph indicates that the total number of transistors has doubled almost every 2 years. This is in line with the prediction made by Gordon Moore in 1965 (often called the Moore's law) [8]. Fig. 1.6 shows that the microprocessor frequency has doubled in every generation and Fig. 1.7 demonstrates the increase in the size of the first and second level caches for the 7 generations of Intel microprocessors [9].

Underlying these revolutionary changes - increasing transistor count and improving speeds, are the advances in the device manufacturing technology, which allow for a steady reduction in the minimum feature size, such as the minimum

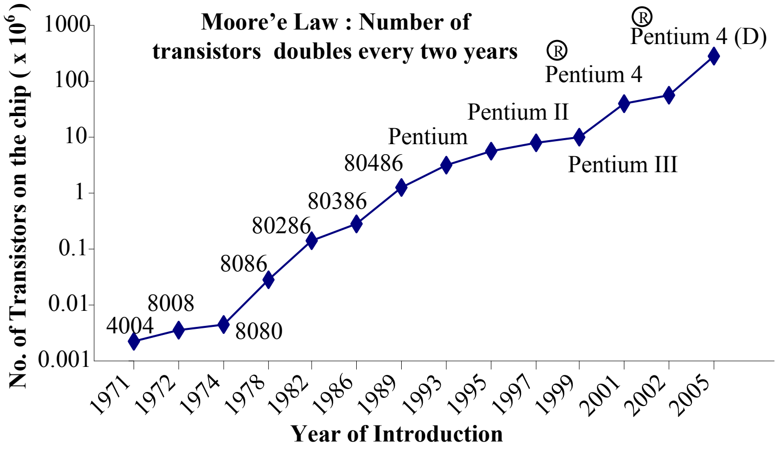


Figure 1.5: Moore's Law. Transistors on a chip [7]

transistor channel length realizable on a chip [10]. The set of manufacturing processes and techniques, which are used to achieve the minimum feature size, are referred to as the "technology node". As the manufacturing processes are improved to reduce the minimum realizable feature size, the technology is said to "scale" from one node to the next.

This scaling of the transistor dimensions ($\sqrt{2}$ shrink in each lithographic dimension - length, width and effective oxide thickness of transistors) is accompanied by a scaling of the supply voltage to keep the dynamic power consumption under control. Hence, the transistor threshold voltage is also commensurately scaled to maintain a high drive current. Overall, this paradigm of technology scaling results in a reduction in the intrinsic capacitance, which enables a faster switching time. This provides increased performance and reduced power consumption ($Power = CV^2f$), while packing in more devices in the same area, which effectively lowers the cost per transistor. Growing logic and memory density enables increasingly complex products. Moreover, many of the off-chip components can now be integrated on-chip, which further reduces cost. Therefore, the idea of technology scaling is a very attractive. The IC industry has worked aggressively to continue this trend of

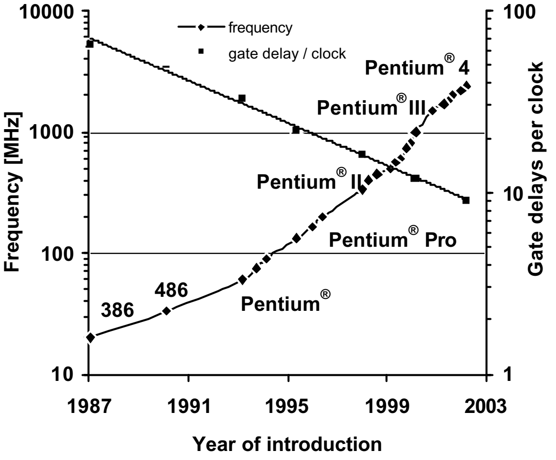


Figure 1.6: Frequency doubled and number of gates per clock reduced by 25% per generation [9]

technology scaling, and endeavours to do the same in near future.

1.1.3 Variability

However, with the scaling of transistor dimensions, in the nanometer regime, fundamental limits are being approached [10]-[12]. It is becoming increasingly difficult for the process engineers to control certain device and interconnect parameters such as channel lengths, interconnect dimensions, contact shapes and parasitics, inter-layer dielectric thicknesses and dopant concentrations. This is because of the fact, that the manufacturing precision has not scaled proportionately with the device and interconnect dimensions. As a result, the relative variation in the device and interconnect geomtery has increased. For instance, a 2nm variation in the channel length may not be an important factor at 180nm generation (target channel length = 180nm), but becomes significant at the 45nm generation. Additionally, growing die size has contributed to an increase in the within-die variations. Therefore, in the modern nanometer era, the circuit performance traits such as delay and power

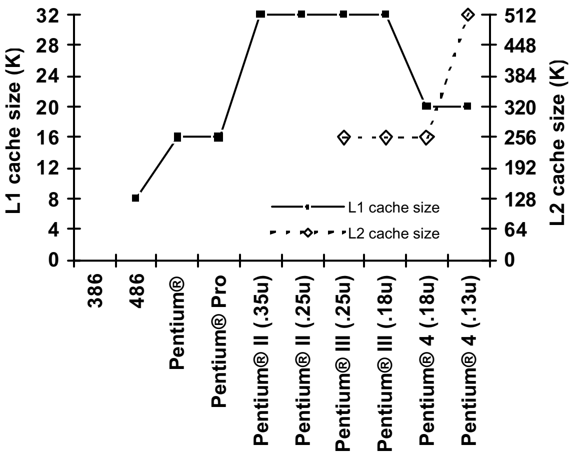


Figure 1.7: Increasing on-chip cache size [9]

become increasingly sensitive to variability. In fact, variability has been elevated to a first-order limitation to continued technology scaling. This process and device variability challenge to continued technology scaling is the most urgent problem confronting the designers. The problem is even more serious for the SRAM array, because it employs the minimum sized transistors and because the SRAM increasingly occupies a greater percentage of the chip area. Therefore, a variability aware design of the SRAM array is essential to achieve a high SRAM yield, and to enable continual technology scaling.

1.2 Contributions of this work

- In this work, an optimization framework, for the statistical design of the SRAM array, is proposed. The objective is to provide an efficient, yet simple and easy to deploy design technique, for the SRAM circuit designers. The proposed method addresses the increasing process variability considerations, upfront, during the design phase to generate an optimal SRAM bit-cell design, which is robust enough to withstand the process variations in the transistor geometrical dimensions and intrinsic threshold voltage fluctuations, and therefore, has a high yield. The resultant optimal design also meets the desired specifications of area, stability, functionality, speed and leakage. With the proposed method, optimal SRAM designs are obtained and the yield verified using Monte Carlo simulations. With the results, it is shown that the conventional sizing strategy is no longer sufficient to ensure high yielding bit-cell designs, and a statistical design approach is essential in the latest technologies.

- An improved problem formulation for the statistical design method is presented. Because the SRAM bit-cell is arrayed to achieve large quantities of memory, the area of the SRAM bit-cell is very important from the economic point of view. Traditionally, SRAM bit-cell area has scaled by 50% every technology [13], and this is the most important design requirement. This work proposes two ways to achieve 50% scaling in the nanometer regime. These are (a) use of progressively longer transistor lengths and (b) partitioning. Use of longer transistors to improve scaling seems counter-intuitive. However, it is shown in subsequent chapters, how this concept works. Well-scaled designs in the 65nm, 45nm and 32nm technology nodes are achieved by employing these two principles. The impact of technology scaling is investigated.

- Additionally, the impact of voltage scaling on the SRAM array design is also studied. Relaxing performance requirement, in the face of voltage scaling, helps achieve smaller area for the SRAM bit-cell. But the area benefit diminishes at 32nm technology, when the design yield is limited by static noise margin and not performance.

1.3 Organization of the Thesis

The remainder of this thesis is organized as follows

Chapter 2 provides the background for this work. In the first section, SRAM basics, including SRAM components, organization and operation are described. The four metrics of the SRAM array design - static noise margin, write switching voltage, read speed and leakage are explained in the next section. Subsequently, the various sources of variability are described. The increasing impact of variability on the SRAM design metrics is demonstrated to motivate the development of a statistical design procedure for the bit-cell.

Chapter 3 explains the proposed method. The constraints of the design problem are formulated. The design yield is defined and the optimization framework is developed. The results of optimization are presented for a set of requirements in the 45nm technology. The optimality of the resultant 45nm design is also verified.

Chapter 4 presents an improved version of the statistical bit-cell design method. Optimal designs in the 65nm, 45nm and 32nm technologies are derived with the improved method, and analyzed for the area and yield trade-offs. Two ways - progressively longer transistors and partitioning, to improve the area scaling of the SRAM bit-cell are then explained. The resultant optimal designs, with these two principles, scale as per expectations. The impact of voltage scaling is analysed and finally, summarised recommendations are made for SRAM array design.

Chapter 5 concludes this dissertation and outlines future work.

Chapter 2

Background

In the first section, the basic single port SRAM architecture is described. The SRAM read and write operations are explained in detail. This provides the requisite background to discuss the design care abouts for the SRAM array. In the next section, the major sources of variability are discussed. The impact of variability on the SRAM design metrics is demonstrated to build the case for a statistical approach for the design of the SRAM bit-cell.

2.1 SRAM Concepts

2.1.1 SRAM Architecture

Fig.2.1 presents a static random access memory (SRAM) of size (or number of bits stored) $m \times n$, where m is the number of words and n is the number of bits per word. The figure indicates the main inputs for a synchronous, single port memory: CLK (input clock), *Addr* (address of the memory location, which is accessed for read or write), R/W (control signal specifying read or write), EN (memory enable, a memory operation can be performed at the CLK edge, only when EN is asserted)

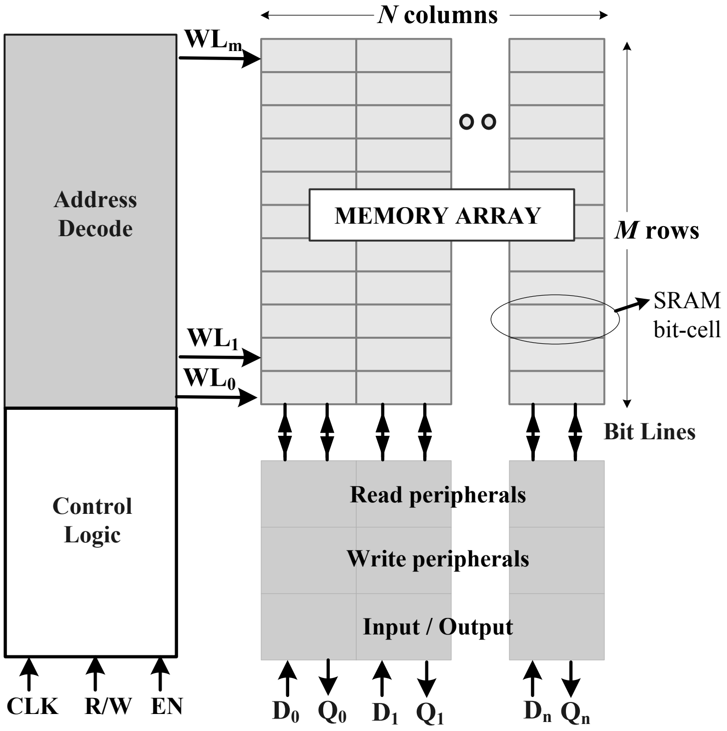


Figure 2.1: Basic SRAM architecture

and the data lines - $D_0, D_1, .., D_n$, which hold the input data for the write operation. The lines - $Q_0, Q_1, .., Q_n$ constitute the outputs of the memory [12], [14].

In addition to the memory array, which stores data, the other fundamental building blocks of the SRAM are the row and column peripheral circuits and the control block. When the word line of a row is turned 'ON', all the memory bits in the selected row become "active" and can be accessed for read or write operations. To decode m word lines, one needs $log_2 m$ address bits. The address latches and predecoders as well as internal clock generation circuits (for sequencing read/write sub-operations) are placed in the control block. The row peripheral circuits, adjacent to the array, consist of the word line decoders and drivers.

The column periphery sits at the bottom of the memory array. The information read out on the bit-lines during the read operation (explained later) is amplified by the circuits in the read peripherals, and buffered out onto the output lines $(Q_0, Q_1, .., Q_n)$. During the write operation, the information on the data lines - $D_0, D_1, .., D_n$, is processed in the write peripherals and presented on the bit-lines for a subsequent write. The column periphery usually contains other circuits for redundancy, built-in-self test collar, selective write, etc. These are not central to this dissertation and are therefore, not discussed.

2.1.2 Read Operation

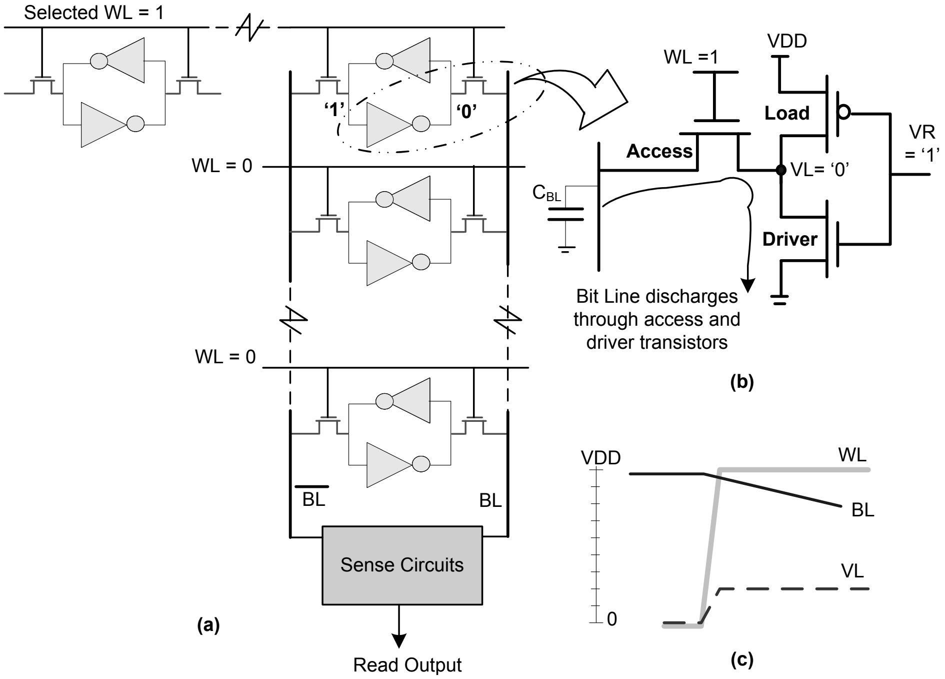


Figure 2.2: Read Operation (a) SRAM components (b) Voltage divider action (c) Transient simulation waveforms to show bit line discharge and rise of node VL to intermediate voltage

When the memory is not accessed for any operation ($EN = 0$), the bit lines are precharged to logic "1". At the onset of "read" or "write", the precharge is released. Because a bit-line runs through all the bit-cells in a column, the resultant bit-line capacitance is large, and therefore, the precharged state on the bit-line is maintained due to charge storage. Subsequently, the selected word line is turned ON to enable the access transistors of all the bit-cells in the corresponding row. This connects the cell internal state to the respective bit lines. This is depicted in Fig. 2.2 (a).

Fig. 2.2 (b) shows the schematic half-cell view of a bit-cell, which is accessed for "read". Node VL stores "0". The stack formed by the access and the driver transistors provides a discharge path for the bit-line capacitance. In principle, the complement bit line remains high, though it also goes down a little bit because of the coupling with the true bit-line. Fig. 2.2 (c) depicts the waveforms during the read cycle. When a sufficient voltage differential develops between the true and the complement bit-lines, the sense amplifier is enabled. The amplified signal is buffered out as read output. The required input differential for the sense-amplifier ranges from 60-200 mV, which is much smaller than what would be needed to trip a logic sense inverter (about half of VDD). Since the bit-line discharge rate is quite small (in the range of 10mV/100ps for large memories), sense amplifiers significantly speed up the read operation [15]-[17]. The array bit lines are usually isolated from the sense bit lines to reduce the load on the sense bit lines. This is easily achieved as memories usually have column multiplexing (discussed later). The sense amplifier enable signal should be asserted at just the right time. If it is too late, it compromises the performance. If it is asserted too early, insufficient input differential voltage may result in erroneous read. To achieve the right timing, self-timing and dummy tracking circuits are employed commonly [12]. These are not discussed in this dissertation.

It can be observed from Fig. 2.2 (c), that VL, storing "0", rises to an intermediate voltage level due to the potential divider action between the driver and the access transistors. This rise should be small; if the voltage at VL becomes higher than the trip point of the other inverter, the bit-cell can potentially flip. Therefore, for a non-destructive "read", the driver should be sized stronger than the access transistor to ensure that the node VL remains closer to the ground level during "read". As will be shown in the subsequent sections, degradation of "0" or "1" logic levels reduces the static noise margin, which can cause stability issues.

2.1.3 Write Operation

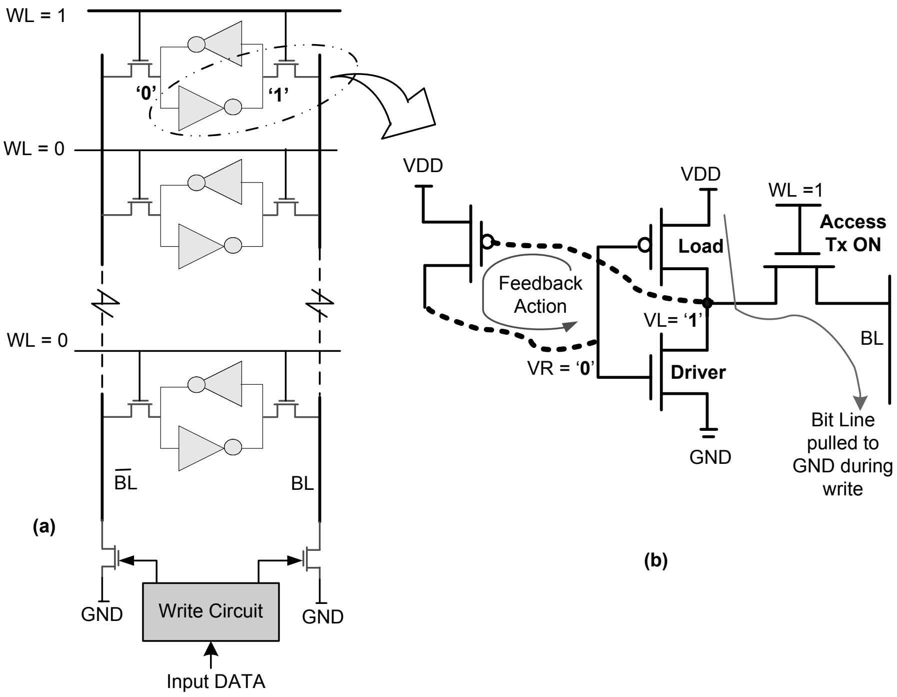


Figure 2.3: (a) SRAM write operation (b) Bit-cell dynamics during write operation

The memory write is usually a write "0" operation, i.e., logic "0" is written to overwrite the node storing logic "1". The input data is decoded to pull the appropriate bit-line (true or complement) to ground through a strong NMOS device, as depicted in Fig. 2.3(a). The operational stack during the write operation is formed by the load and access transistors, in series. This is demonstrated in Fig.

2.3(b). The PMOS load transistor must be overpowered to overwrite logic “1” at node VL. Therefore, the access transistor is made stronger than the load transistor. As VL falls below the threshold voltage of the PMOS of the other inverter, feedback action takes over to write “1” into the complement node - VR, and “0” into VL. It can be deduced why write “0” is the preferred mode of writing into the bit-cell. Writing “0” requires that the bit-line be pulled to ground by an NMOS device, which can be sized smaller than the corresponding PMOS device, which would be needed if a write “1” mode is employed.

2.1.4 Memory Organization

The column peripheral circuits such as the sense amplifier and the write drivers consist of large sized transistors. It is usually not possible to lay out these peripherals in the same pitch as that of the single bit-cell, because the bit-cell area is optimized to be the minimum. Therefore, the column periphery is shared by multiple cells, usually 4, 8 or 16 (a power of 2), in the same row. This concept is demonstrated is Fig. 2.4. Bit lines of the four successive cells, in the accessed row, are multiplexed through the 4 to 1 column select logic, to finally interfaces with the read/write peripheral circuits. This kind of array multiplexing provides variable aspect ratios and power-performance trade-offs for the customer.

2.2 SRAM Array Design Metrics

The quality of the SRAM array design is assessed by measuring certain design metrics. The key design metrics are the static noise margin (SNM), write switching voltage (Vtrip), read current and leakage. Of these, SNM and Vtrip are the functional metrics. With these, one can analyse whether the memory has enough noise margins, and whether it is possible to read or write into the memory successfully. Read current and leakage are the performance metrics. The specifications for the performance metrics depend on the overall desired memory performance and leakage numbers.

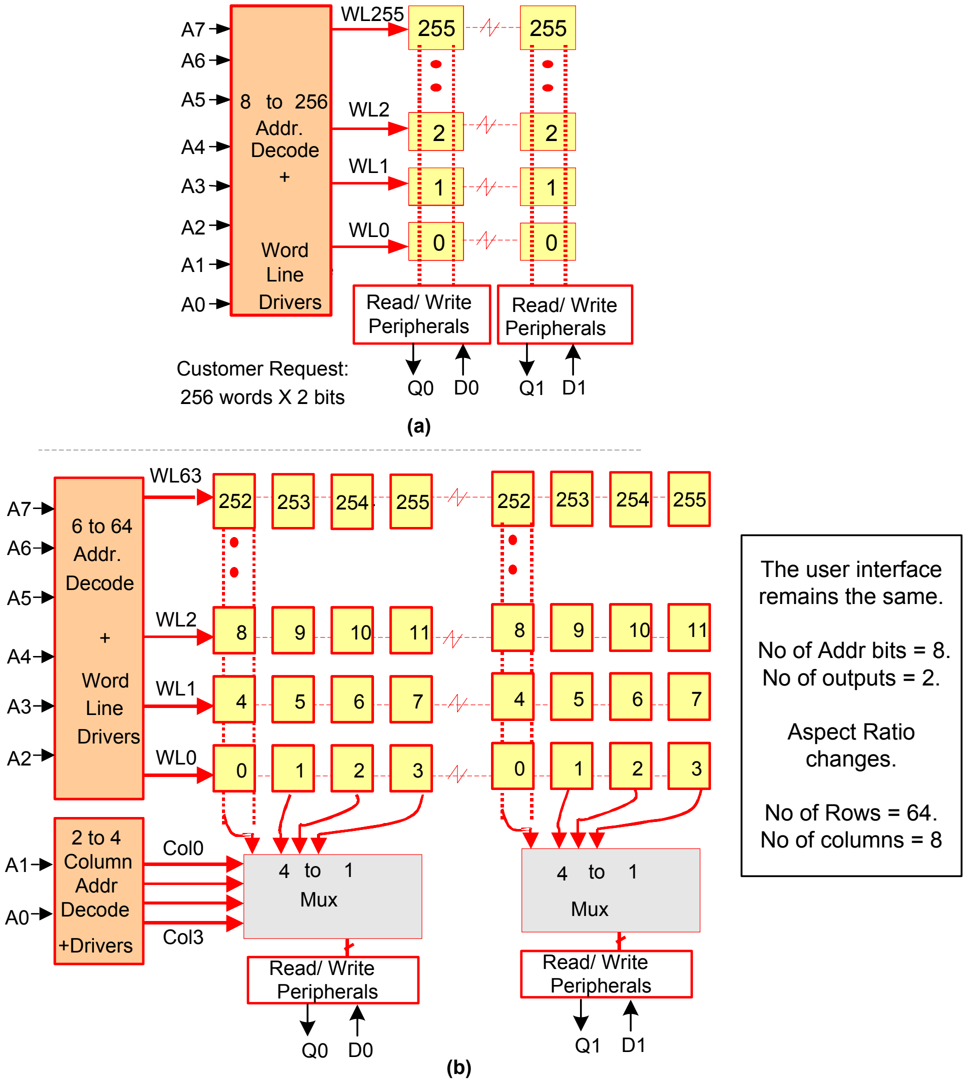


Figure 2.4: (a) SRAM without multiplexing (b) SRAM with mux-4 architecture

2.2.1 Static Noise Margin

Definition

Static Noise Margin (SNM) is defined as the maximum static spurious noise that the bit cell can tolerate while still maintaining a reliable operation [18]-[19]. It is called static as it considers the DC sources of noise (static in time) such as variations in the transistor sizes due to process spread, supply voltage degradation due to IR drop, threshold voltage mismatch in the devices due to random dopant fluctuations and layout differences such as poorly formed contacts and vias. However, the SNM of a good design should be sufficient to withstand dynamic noise sources such as coupling, soft errors, supply voltage fluctuations, change in voltage dependent capacitances in the bit cell, slope of the word line, etc. In this work, SNM refers to the noise margin, when the word-line is turned ON. Retention Noise Margin (RNM) refers to the noise margin, when the bit-cell is not accessed, i.e., when the word-line is OFF. It is explained in the next section that the noise margin degrades when the word-line is turned ON, therefore the SNM is smaller than the RNM.

Measurement

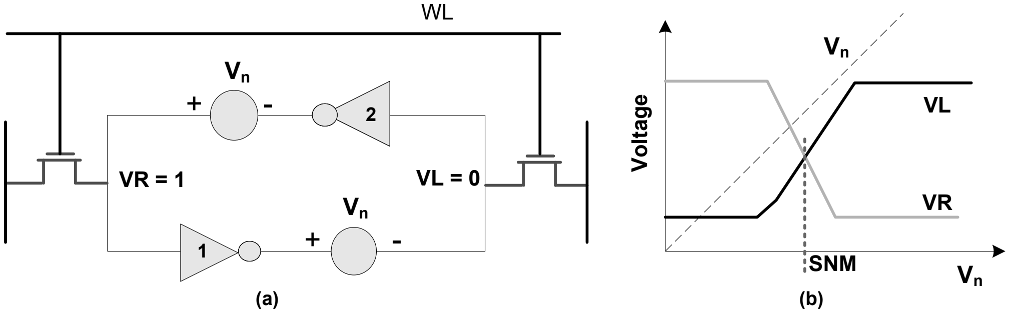


Figure 2.5: (a) Circuit to measure SNM (b) DC simulation

The DC sources of noise can be modeled as voltage sources V_n connected in the feedback path as shown in Fig. 2.5. The polarity of the noise sources is such as to worsen the voltage levels at both 'true' and 'complement' nodes, at the same time. This is done to apply the worst-case DC noise to the system. Here, the worst-case implies a state of the system, which would become unstable with the minimum noise. For example, if the noise source V_n is applied just to worsen logic "1" and not logic "0", then a certain V_n voltage would trip the cell. This, however, would not be the worst-case cell flip scenario, because the bit-cell can potentially flip for a smaller V_n, if the noise impacts the 'true' and 'complement' nodes simultaneously [19].

Since a CMOS inverter is also an amplifier and the condition $R_{in} >> R_{out}$ is applicable (gate current is nearly 0, which makes the input resistance = V/I, infinitely large as compared to the output resistance), the shape of the transfer curves does not change with noise and this kind of modeling is valid. Fig. 2.5 shows how SNM can be measured using a DC circuit simulator. A DC sweep is applied at V_n. The minimum value of V_n, for which the cell flips or gets disturbed, is the minimum noise margin that the bit-cell can tolerate. This is therefore, the SNM.

Qualitatively, SNM can be understood by plotting the transfer curves of inverters 1 and 2, in Fig. 2.5(a), super-imposed on each other. This is depicted in Fig. 2.6. Solid Curves I and II correspond to voltage transfer curves of inverters 1 and 2 respectively.

The transfer curves intersect at three points- A, B and C. However, point C has a very huge gain and is a metastable point. Therefore the system has two stable states; first when VL = 0 and VR = 1 (point A) and second, when VL = 1 and VR = 0 (point B). The bit-cell in Fig. 2.5 (a) rests at point A. During the 'read' operation, the voltage at node VL, which stores "0", rises to a non-zero value as mentioned before. Therefore, it can be observed that the VL voltage at point A is more than 0. This explains why the noise margin becomes worse when the

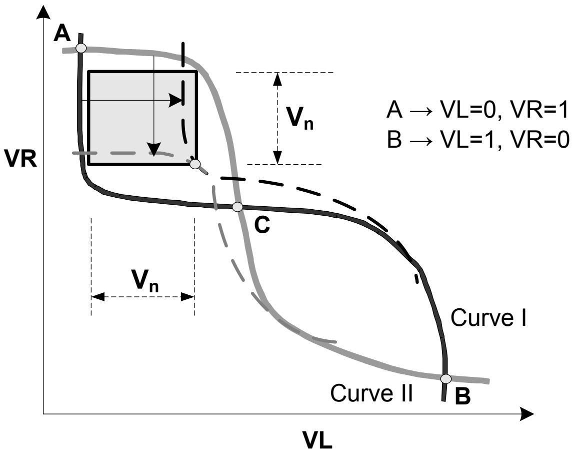


Figure 2.6: SNM measured graphically, as the side of the largest inscribed square within the transfer curves

word-line is turned ON.

Logic “0” can further degrade because of noise. This deterioration in the voltage for logic ’0’ is represented by dashed curve I, which is the horizontally shifted version of the solid curve. The “shift” equals the noise inflicted at node VL. Similarly, the voltage at node VR (logic “1”) can degrade because of noise. This is represented by the shifted dashed curve II, where the downward shift is the noise at node VR.

A certain amount of inflicted noise can shift the curves such that the points A and C coincide, which would force the system to have just one stable state - point B. This is shown in Fig. 2.6. This implies that the bit-cell would flip to state B, if this amount of noise is applied. The noise sources that cause a shift in the solid curves are equivalent to the sides of the inscribed rectangle as indicated in Fig. 2.6. Because the worst-case condition occurs when the noise affects both the nodes simultaneously, it is appropriate to consider a square. Therefore, the SNM can be measured graphically, as the side of the largest inscribed square between the transfer curves. This also implies that the worst-case condition for SNM is when the word line is turned ON, because this degrades logic “0”.

Several ways to model SNM have been proposed [19]-[20]. For this work, SNM is measured by DC simulations. SNM varies with supply voltage, temperature and transistor sizes. SNM is also strongly impacted by the process variations. The SNM can be controlled by the SRAM designer through transistor sizing.

2.2.2 Write Switching Voltage

To write into the bit-cell, one of the bit-lines is pulled to ground. This overwrites logic '1' to logic '0'. The maximum bit-line voltage at which the bit-cell flips (or is written into) is the write switching voltage [21]-[22] or $Vtrip$. The bit-cell should be designed such that the Vtrip is not too high because this can lead to unintended write during the read cycle. At the same time, the Vtrip should not be too low, because driving the precharged bit-line with a huge capacitance, to a voltage closer to ground would take longer and increase the memory write time. Moreover, it may not be possible to pull the bit-line all the way to ground, because the bit-lines of large memories can be a few hundreds of micrometers long. This increases the IR drop on the bit-line, and the resultant voltage at the bit-line, within the bit-cell, may always be a few millivolts above the ground. Therefore, the bit-cell design should provide a Vtrip, which ensures a successful, intended and timely write operation. Again, as in the case of SNM, the designer can control the Vtrip by transistor sizing. Fig. 2.7 demonstrates the measurement scheme for Vtrip.

2.2.3 Read Saturation Current

Fig. 2.8 depicts the half bit-cell circuit during the read operation. The bit-line capacitance discharges though the series access and driver transistors, to develop a bit-line differential, which is amplified by the sense amplifier. Therefore, the memory read out time is strongly influenced by the discharge time of the bit-line. The bit-line discharge time can be expressed as follows:

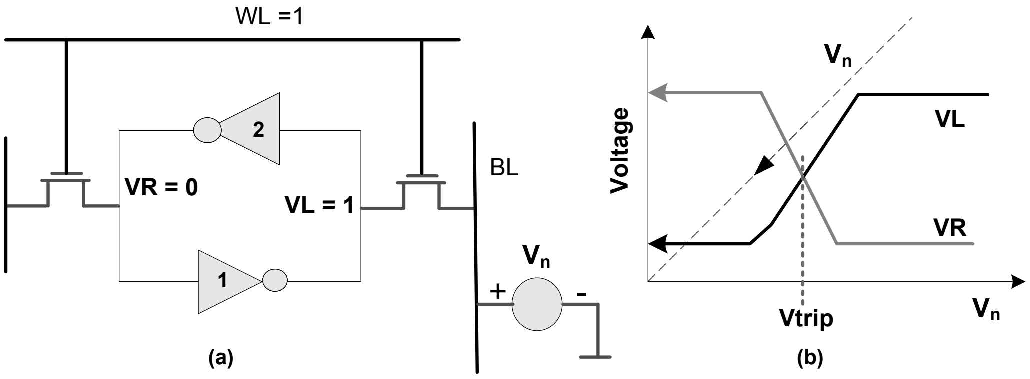


Figure 2.7: (a) Vtrip measurement (b) DC simulation waveforms

$$T_{discharge} = \int \left(\frac{C_{BL}}{I_{read}} \right) dV_{BL}. \tag{2.1}$$

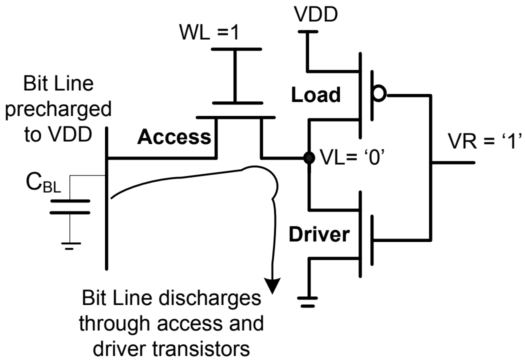


Figure 2.8: Read operation

In equation (2.1), I_{read} is the read current in the driver-access stack. A larger I_{read} can lower the bit-line discharge time - $T_{discharge}$. Since node VL rises to a few hundred millivolts during read, the driver transistor operates in the linear region. Because the bit-line is made to discharge only about 60-200mV, the drain to source voltage of the access transistor remains more or less higher than or equal to its gate overdrive. Therefore, the access transistor operates in the saturation region. Neglecting the channel length modulation effect, the read current through the saturated access transistor can be assumed to remain constant during the entire discharge time. This current is used as a reliable metric for the memory read performance [21]-[23].

2.2.4 Leakage

Leakage is the main cause of power dissipation in the SRAM due to the lower switching activity per bit-cell. Fig. 2.9 shows the paths of two major leakage components - subthreshold leakage and gate leakage. There are other sources of leakage as well, such as the junction leakage.

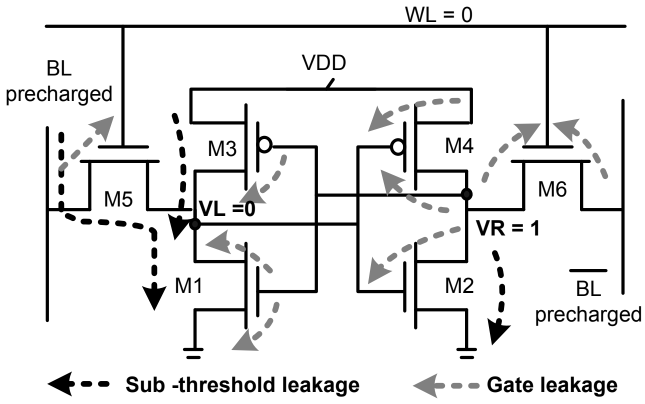


Figure 2.9: Major leakage paths

Fig. 2.10 demonstrates that the entire array except the accessed word, leaks during a normal memory operation. The architecture level leakage reduction techniques such as applying a diode drop in the array supply voltage [24] can only be used in the retention modes , when no read or write is being performed. With as many as 1MB bit-cells in the array, the cumulative array leakage in the read or write modes, can be very high. Therefore, intrinsic bit-cell leakage is an important

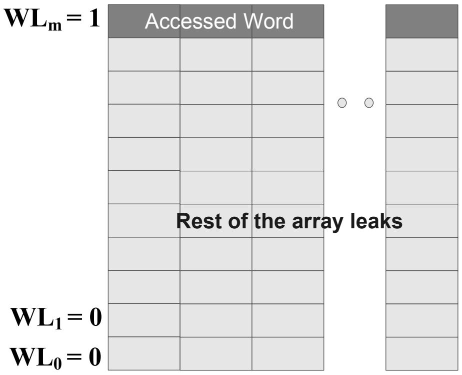


Figure 2.10: Array leakage

metric for the bit-cell design.

It has been emphasized earlier that the bit-cell area is very important from the economic perspective. For a good SNM, the driver transistor should be sized stronger than the access transistor. Because of area concerns, the designer cannot size up the driver transistor too much. The alternative is to reduce the strength of the access transistor. However, the access transistor cannot be made too small since this would degrade the read current. Additionally, the access transistor should be reasonably strong to enable a successful write operation. Similarly, the strength of the load transistor can be reduced to improve the Vtrip, but a very weak load deteriorates the SNM, although the impact is small. The lengths of the driver and the access transistors can be reduced to improve the read performance, but this adversely impacts the leakage, which has become a serious concern these days. Therefore, even for a deterministic design, it is difficult to choose the optimal sizes of the bit-cell transistors, such that all the design metrics meet specifications. The design problem is further compounded by process variations, because of which the design metrics vary from their respective target values. Therefore, statistical bit-cell design is imperative to achieve an optimal, high-yielding design. The next section discusses variability.

2.3 Variability

If a particular performance trait, say a propagation delay, of a population of VLSI circuits (e.g. 1000 samples of a delay chain, with exactly the same layouts, and intended delays) is sampled, a distribution of propagation delays is likely to emerge. The propagation delays are not exactly the same, because of inherent fluctuations in the manufacturing process or "variability". The measurable effect of variability may be a substantial deviation of the circuit behavior from the expected or nominal response. Therefore, only those samples, whose propagation delay is less than the maximum delay specification, can be termed as "acceptable". In this work, "yield" is defined as the ratio of the chips that are "acceptable" (i.e., all the performance traits satisfy their respective specifications) to the total chips that are manufactured. Design for manufacturability, thus, involves choosing a nominal design so that the vast majority of the fabricated circuits (e.g. 99%) would meet the maximum or minimum acceptable specifications for circuit performance traits, while keeping the area overhead minimal.

The next few sections discuss various sources of variability, the impact of variability on the transistor metrics and the modeling of variability. There are multiple criteria, which can be used to classify and understand variability. Variability can be "temporal" or "spatial" in nature. Furthermore, "temporal" variability can be reversible or irreversible. Spatial variation occurs between wafers, between chips, between circuits and between devices.

2.3.1 Temporal variation

Dynamic or time dependent delay and/or power variability in CMOS devices is termed as "temporal". It can occur because of changes in the operating environment [25], that is, the supply voltage fluctuations and temperature variations.

Temporal variability can also get induced by use and aging effects. Several examples of temporal sources of variability can be observed. Additional delay is needed to discharge the residual charge trapped in capacitance between devices in NAND gate stacks. Similarly, self heating (device heating caused by extended periods of high device current) and silicon-on-insulator history effect are examples of application/use dependent sources of temporal variability. Aging related sources of temporal variability are negative bias temperature instability (NBTI), hot electron effects, time-dependent dielectric breakdown (TDDB) and electromigration. NBTI affecting PMOS and hot electron effects impacting NMOS, both elevate device thresholds over a period of time, degrading device and circuit performance [26]-[27]. Because of high current densities over a prolonged interval of time, electromigration results in a slow physical displacement of metal from one part to the other, which severely degrades the metal width and hence the conductivity of the interconnect [28]. TDDB can occur because of prolonged application of a high voltage across the oxide layer, causing a 'weak' spot within it which allows the flow of current. This current flow, which is basically due to the loss of dielectric isolation at that spot, causes localized heating, which induces the flow of a larger current. A vicious cycle of increasing current flow and localized heating ensues, eventually causing a meltdown of the silicon, dielectric, and other materials at the 'hot spot'. This meltdown creates a short circuit between the layers supposedly isolated by the oxide.

On-die hot spots (regions of excessive local heating because of high power dissipation) [29] and activity factor (related to frequency) are other sources of temporal variability. Of the examples mentioned above, NBTI, hot electron effects and electromigration cause irreversible change in the device/interconnect parameters. The impact of self-heating, activity factor and on-die hot spots can be reversed.

2.3.2 Spatial variation

Spatial variation refers to lateral (planar) and vertical differences from intended polygon dimensions and film thicknesses that set in between devices, circuits, wafers and lots during the lifetime of a particular fabrication system [30]. But once the fabrication process is complete, the spatial sources of variation do not change with time or use. For example, the fabricated channel geometry of similar devices can differ across the chip, but for a particular device, the channel geometry would not change with time. Spatial variation can be broadly categorized into inter-die and intra-die variation.

Inter-Die variation

Die to die, wafer to wafer and lot to lot variation , all are clubbed together as inter-die variation. The inter-die variation in a parameter, say threshold voltage or V_{th}, modifies the V_{th} of all the transistors in a die in the same direction, i.e., the threshold voltage of all the transistors in the die, either increases or decreases. This shifts the mean chip threshold voltage, because of which, different chips acquire a different mean threshold voltage. However, this does not cause a mismatch between different transistors on the same die. The inter-die variations are generally assumed to have a simple distribution such as gaussian, with a given variance. These variations may have systematic trends across dies, and can be predicted if the specific orientation and location on the wafer for the die are known. However, the circuits need to run for all the dies, irrespective of their placement on the wafer. Moreover, information such as die position is not available at the design time, and therefore, impact of inter-die variations on process parameters must be captured by using random variables. This is usually done by using corner models [31].

Inter-die variations can occur because of by-wafer and by-reticle process steps. By-wafer processing steps that cause inter-die variation include (a) rapid thermal annealing, when temperature gradients appear across the wafer (b) photoresist development and (c) etching. By-reticle, the photolithography process contributes to variability if the focus changes as the mask is stepped across the wafer. Focus variation can be caused by aberrations of the lens system and/or by wafer nonplanarity.

Intra-Die variation

The intra-die or within-die component of variations can shift the process parameters of transistors at different locations, within the same die, in different directions [31]-[32]. For instance, the threshold voltages of some transistors can increase whereas those of some others can reduce. Within-die variability can be systematic, meaning that there is a well-understood relationship between the placement or layouts of devices and the resulting parameter values. For example, the channel length of transistors in close proximity can be highly correlated. Within-die variability between transistors can also be totally random , e.g. the variation in the threshold voltage of transistors because of the random variations in the number and location of the dopant atoms in the channel region. The systematic intra-die variations do not result in large differences between two transistors that are in close spatial proximity, but the random component of the intra-die variation can result in a significant mismatch between the neighboring transistors in a device.

2.3.3 Process Parameters

All the spatial sources of variability -inter die and intra die, manifest as process variations in the device and interconnect parameters. Some of these parameters are geometrical, while others are statistical. Variations in the geometrical parameters are usually caused by extrinsic sources, whereas the statistical parameters vary because of intrinsic reasons.

Geometrical process parameters

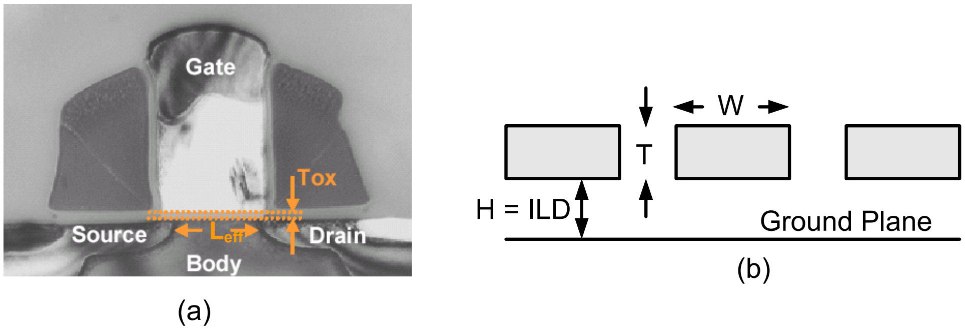


Figure 2.11: (a) Cross section showing transistor geometry (b) Cross section showing interconnect geometry

Extrinsic variability is due to unintentional shifts in the contemporary process conditions, it is typically not associated with the fundamental atomistic problems, but rather with the operating dynamics of the fabricator [33]-[36]. Device and interconnect parameters which are subject to extrinsic sources are displayed in Fig. 2.11 (a) and (b). These are the device length, width and oxide thickness; and interconnect width, thickness and inter-layer dielectric thickness. The various causes of variability in the transistor dimensions are sub-wavelength lithography, proximity effects and lens aberrations. In the sub-wavelength lithography, the minimum feature dimensions and spacings decrease below the wavelength of the light source. Pattern fidelity degrades markedly in this regime, leading to the use of compensation mechanisms, such as optimal proximity correction and phase shifting masks. However, because of these compensation techniques, the layout polygon geometries in the polygon layout tool are no longer consistent with the mask layout geometries, which in turn are no longer consistent with the actual fabricated geometries. Line-end shortening, corner rounding, local context dependent linewidth variations are all fundamental consequences of subwavelength lithography [33]. The proximity effect causes the linewidths in the dense areas to be different than the linewidths in the isolated areas. This is caused by variations in the light intensity during exposure of the photoresist, resulting from the presence of neighboring features. Variations in the interconnect are largely caused by chemical mechanical polishing and lithography [34].

Of all the geometrical parameters, the circuit designer can only control the length and width of the device and the interconnect. It is too tedious to model all of the above mentioned effects accurately- the corner rounding, dog bone effect (the channel length is smaller at mid-width), etc. All the relevant effects can be assumed to have an overall impact on the effective length and width of the transistor channel and the interconnect. These variations are often expressed as a fraction - $3\sigma/\mu$, where μ and σ, are the mean and the standard deviation of the process parameter, which has a gaussian distribution [31]. Therefore, $\pm 3\sigma$ is considered to be the spread of the design parameter around it's mean value. This concept is used to model geometrical parameter variations in this dissertation.

Statistical process parameters

Intrinsic variations are caused by atomic level differences between devices that occur even though the devices may have exactly identical layout geometries and environment [36]. These stochastic or statistical differences appear in the dopant profiles, film thickness variations and line edge roughness. These result in intrinsic variations, mainly in the device threshold voltage. This is demonstrated in Fig. 2.12, which plots the V_{th} of 3500 identical n-MOSFETs laid out in a compact array. Because of close proximity, there is a high spatial correlation between devices, hence there is no systematic width or length differences between the FETs. The wide threshold voltage distribution is because of intrinsic sources of variability such as the random dopant fluctuations (RDF), line edge roughness and intrinsic oxide thickness variations.

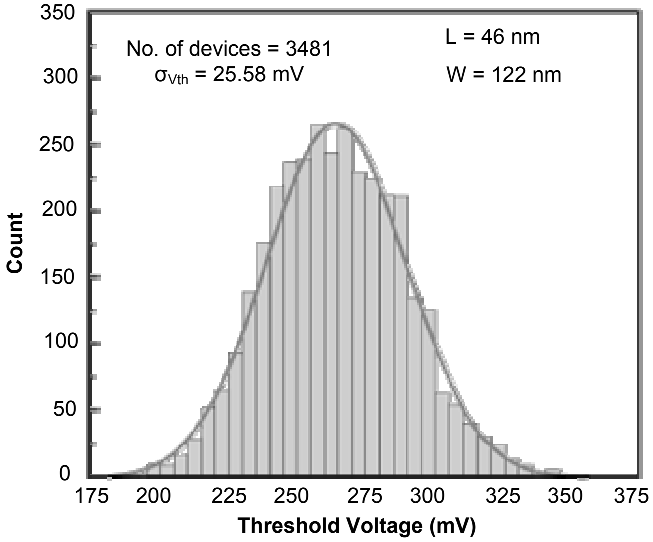


Figure 2.12: Threshold voltage histogram of the transistors in the 90nm technology [36]

However, the most significant source of intrinsic V_{th} variations is the RDF [43]-[44], which is the focus of this work. In the nanometer regime, due to the small channel area, the dopant distribution in the channel acquires a discrete character. For example, consider a uniformly doped NMOS, where $L_{gate} = 45nm$, $W = 3L_{gate}$, impurity density $N_{ch} = 10^{18}cm^{-3}$ and $W_{dep} = 35nm$. The average number of acceptor atoms $N = N_{ch}L_{gate}WW_{dep}$ is approximately 200 [37]. Therefore, only a few ionized acceptors in the body of the transistor are responsible for setting the threshold voltage. It is also noteworthy that the dopant implant and anneal process results in the placement of a random number of dopant atoms in the channel and in the random positioning of these atoms. Fig. 2.13 illustrates a 3D perspective of the dopants in the source, drain and channel region of the transistor [38] . As shown, the source and the drain doping is quite dense, but the channel doping is sparse and subject to statistical variation. The acceptor dopants in the channel are subject to Poisson statistics, which can be represented by a normal distribution with a standard deviation of $N^{1/2}$. Therefore, the percentage variation of N in-

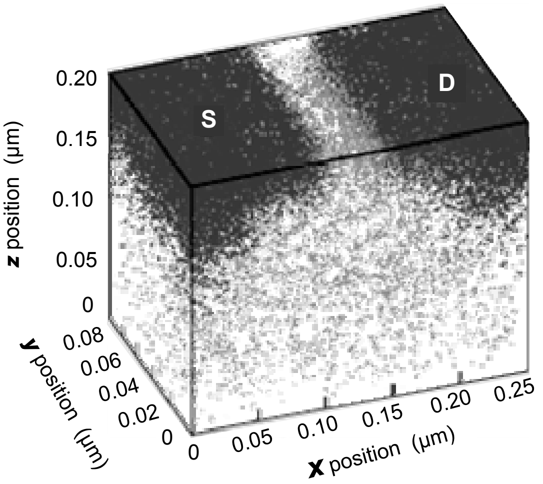


Figure 2.13: Randomly placed dopants in a 50nm channel length MOSFET [38].

creases as the devices are scaled. That, this could cause significant variations in the device threshold, was first realized by Keyes [39] in 1975, who presented a model to estimate the threshold voltage variation due to RDF. The intrinsic V_{th} distribution was demonstrated to be gaussian in nature.

The random dopant fluctuation (RDF) induced variability in the device threshold has been a subject of vigorous research [37]-[45]. Researchers have assumed uniform or non-uniform, 2-D or 3-D dopant distribution profiles to derive analytical expressions for the σ_{Vth} - standard deviation of the threshold voltage distribution due to RDF. The impact of V_{th} variations on the device current and leakage is investigated using these models. Most of the analytical models for σ_{Vth} are of the following form - $t_{ox}\sqrt{\frac{N_{ch}}{WL_{eff}}}$. This expression implies that the σ_{Vth} is inversely proportional to the channel area. This is turn indicates that as the channel area reduces because of technology scaling, the RDF induced V_{th} variation would become more serious [44]. Silicon test structures have been fabricated to verify the modeling of intrinsic V_{th} variations [46]-[47]. It has been confirmed that the primary component of the σ_{Vth} is inversely proportional to the channel area.

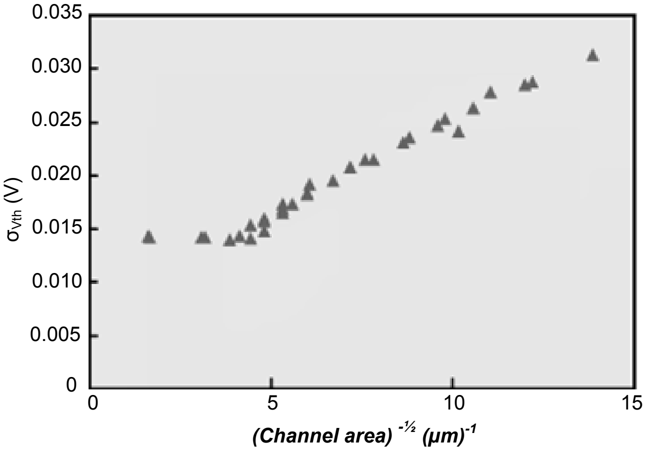


Figure 2.14: σ_{Vth} vs. $(channelarea)^{-1/2}$ for nMOS populations in 90nm technology. Each point is a different length × width geometry [36].

An example is demonstrated [36] in Fig. 2.14. Here, there are 32 populations of 1500 identical FETs in a compact array. Each population has a different length × width combination for the transistor channel. The σ_{Vth} of each of the 32 distributions has been extracted and plotted to show the dependence on channel area. Many analytical expressions have been derived for σ_{Vth}. Comparing with Fig. 2.14, it can be seen that the inverse proportionality with channel area is indeed realized in the data.

To summarise, variability can be broadly grouped into (a) environmental parameters-related to supply voltage, temperature, specific application, activity factor, aging, etc (b) geometrical parameters - physical device and interconnect lateral and vertical dimensions influenced by various sources of process variations (c) statistical parameters - stochastic variations in the device threshold voltage because of atomistic level random variations such as the RDF, line edge roughness and oxide thickness variations. Each of these parameters is considered in the design of the SRAM array. Of the environmental factors, voltage and temperature have been considered in this work. Dynamic factors such as the activity factor are small for the SRAM cells.

For the geometrical parameters, as mentioned above, only the transistor length and width - the lateral dimensions can be controlled by the designer. These are also the most significant geometrical parameters. RDF - the major source of statistical variations in the threshold voltage, is considered in this work. The detailed modeling is explained in the next chapter. The next section describes the impact of variability on the SRAM.

2.4 Impact of Variability on SRAM

Fig. 2.14 emphasizes the relationship between the σ_{Vth} and the transistor size. Smaller the transistor, larger is the standard deviation of the intrinsic threshold voltage variations due to RDF. Therefore, SRAM bit-cells, which are designed with the smallest possible transistors (to economize array area), are especially susceptible to large threshold voltage deviations. The impact of V_{th} fluctuations due to RDF on the SRAM was first described in [48], for an SRAM cell with resistor load. It is shown that the V_{th} mismatch distribution between two matched pair of devices has a measured standard deviation of 17.3mV. For an SRAM array with 4 million matched pairs, two pairs can have a mismatch of $4.9\sigma_{Vth} = 85mV$. The measured V_{th} mismatches on a bit-cell, which failed at 3.4V were as high as 90mV and 60mV for the driver and transfer devices respectively (for 0.35μ technology). Such a high magnitude of threshold voltage variations in the SRAM bit-cell transistors results in a serious deviation of the bit-cell performance characteristics such as the SNM, Vtrip, read current and leakage from their respective expected values.

It is important to understand that the individual logic circuits are also impacted by the V_{th} variations due to RDF, which causes variability in the drive currents and propagation delays. However, this variation tends to average out over a chain of logic circuits. This is not the case with the SRAM array, where each bit-cell can be accessed independently, and therefore, the performance characteristics of all the bit-cells must be within desirable limits. Therefore, with shrinking transistor geometries and increasing σ_{Vth}, it is becoming increasingly important to consider the impact of variability on the SRAM characteristics, during the design phase.

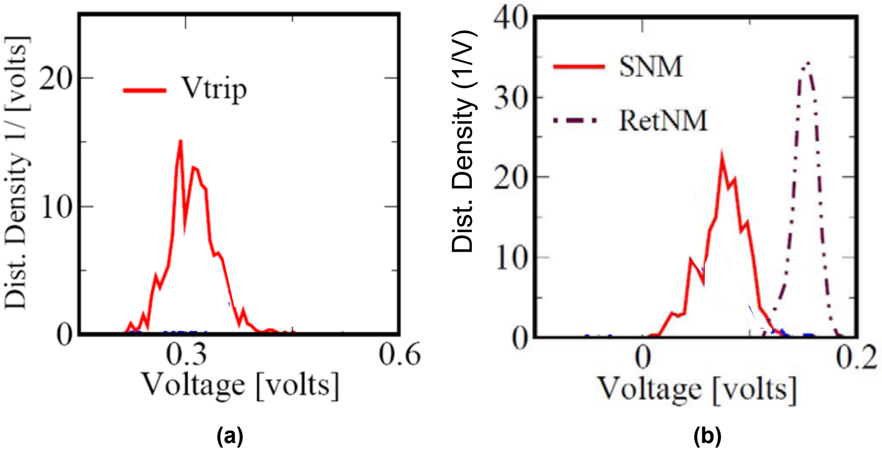


Figure 2.15: (a) Measured Vtrip distribution (b) Measured SNM and RNM distributions [49].

In [49], the bit-cell characteristics of 512 identical SRAM bit-cells, in a test structure designed in a 65nm technology, are measured. Due to intrinsic V_{th} fluctuations, the bit-cell characteristics such as the SNM and the Vtrip of the 512 cells, vary significantly and their distributions are observed to be normal. The distributions of SNM and Vtrip from [49] are reproduced in Fig. 2.15. RNM or the retention noise margin, is the noise margin when the bit-cell is not being accessed for the read operation. In such a scenario, the word line is in the OFF state. Therefore, the nominal RNM is higher than the nominal SNM. Fig. 2.15 indicates a fraction of the SNM distribution to the left of the origin (SNM ≤ 0), which corresponds to the failing bit-cells. Therefore, the bit-cell should be designed, so as to shift the SNM distribution towards the right, to decrease the number of failures.

The impact of variability on the bit-cell characteristics has been studied extensively [20],[50]-[54]. Similar to SNM and Vtrip, the read current also acquires a normal distribution. The leakage distribution is lognormal, because of exponential dependence of the sub-threshold leakage on the threshold voltage. The simulation results depicting some of these distributions are presented in the next chapter in the relevant sections. Armed with the understanding of the primary sources of variability (e.g. transistor width, length and V_{th}), and how SRAM characteristics depend on these parameters (e.g. SNM improves with increasing driver width), some researchers have developed models to predict the impact of variability on the SRAM design metrics. However, these models are analytical in nature, and can only model the variability in a particular chosen design. These works do not provide an optimization framework, which would automatically provide an optimal, high yielding design by using the variability information. The next chapter proposes the novel statistical bit-cell design method, which provides such a framework.

Chapter 3

Statistical Design of the 6T SRAM Bit Cell

Technology scaling has a two-fold impact on the SRAM design. First, increasing σ_{Vth} of the scaling SRAM transistors, increases the variance of the distributions of the design metrics such as the SNM, vtrip and read current. Secondly, growing memory density at each successive technology generation requires that the bit-cell be designed to tolerate a larger number of sigma variations (e.g., 4σ to 5σ), in the design characteristics, to ensure a satisfactory memory yield. The extent of variations in the bit-cell design metrics, in large measure, is a function of the bit-cell transistor sizes. Therefore, to meet the specifications of all the bit-cell design metrics for all the fabricated cells, amidst variability (yield), and within the minimum possible bit-cell area; the widths and lengths of the bit-cell transistors must be chosen optimally. This requires consideration of the impact of variability up front, that is, during the design phase.

In this chapter, a statistical method to design an SRAM bit-cell is proposed. Previous literature in this field is reviewed and compared with the proposed method. Subsequently, the bit-cell design problem is formulated by modeling the variability and a yield maximization approach is presented. Optimal bit-cell designs in the 45nm technology are derived with the proposed method. The results and observations are discussed and the optimality of one of the designs is verified by Monte Carlo simulations.

3.1 Current Industrial Design Practice

Typically, SRAM is a pilot product in a new technology generation, and therefore the SRAM design proceeds in parallel with the process development. It involves evaluation of several bit-cell architectures and layout topologies for process-layout interactions. It also requires choice of SRAM specific physical design rules and assessment of their robustness to minimize the occurrence of hard failures such as opens and shorts. At this level, process and technology developers play a critical role in the bit-cell design. For the circuit designer, the bit-cell design entails optimal selection of the transistor sizes to avoid failures such as the destructive read, write failure, access failure and excessive leakage, which can occur due to variations in the transistor parameters. Such failures are called parametric failures. The circuit designer interacts with the process developers on one hand, to ensure a highly manufacturable design in the minimum possible area; and with the system developers on the other hand, to consider the SRAM environment, SRAM array size, supply voltage and performance specifications. Therefore, the task of the circuit designer, i.e., the selection of the optimal sizes for the bit-cell transistors, is quite significant.

Along with the design of the bit-cell, the complete SRAM design also involves the design of the periphery and control logic, e.g. the design of the sensing strategy, circuits for tracking and self-timing, for write, redundancy, test collar, precharge, address decode, etc. In this work, a specific aspect of the SRAM design, which is the bit-cell design, is focussed upon.

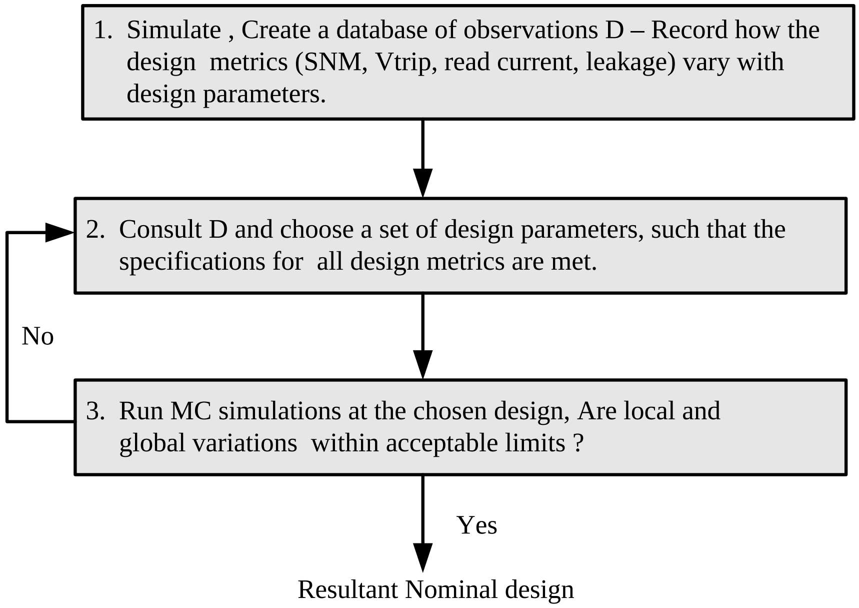

1. Simulate , Create a database of observations D – Record how the design metrics (SNM, Vtrip, read current, leakage) vary with design parameters.
2. Consult D and choose a set of design parameters, such that the specifications for all design metrics are met.
No
3. Run MC simulations at the chosen design, Are local and global variations within acceptable limits ?
Yes
Resultant Nominal design

Figure 3.1: Current Bit-cell design method

The current industrial approach to determine the sizes of the bit-cell transistors is outlined in the flowchart in Fig.3.1. Step 1 involves preparation of an exhaustive database to record the variations of the design metrics (SNM, vtrip, read current and leakage) with each of the design parameters. The design parameters are the width and length of the bit-cell transistors. For example, the database can record the SNM variation with the driver width at different combinations of the width and length of the access transistor. This database is used as a reference throughout the design procedure. Step 2 involves extensive and careful consultation of the database created in Step 1, to choose a set of design parameters, which would satisfy the specifications for all the design metrics. This is difficult even for a deterministic design, because the design metrics - SNM and Vtrip, read current and leakage, are conflicting in nature. For example, increasing read current would increase leakage also. Therefore, an optimal selection of the design parameters is required. Moreover, at this point, the designer can only observe the nominal values of the design metrics. It still needs to be verified that for the chosen design parameters, the local and the global variations in all the four design metrics are within the desirable limits.

In Step 3, Monte Carlo simulations are run at the chosen sizes to analyze the impact of variability on the design metrics. If the spread of any of the design metrics is too large, the transistor sizes should be increased to reduce the intrinsic threshold voltage variations due to RDF. This increases area and can have an adverse impact on some other design metric. Therefore, the database, generated by Step 1, is consulted to judiciously choose a new set of transistor sizes. The designer loops on Steps 2 and 3, and is aided by the database created in Step 1, to derive the final bit-cell transistor sizes. The above procedure is iterative, time consuming and requires manual intervention. In addition, the chosen design need not be optimal. In fact, it can be an over-design with larger area.

However, little work has been carried out that incorporates up front, the statistical information about the variations in the performance targets, into the design in a systematic way. Statistical analysis and Monte Carlo simulations are performed, but the results are not applied in a systematic manner, to arrive at an optimal design point. Some of the proposed approaches for robust SRAM design are discussed below.

3.2 Related Work

Several researchers have proposed ways to improve the immunity of the SRAM to process variations. Most of these focus on improving the SNM and write margin yield. One approach is to provide additional circuitry to reduce the swing on the word-line during the read operation to improve SNM, and to alter the memory power supply during the write operation for better write switching voltage [55]. Separate array and logic supply voltages have been proposed to enable better write margins [56]. This also helps to reduce leakage power in different operating modes. However, the disadvantage is the extra cost and complexity associated with adding an extra supply, such as the use of level shifters and isolation circuits.

Another proposal to improve the SRAM yield is to dynamically detect faulty cells and replace them by adaptively resizing the cache memory [57]. Additional column address bits are added to the tag to modify the mapping scheme of the cache. This architecture downsizes the cache to avoid faulty blocks, therefore it increases the cache miss rate, affecting the processor performance. Another proposal is to use body bias for NMOS and well bias for PMOS to shift the threshold voltage higher or lower based on the inter-die process corner [58]. Leakage and ring-oscillator delay monitoring is used to determine the inter-die process corner. A circuit to select the proper body bias to minimize the impact of Vt shift is activated to apply the forward or reverse body bias, as the case may be. This approach works for global variations, and not for within-die variations.

A bit-cell level optimization approach is described in [59]. Driveability ratio - the ratio of currents of the access and load transistors, is introduced as a parameter to relate the write margin to the transistor size. It is shown that the designers can employ driveability ratio variation along with write assist [60] circuits to trade-off between SNM and write margin. However, the selection of the transistor sizes is still based on observation of exhaustive data. The SRAM design is optimized at the process level in [61]-[62]. The methodology involves choice of critical physical design rules in conjunction with judicious application of optimal proximity correction, comparative analysis of different architectures and metal routing strategies, process understanding and continuous monitoring of electrical test data as feedback for process improvement.

The above proposals are either architecture level modifications or post-silicon tuning techniques to improve parametric yield. Some of them are proposals at the bit-cell level, but they only present insights that should be kept in mind during the size selection process. Some proposals are about layout and process optimization. They do not address the design problem confronted by the circuit designer, which is to provide a system for the statistical design of the bit-cell.

Mukhopadhyay et al. [63] have used the concept of failure probability to statistically design the SRAM cell. However, [63] only considers the within-die variation due to RDF. It does not consider the impact of inter-die variation in the transistor dimensions, on the bit-cell design metrics. Secondly, the problem formulation is such that the optimal design would depend on the number of rows and columns in the memory. The transient design metrics like the access time and the write time have been used, which depend on the memory size, circuit capacitances and the slopes of the input signals which vary with the size and the layout of the peripheral drivers and the bit-cell. However, for an embedded SRAM, the bit-cell needs to be designed for a range of memory sizes. Therefore, DC parameters such as the SNM, write trip voltage and read saturation current have been traditionally adopted as design constraints for the bit-cell in embedded SRAM [21]-[23].

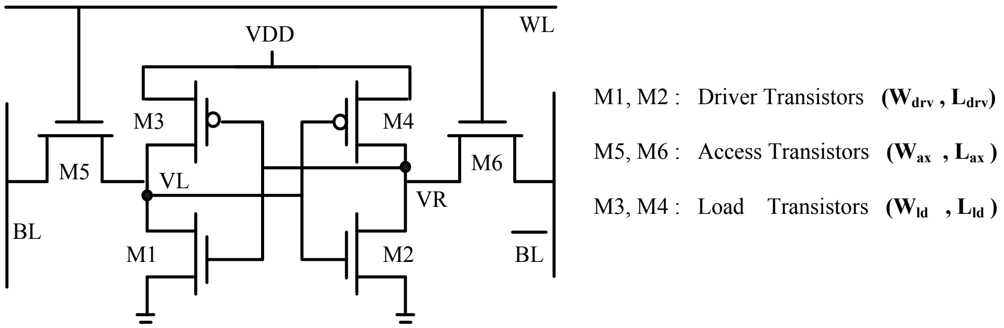
Additionally, the development of semi-analytical models in [63] for the transistor characteristics (such as saturation current and leakage models) not only induces approximation errors, but also has limited usage in the industry, because designers prefer available SPICE models. Also, analytical modeling in [63] for the design metrics such as the write time is not accurate. At the beginning of the write operation, the bit line is assumed to have been completely pulled to ground by the write driver. This is over-simplified as the bit line capacitance can be quite high (0.1pF) for large memories and the word line is usually turned on, well before the bit line completely discharges to ground, to gain cycle time. This can be observed in [60], Fig 3a (signals *wl* and *blt*). Other causes of inaccuracy are the position of the bit-cell in the array (top/bottom, close to word line driver or away from it) and the bit line driver size.

The joint failure probabilities (e.g. read and access failure occurring together) have been ignored in [63] in the formulation of the objective function because of computational complexity and therefore, the obtained design solution need not be optimal. The proposed method in our work formulates the cell failure due to within-die variations in design metrics as constraints. As the solution of the optimization problem requires that all the constraints be satisfied simultaneously, the failure of the bit-cell due to simultaneous occurrence of two or more reasons is also accounted for.

The bit-cell consists of the driver, the access and the load transistors. Considering the width and the length of each of these transistors as a design parameter, the bit-cell design problem lies in a six-dimensional parameter space, defined by the dimensions - W_{drv} ,W_{ax} ,W_{ld} ,L_{drv} ,L_{ax} ,L_{ld}. The statistical bit-cell design method proposed in this dissertation, inscribes a maximum yield box in this six-dimensional space. For a given distribution of the widths and lengths, the method derives the nominal transistor dimensions that provide the maximum immunity to the variability in the transistor dimensions (inter-die) and intrinsic threshold voltage fluctuations due to RDF. The proposed method involves a minimal initial infrastructure in terms of model building and mathematical computations and uses readily available models and tools in the industry for simulation. No analytical modeling for either the transistor characteristics or the design metrics is involved. This reduces approximation and makes the method attractive and practical for industrial usage. Also, the proposed problem formulation imparts the necessary flexibility to tune the design corresponding to the specifications, as demonstrated in Section 3.5. High performance-moderate leakage and low leakage-moderate performance bit-cells in the 45nm CMOS technology are designed and analysed. It is shown that the conventional sizing is no longer sufficient to ensure a high yield for a low leakage bit-cell design. The results are verified by Monte-Carlo simulations to show the optimality of the chosen design.

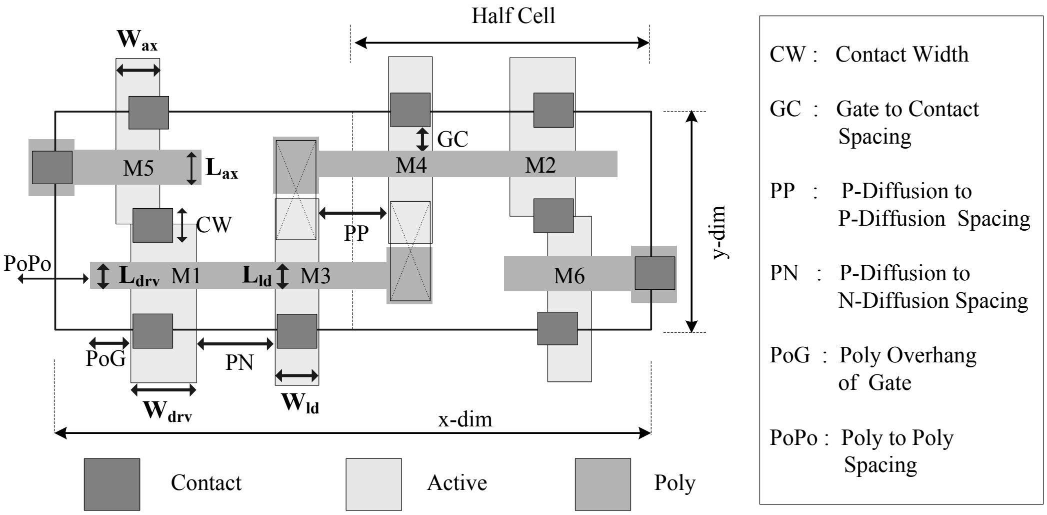
3.3 Preliminaries

3.3.1 Design Metrics



VDD
WL
M3
M4
M5
M6
VL
VR
BL
M1
M2
BL
M1, M2 : Driver Transistors (W_drv , L_drv)
M5, M6 : Access Transistors (W_ax , L_ax)
M3, M4 : Load Transistors (W_ld , L_ld)

Figure 3.2: 6T SRAM bit-cell schematic.



Half Cell
W_ax
M5
L_ax
CW
GC
M4
M2
PP
PoPo
L_drv
M1
L_ld
M3
M6
y-dim
PoG
PN
W_drv
W_ld
x-dim
Contact
Active
Poly
CW : Contact Width
GC : Gate to Contact Spacing
PP : P-Diffusion to P-Diffusion Spacing
PN : P-Diffusion to N-Diffusion Spacing
PoG : Poly Overhang of Gate
PoPo : Poly to Poly Spacing

Figure 3.3: 6T SRAM bit-cell sample layout, from [20].

The schematic of a typical 6T SRAM bit-cell is drawn again in Fig.3.2, for convenience. Transistors M1 and M2 are referred to as drivers, M3 and M4 are load transistors and M5 and M6 denote the access transistors. The output nodes of inverters 1 (M1 and M3) and 2 (M2 and M4) are called VL and VR, respectively.

For subsequent discussions, it is assumed that VL is at logic “0” and VR is at logic “1”.

Four of the bit-cell designs metrics - SNM, Vtrip, read current and bit-cell leakage, have been discussed in chapter 2. These are measured by DC simulation. Another important metric is the bit-cell area. The bit-cell area depends on the chosen layout topology. The bit-cell can have different types of layout [61],[64]-[65]. Researchers at IBM [67], Intel [68] and TI [69], have proposed an approach of Restrictive Design Rules (RDRs) such as single-orientation poly-silicon gates , resulting in layout geometries that are more regular with enhanced manufacturability and support a more exhaustive checking of the algorithms for resolution enhancement techniques [66]. Some of these have already been adopted as best practices for memory [69]. The layout topology in 45nm technology from [65] is used in this work and reproduced in Fig.3.3. The corresponding tight design rules are also mentioned in [65]. The x and y dimensions of the bit-cell layout - x_{dim} and y_{dim}, respectively, are calculated as a function of the layout rules as shown below. Because the bit-cell is symmetrical, the x_{dim} is the twice of the x dimension of the half-cell, which can be determined in two ways - $x1$ and $x2$. The greater of $x1$ and $x2$ determines the x_{dim}. Similarly, the y_{dim} is determined by the comparing the calculated values of $y1$ and $y2$.

$$\begin{aligned}
&Area = x_{dim} \times y_{dim},\; x_{dim} = 2\,max(x1, x2),\; y_{dim} = max(y1, y2),\\
x1 =&(\frac{1}{2})(PP) + W_{ld} + PN + W_{drv} + PoG + (\frac{1}{2})(PoPo),\\
x2 =&(\frac{1}{2})(PP) + W_{ld} + PN + W_{ax} + PoG + (\frac{1}{2})(CW),\\
y1 =&2[(\frac{1}{2})(CW) + 2(GC) + L_{ld}] + CW,\\
y2 =&2[(\frac{1}{2})(CW) + 2(GC)] + L_{drv} + L_{ax} + CW. \qquad (3.1)
\end{aligned}$$

In Fig. 3.3, all diffusion contacts have diffusion layer underneath (not visible), but there is no diffusion layer overhang around the contact. This is because the design rules for the SRAM are scaled beyond those of standard logic-process design rules, to achieve competitive bit-cell area [61]. Several design rules are violated within the cell array. An imaginary layer (e.g. it can be called SRAM_ARR) is drawn on top of the array during layout design, so that the design rule checker tool would identify the array portion of the SRAM and not flag these violations. Post-layout comprehensive lithographic correction strategies are used to ensure a robust bit-cell layout [62]. The rectangular contacts are the coupled contacts, which are used to strap poly and diffusion for cross-coupling without using metal [61]. This enables the use of relatively looser metal 1 pitch.

3.3.2 Preparatory work

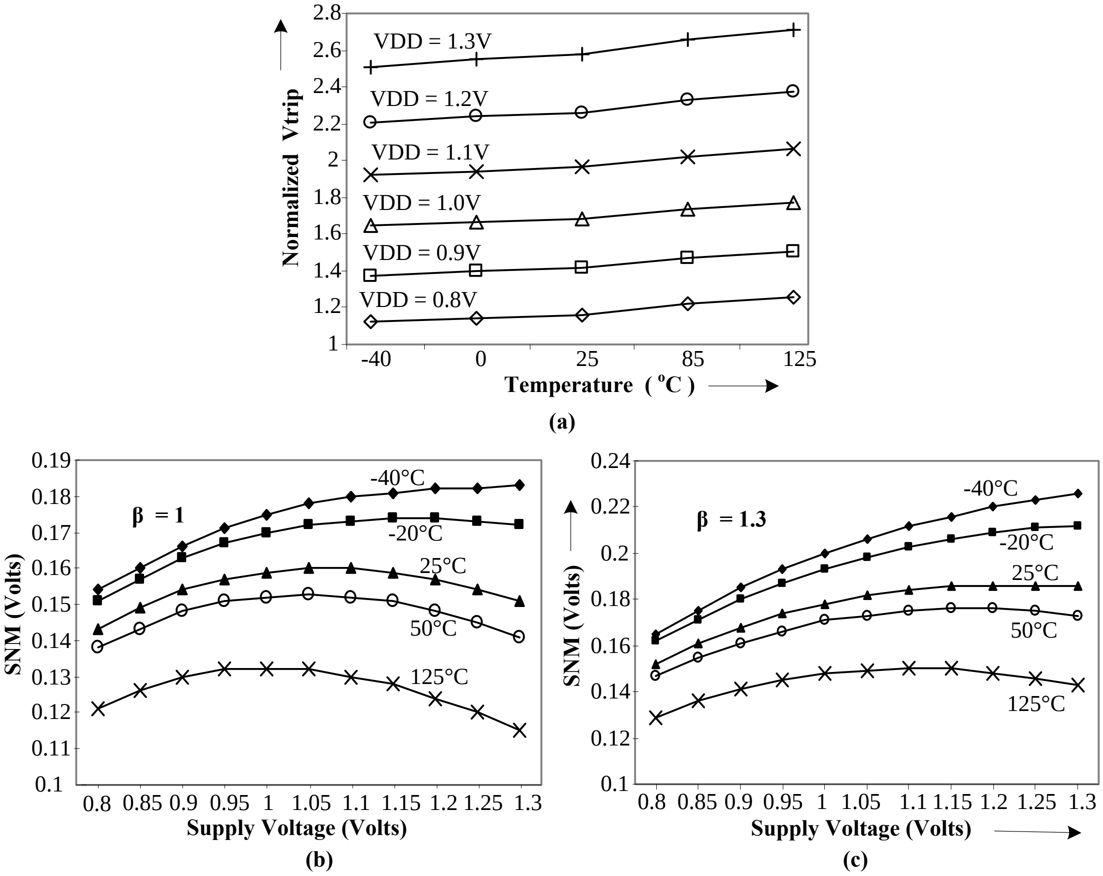
The design metrics (SNM, Vtrip, read speed and leakage) are impacted by the operating parameters (supply voltage and temperature), design parameters (the width and length of the transistors), and statistical parameters (e.g. process parameters such as the threshold voltage). Therefore, the variations in the operating, design and statistical parameters must be considered during the bit-cell design procedure.

Operating Parameters

These are often critical and are accounted for, by evaluating the design metrics at their respective worst-case operating conditions. Table 3.1 documents the worst-case operating conditions (lower or higher than the nominal voltage, low or high temperature) for all the design metrics. The nominal voltage is the applied voltage, but the actual operating voltage available at the SRAM power rails can change due to factors such as the IR drop, operating frequency and the temperature.

Table 3.1: Worst-Case Operating Conditions for Design Metrics

Operating Condition	SNM	Vtrip	Read Current	Leakage
Voltage	L/H	L	Performance Corner	H
Temperature	H	L		H



(a)

(b)

(c)

Figure 3.4: For a 45nm design,(a)Variation of vtrip with supply voltage and temperature (b) Variation of SNM with supply voltage and temperature for $\beta = 1$, and (c) for $\beta = 1.3$.

For example, for general-purpose applications, the possible worst-case for leakage is 10% higher than the nominal voltage, and a high temperature of 85^o C [6]. Similarly, read current simulations should be carried out at the performance corner to meet the timing goal. The performance corner usually consists of lower than the nominal voltage, and a high temperature. The number of performance corners and the voltage and temperature for each of the performance corners is determined by the intended set of applications for the memory. For example, for mobile applications, the operating temperature will be lower than that for server (high-performance) applications. The chip vendor might also choose to check chip timing at multiple performance corners for reliability standards.

The variation of vtrip with supply voltage and temperature is depicted in Fig 3.4 (a), for the 45nm technology (Predictive Technology Models). It exhibits that the vtrip is the lowest , i.e., the worst, for low voltage and low temperature. The worst-case voltage condition for SNM is interesting. At high voltage and high temperature, the SNM begins to degrade, as shown in Fig. 3.4 (b), because the node VR, which stores "1", leaks excessively through M2. This behavior can be arrested, if the cell ratio β $(= (W_{drv}/L_{drv})/(W_{ax}/L_{ax}))$ is increased, as depicted in Fig 3.4 (c). With a higher β, the node VL, which stores "0", remains closer to ground (stronger "0") and the gate voltage of M2 reduces, thereby, shutting it off more effectively. However, since the proposed method explores the entire space of the allowable transistor sizes (i.e., all β values), SNM is simulated at both, the high and low voltages. A nominal supply voltage of $1V$ is assumed for the 45nm technology [6]. Most of the trends in Table 3.1 should remain the same for all technologies.

Design Parameters(Inter-Die) $\{W_{drv}, W_{ax}, W_{ld}, L_{drv}, L_{ax}, L_{ld}\}$

Variations in the transistor widths and lengths are considered to be the main source of inter-die (global or die to die) variations in this work. The inter-die V_{th} variations are accounted for implicitly, because these are predominantly caused by the variations in the gate length [70]. According to the *International Technology Road Map for Semiconductors* [6], the gate dimension variations are assumed to have a 3σ value of $\pm 12\%$ of the physical gate length. With a physical gate length of 25nm in 45nm technology [63],[6], the 3σ variation in the gate dimension is selected as 3nm (a different transistor length can be chosen for the design, but the 3σ variation remains fixed at 3nm).

Statistical parameters (Intra-Die)

Because of the small area of the SRAM bit-cell and proximity of the transistors, the usage of restricted design rules, a highly regular layout and fairly controlled process for the SRAM array fabrication, the impact of the intra-die variations in the channel length and width is small and negligible [63]. Therefore, in this work, intrinsic V_{th} variations due to RDF are considered as the major source of intra-die variations in the design metrics. However, the proposed method can be extended to incorporate other sources (such as intra-die width and length variation, or these can be included as additional components of the intra-die V_{th} variation).

The threshold voltage variations of six transistors are considered to be six independent and un-correlated Gaussian random variables [20], [63]. This assumption is justified, since primarily, the effect of RDF is considered. The placement and the number of dopants in the channel of one transistor depend only on the geometry of that transistor, and are independent of the placement and number of dopant atoms in the channel of any neigbouring transistor. It has been described earlier that the distribution of the V_{th} due to RDF is normal. The standard deviation of the V_{th} distribution (σ_{Vth}) due to RDF is a function of the doping profile, manufacturing process and the transistor geometry. The σ_{Vth} of a minimum sized transistor (σ_{Vth0}) is usually available in the process development kits of vendors and is used as an input parameter in this work. Then, the σ_{Vth} (due to RDF) is related to the transistor size as follows [63],[36]:

$$\sigma_{Vth} = \sigma_{Vth0} \sqrt{\frac{W_{min}\, L_{min}}{W\, L}}. \tag{3.2}$$

The circuit designer can specify only the nominal values of the geometrical transistor layout dimensions (design parameters), and has little control over statistical parameters such as the V_{th} variations due to mismatch. However, as shown by equation 3.2, the choice of design parameters can be used during the design phase, to control the extent of device mismatch.

3.4 Problem Formulation

3.4.1 Intra-die Variations

Variability Modeling

Before modeling the intra-die variations in the design metrics statistically, some mathematical concepts are presented below. Consider $x_1, x_2, .., x_n$ as n independent, un-correlated Gaussian random variables. Assume that the statistical means of these random variables are $\mu_1, \mu_2, .., \mu_n$, respectively. The corresponding standard deviations are $\sigma_1, \sigma_2, .., \sigma_n$, respectively. Now consider a dependent variable y, which is a function of $x_1, x_2, .., x_n$, that is, $y = f(x_1, x_2, .., x_n)$. Then, the mean (μ_y) and the standard deviation (σ_y) of the distribution of y can be estimated using the multi variable Taylor series expansion [71], as:

$$\mu_y = f(\mu_1, \mu_2, .., \mu_n) + \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\partial^2 f(x_1, x_2, .., x_n)}{\partial x_i^2} \right)\bigg|_{\mu_i} \sigma_i^2 \tag{3.3}$$

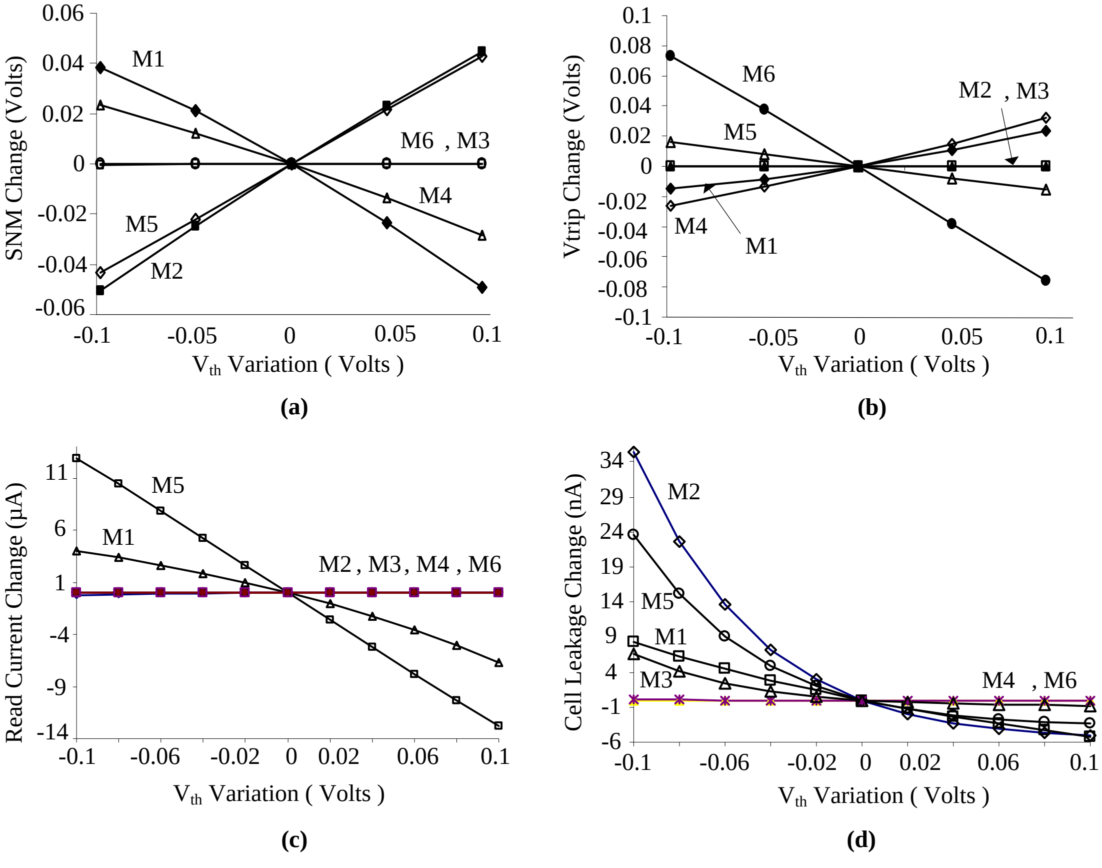
$$\sigma_y^2 = \sum_{i=1}^{n} \sigma_i^2 \left(\left. \frac{\partial f(x_1, x_2, .., x_n)}{\partial x_i} \right|_{\mu_i} \right)^2 \tag{3.4}$$

Equations 3.3 and 3.4 suggest that if the mean and variance of the distributions of $x_1, x_2, .., x_n$ are known, then the statistical mean and variance of the distribution of the variable y, which is a function of $x_1, x_2, .., x_n$, can be estimated. These expressions, derived from the Taylor series expansion, can be used to model any continuous multi-variable function. Because the intra-die variations in the design metrics -SNM, Vtrip, read current and leakage, are a function of the intrinsic V_{th} variations due to RDF, the Taylor series expansion can be used to model the distributions of the design metrics. The V_{th} of six transistors are independent and un-correlated random variables with gaussian distributions (similar to $x_1, ..., x_n$), whose standard deviations can be estimated from equation 3.2. This implies that the statistical mean and variance of the intra-die distribution of the design metrics can be estimated as follows (shown for SNM):

$$\begin{aligned} SNM_{mean} = SNM_0 + \frac{1}{2}\sum_{i=1}^{6}\left(\frac{\partial^2 SNM}{\partial Vth_i^2}\right)\sigma_i^2 = SNM_0 + \frac{1}{2}\Bigg(\left(\frac{\partial^2 SNM}{\partial Vth_1^2} + \frac{\partial^2 SNM}{\partial Vth_2^2}\right)\sigma_{drv}^2 \\ + \left(\frac{\partial^2 SNM}{\partial Vth_3^2} + \frac{\partial^2 SNM}{\partial Vth_4^2}\right)\sigma_{ld}^2 + \left(\frac{\partial^2 SNM}{\partial Vth_5^2} + \frac{\partial^2 SNM}{\partial Vth_6^2}\right)\sigma_{ax}^2\Bigg) \end{aligned} \tag{3.5}$$

$$\begin{aligned} \sigma_{SNM}^2 = \sum_{i=1}^{6}\sigma_i^2\left(\frac{\partial SNM}{\partial Vth_i}\right)^2 = \left\{\left(\frac{\partial SNM}{\partial Vth_1}\right)^2 + \left(\frac{\partial SNM}{\partial Vth_2}\right)^2\right\}\sigma_{drv}^2 \\ + \left\{\left(\frac{\partial SNM}{\partial Vth_3}\right)^2 + \left(\frac{\partial SNM}{\partial Vth_4}\right)^2\right\}\sigma_{ld}^2 + \left\{\left(\frac{\partial SNM}{\partial Vth_5}\right)^2 + \left(\frac{\partial SNM}{\partial Vth_6}\right)^2\right\}\sigma_{ax}^2 \end{aligned} \tag{3.6}$$

The mean and the variance of the distributions of Vtrip, read current and leakage can be estimated in a similar manner. In the above expressions, SNM_0 is the simulated SNM or the nominal SNM at the mean V_{th} values for all the six transistors. Also, the partial derivative terms are computed numerically, by simulation, at the mean V_{th} values. Fig. 3.5 demonstrates the variation of the design metrics (sensitivity) with the variation in V_{th} of each transistor. As can be observed from Fig. 3.5(a),(b) and (c), SNM, Vtrip and read current vary almost linearly with V_{th}. As a result, the second order partial derivative terms, on the RHS. of equation 3.5, are quite small when compared with the nominal value of the SNM, Vtrip or read current (e.g. SNM_0). However, for leakage, this is not the case, because leakage varies non-linearly with the threshold voltage. Therefore, for estimation of the statistical mean of the leakage distribution, the second order partial derivative terms should be calculated carefully.



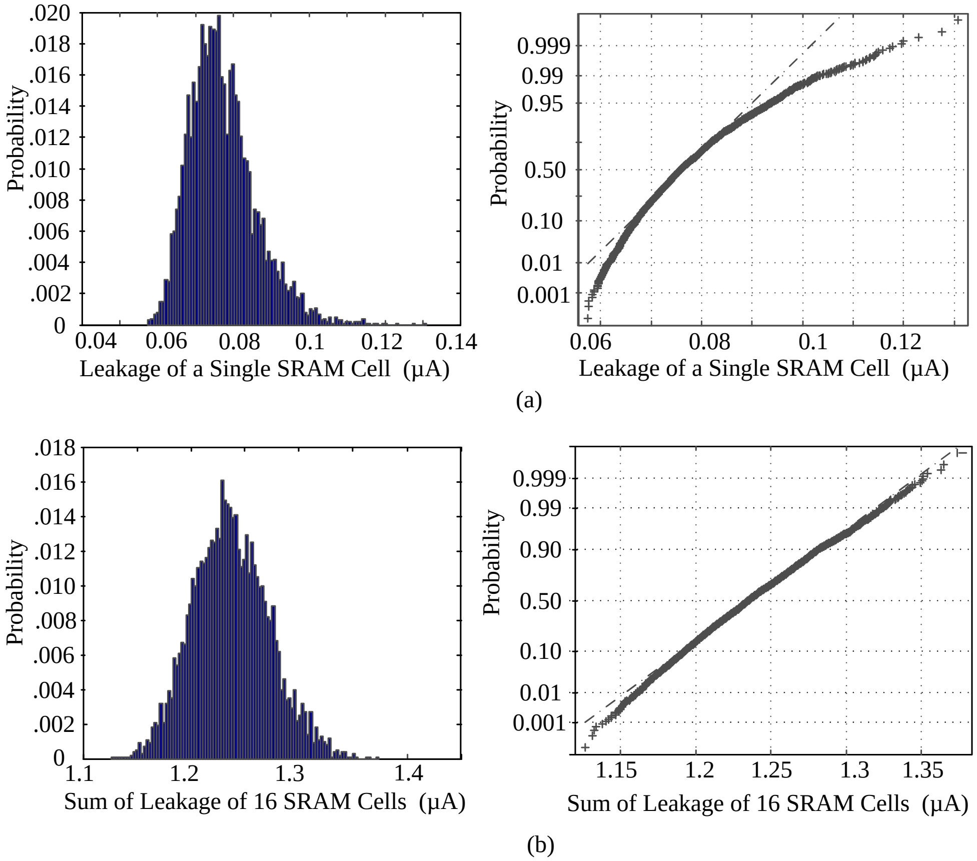
(a)

(b)

(c)

(d)

Figure 3.5: Variation of the design metrics with V_{th} of each transistor.



(a)

(b)

Figure 3.6: Leakage (μA) distribution histogram and normal probability plot for (a)single memory cell and (b)sum of the leakage of 16 cells.

The results of modeling are verified by Monte-Carlo (MC) simulations, where 10000 points are generated with independent and gaussian variation for the V_{th} of six bit-cell transistors. For example, for the leakage MC results (for a 45nm bit-cell design) shown in Fig.3.6 (a), the average leakage and standard deviation are 77.2nA and 8.9nA, respectively. The estimated values (from equations 3.5 and 3.6)are 77.7nA and 8.57nA, respectively. Similarly, from Fig.3.7(c), the SNM average and standard deviation from MC simulations are 131.8mV and 22.1mV, respectively. The corresponding estimated values from modeling are 132.1mV and 21.6mV, respectively. For our purpose of modeling, this level of accuracy is sufficient. If greater accuracy is desired, higher order terms in the Taylor series expansion can be included.

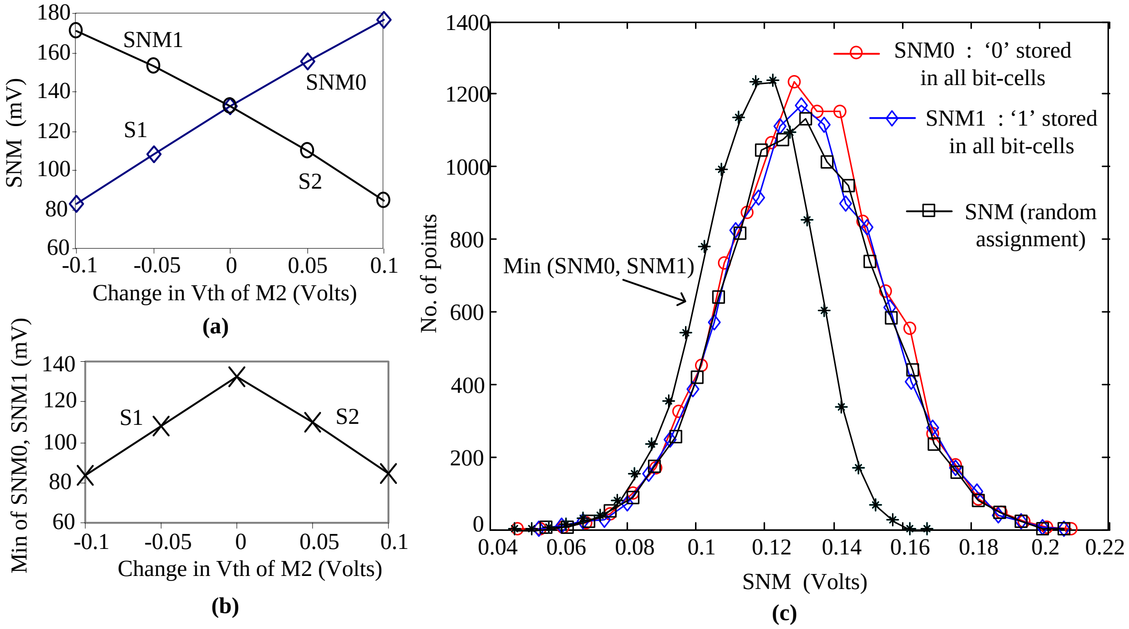


Figure 3.7: (a) Variation of SNM with V_{th} variation in M2 when "0" or "1" is stored in the bit-cell (b) Variation of min(SNM0,SNM1) with variation in V_{th} of M2 (c) Frequency distribution for SNM: SNM0 - when "0" stored in all cells, SNM1 - when "1" stored in all cells, SNM - random assignment and for minimum(SNM0, SNM1)

The plots in Fig. 3.5 are for a given data value stored in the bit-cell (logic "0" at node VL). For instance, Fig. 3.5 (a) shows that the V_{th} of transistors M3 and M6 have no impact on the SNM variation. In other words, $\frac{\partial SNM}{\partial Vth_3}$ and $\frac{\partial SNM}{\partial Vth_6}$ are nearly 0, and do not contribute to σ_{SNM} in equation 3.6. However, if the opposite data value is stored in the bit-cell, $\frac{\partial SNM}{\partial Vth_3}$ and $\frac{\partial SNM}{\partial Vth_6}$ will be non-zero. Therefore, we need to make sure that the sensitivity modeling for σ_{SNM} accounts for both the data values.

To investigate this, consider Fig. 3.7 (a), which depicts the SNM variation with the V_{th} variation in the driver transistor M2. SNM0 represents the case when "0" is stored in the bit-cell. SNM1 represents the case when "1" is stored in the bit-cell. If our concern is to model the V_{th} sensitivity of the worst-case static noise margin of a single fabricated bit-cell, at all instants of time, then we calculate $SNM_{worst-case} = minimum(SNM0, SNM1)$. The variation of $SNM_{worst-case}$ with V_{th} variation of driver transistor M2 is a 'V' shaped curve, as depicted in Fig. 3.7(b). This curve can be obtained by taking the smaller of SNM0 and SNM1 values, from the curves in Fig. 3.7(a), at every V_{th} point. However, we are interested in the statistical distribution of SNM and not the worst-case SNM. At any instant of time, the bit-cell will store either "0" or "1". In other words, the slope - $(\partial SNM)/(\partial Vth_2)$ will be either S1 or S2, but not both. If the V_{th} variation of the second driver transistor M1 is considered, then the SNM0 and SNM1 curves, in Fig. 3.7(a), are interchanged. This is because the bit-cell is symmetrical. For an actual fabricated bit-cell, the two halves have mismatch due to layout differences. But DC simulation treats the two halves as symmetrical, unless deliberate layout mismatch is induced. Therefore, the slope - $(\partial SNM)/(\partial Vth_1)$, will be either S2 or S1, depending on whether "0" or "1" is stored in the bit-cell, respectively.

The above discussion implies that the slopes of SNM vs. V_{th1} and V_{th2} will be numerically interchanged, if the opposite data value (1 instead of 0) is stored in the bit-cell, but the overall coefficient of σ_{drv}^2 on the RHS. of equation 3.6 - $((\partial SNM)/(\partial Vth_1))^2 + ((\partial SNM)/(\partial Vth_2))^2$, should remain the same. Applying the same argument to the coefficients of σ_{ld}^2 and σ_{ax}^2, it can be concluded that the σ_{SNM} should remain the same, irrespective of the data value stored in the bit-cell. This has been verified by the following experiments.

Consider Fig.3.7 (c). Here, 10000 points (or 10000 bit-cells) are generated with independent gaussian variation in the V_{th} all six bit-cell transistors. MC simulations are run when all the bit-cells store "0" (SNM0) and when all the bit-cells store "1" (SNM1) at node VL. The histogram results show that the SNM distribution in both the cases is the same (Mean and standard deviation for SNM0 are 131.8mV and 22.1mV; for SNM1, these are 131.4mV and 22.1mV, respectively). In another experiment, for each of the 10000 points, "0" or "1" is randomly assigned to the cell. In this case also, the distribution is identical, with the mean = 131.7mV and standard deviation = 22.1mV. This experiment is repeated in other ways, by assigning "0" and "1" to alternate data points, "0" to first 5000 points and "1" to the next 5000 points, etc. The results are always identical. This indicates that the date value, stored in the bit-cells, is not important from the statistical distribution point of view.

What matters, is whether the bit-cell is storing favorable or unfavorable data. For example, assume that for the first case (in 10000 generated cases), the threshold voltages of the six transistors are such that SNM0 is higher than SNM1. Therefore, for this particular bit-cell, "0" is the favorable data value and "1" is unfavorable. For the 10000 different bit-cells, with different V_{th} values for all six transistors, each bit-cell has either "0" or "1" as the favorable data value. Therefore, whether we store all "0"s or all "1"s or random assignment, it is safe to say that half the cells store favorable data and the remaining half store unfavorable data. Therefore, the SNM distribution remains the same, irrespective of the data assignment.

We have also considered the case when all 10000 cells store unfavorable data (the probability of this happening is very small). For this, we pick the minimum of SNM0 and SNM1 as $SNM_{worst-case}$, for each of the 10000 cases. The distribution of $SNM_{worst-case} = Min(SNM0, SNM1)$ is depicted in Fig. 3.7(c). The minimum SNM value, in the distribution of $SNM_{worst-case}$, coincides with the minimum of SNM0 or SNM1 (therefore, the left tail of the $SNM_{worst-case}$ distribution, in Fig. 3.7(c), coincides with that of SNM0 and SNM1). Hence, if we can ensure than the minimum SNM value in the distribution of SNM0 or SNM1 (which is given on the x-axis by, say $SNM_{avg} - 5\sigma_{SNM}$) is within the acceptable limits, then a good SNM yield can be ensured, irrespective of the stored data pattern in the

10000 cells. Similar argument holds true for all other design metrics. The above discussion justifies our modeling strategy to compute the average and variance of the distributions of the design metrics.

To summarize, in this section, a modeling approach is presented, to estimate the statistical mean and variance of the intra-die distributions of the design metrics - SNM, Vtrip, read current and leakage. Intrinsic V_{th} fluctuations due to RDF are considered to be the major source of intra-die variations in the design metrics. The results of modeling are verified with MC simulations. It should be noted that although millions of MC simulations are required to cover a range of 4σ to 5σ variation, a small number of simulations (e.g. 10000) is sufficient to converge on reasonably accurate values of statistical mean and variance.

Constraint Formulation

Because of the within-die variations, each SRAM bit-cell differs from a million others in the array in its characteristics such as the SNM, vtrip, read speed and leakage. The number of identical bit-cells and the expected electrical yield to the specifications determine the number of sigmas - N_{σ}, over which the bit-cell must operate correctly [21],[51]. E.g., N_{σ}=4.763, mathematically corresponds to only a single cell failure in an array of 1024 X 1024 cells [21], that is, a yield of 99.9999%. It should be noted that SNM, Vtrip, read current and leakage, only cause failure on one side of the statistical variation. Therefore, 1.35 cells per 1000 fall outside the $\pm 3\sigma$ range [21]. Typically, the required number of sigmas ranges from 4 to 5. This concept is used to formulate the constraints of the optimization problem as follows:

$$\frac{SNM_{avg} - SNM_{residual}}{\sigma_{SNM}} \geq N_{\sigma} \tag{3.7}$$

Equation 3.7 applies a constraint on the SNM yield instead of the SNM average value. The value of N_{σ} is selected according to the required yield, redundancy and level of integration. The constraint bounds, also called residuals, such as the $SNM_{residual}$ in equation 3.7 impart the necessary flexibility to design different versions of the bit-cell. Such bounds on the read current (lower limit) and leakage (upper limit) enable the design of a bit-cell to have a high or moderate performance, and a low or ultra-low leakage. The residuals on SNM and Vtrip are used to build margin for reliability.

For the constraint in equation 3.7, the residual and the N_σ are the inputs, chosen by the designer. The average and the standard deviation of the distributions (SNM_{avg} and σ_{SNM}) are computed by modeling and are impacted by the choice of the nominal design parameters as observed from equations 3.2, 3.5 and 3.6. Therefore, the extent of the intra-die variations, and hence, the yield, can be controlled by the judicious choice of the nominal design. The constraint is depicted in Fig.3.8.

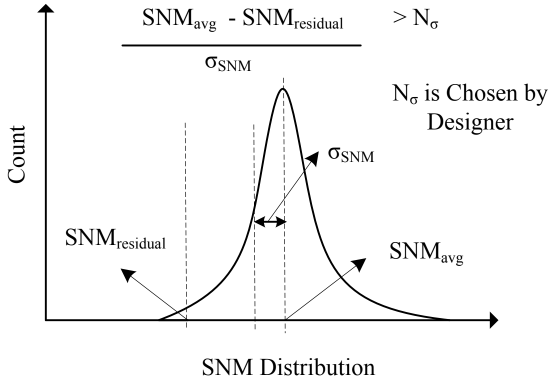

$\frac{SNM_{avg} - SNM_{residual}}{\sigma_{SNM}} > N_\sigma$
N_σ is Chosen by Designer
Count
σ_{SNM}
$SNM_{residual}$
SNM_{avg}
SNM Distribution

Figure 3.8: Pictorial representation of the SNM design constraint.

The constraint formulation, in equation 3.7, is applicable for only gaussian distributions. Consequently, it can be used for the SNM, Vtrip and the read current metrics, but not for the bit-cell leakage. This is because the sub-threshold leakage, which is the primary component of the total leakage, varies exponentially with the threshold voltage. Therefore, the total bit-cell leakage acquires a lognormal distribution with the V_{th} variations.

In such a scenario, the central limit theorem can be applied [63],[71] to model the sum of the leakage of 16 bit-cells as a gaussian distribution. Fig.3.6(b) signifies that the sum of the leakage of 16 bit cells (when the V_{th} of every transistor in each of the 16 bit-cells is a random variable) displays a gaussian distribution. The normal probability plot in Fig.3.6(b), is a straight line, which indicates gaussian distribution. Because of the associated overhead of the peripherals in the memories, the deployment of memories as storage elements is justified for only a certain minimum number of bits (more than 16). Therefore, the use of 16 cells for leakage modeling does not restrict the minimum memory size. The mean and sigma values of the sum of the leakage of 16 cells is given as follows[71]:

$$\begin{aligned} Leakage\_mean_{16cells} &= 16 \times Leakage\_mean_{1cell} \\ and\, \sigma^2_{16cells} &= 16 \times \sigma^2_{1cell}. \end{aligned} \tag{3.8}$$

3.4.2 Inter-Die Variations

Feasible Region

The previous section considers the intra-die variations. Simultaneously, the variability in the design metrics due to the inter-die variations in the transistor dimensions should also be considered for an overall yield maximization. For the sake of simplicity, first assume that the design metrics (SNM, Vtrip, read current, cell leakage and area) are constrained as simple inequalities of the following form - $SNM \geq SNM_{min}$, $Vtrip \geq Vtrip_{min}$, $Iread \geq Iread_{min}$, $Ileak \leq Ileak_{max}$ and $Area \leq Area_{max}$.

Let us call the aforementioned inequalities as performance constraints. Each of the performance constraints is determined by the transistor dimensions, also called the design parameters - W_{drv} ,W_{ax} ,W_{ld} ,L_{drv} ,L_{ax} ,L_{ld}. The design parameters define a six-dimensional parameter space. The circuit specifications ($SNM_{min}, Vtrip_{min}$, etc.) determine a region within the parameter space, where the circuit is acceptable, i.e., all of the inequalities of the performance constraints are satisfied. For illustration, consider Fig. 3.9, which depicts a three-dimensional parameter space defined by W_{drv} ,W_{ax} and W_{ld}. It is not possible to visualize the problem in more than three dimensions. Therefore, for illustration only, it is assumed that only W_{drv} ,W_{ax} and W_{ld} are available as design parameters to the designer (all transistor lengths are fixed).

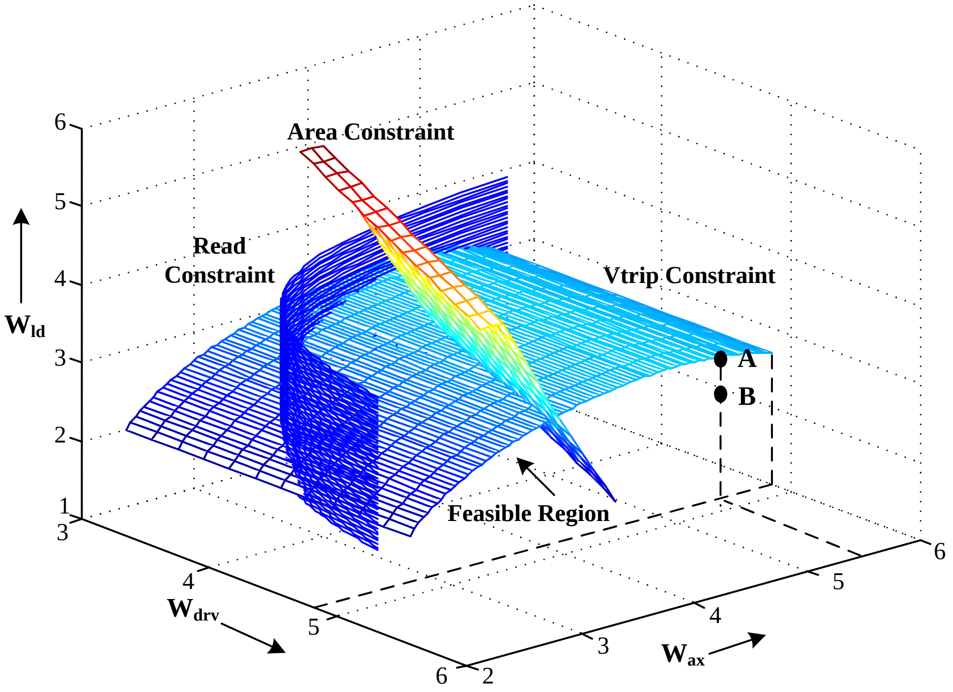


Figure 3.9: Pictorial representation of the feasible region in 3-dimensions

Fig. 3.9 contains three constraint surfaces. For instance, the Vtrip constraint surface is defined by the equality $Vtrip(W_{drv}, W_{ax}, W_{ld}) = Vtrip_{min}$. For different combinations of W_{drv} and W_{ax}, the numerical equation solver is used to obtain the W_{ld}, which satisfies the equality $Vtrip(W_{drv}, W_{ax}, W_{ld}) = Vtrip_{min}$. This procedure (with simple models for the transistor, and Vtrip dependence on the transistor widths) provides all the points (e.g. point A) on the Vtrip constraint surface. All the points, under the Vtrip constraint surface, satisfy the inequality $Vtrip \geq Vtrip_{min}$. For example, by comparing the coordinates of points A and B, it can be observed that the W_{drv} and W_{ax} at point B are the same as those at point A. But, W_{ld} at point B, is smaller than the W_{ld} at point A. Since Vtrip improves with a reduction in W_{ld}, the Vtrip at point B is more than that at point A. This implies that all the points, under the Vtrip constraint surface, lie in the Vtrip acceptability region. The intersection of the acceptability regions for the Vtrip, read current and area constraints, is indicated as the 'feasible region' in Fig. 3.9. Therefore, the designer must choose W_{drv}, W_{ax} and W_{ld}, such that the design point defined by them, lies within the feasible region, to satisfy the performance constraints.

In the illustration in Fig. 3.9, the feasible region is determined explicitly, that is, by expressing Vtrip, read current and area, analytically as a function of the transistor widths. Explicit methods develop an approximation of the feasible region [72]. One way to achieve this is to approximate the performance constraints by simple analytical expressions in the region of interest. However, very often, as also in our case, it is not possible to accurately express the performance constraints like SNM, Vtrip, read current and leakage as analytical expressions. For example, just the saturation current through a single transistor requires the complex BSIM model to capture all the effects. Various components of leakage through the bit-cell require numerical solution of complex equations. Developing analytical expressions for SNM and Vtrip is also quite involved. Therefore, techniques such as curve-fitting and polytope approximation may be employed. But, with the increase in the dimensionality of the parameter space (e.g. six), these techniques become computationally expensive. Moreover, in the industry, the designers can use the BSIM transistor models for simulation, to obtain the values for SNM, Vtrip, read current and leakage; and do not wish to spend time in developing analytical expressions for the performance constraints. Therefore, in this work, implicit determination of feasible region is applied [72], that is, whether or not a chosen point, in the parameter space, lies in the feasibility region, is determined by simulations and by evaluating the performance constraints (SNM, Vtrip, read current, leakage) at the chosen point, to verify that the specifications are met.

Yield Definition

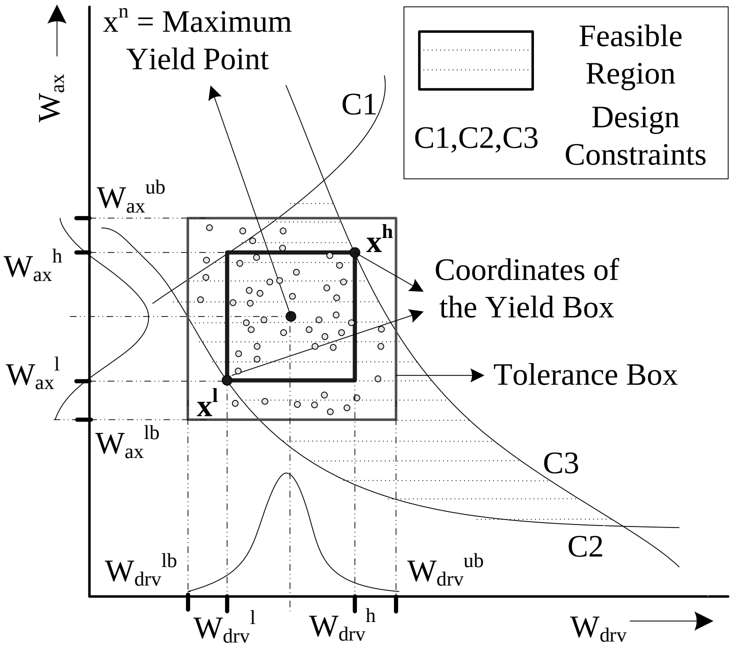


Figure 3.10: Simplified yield maximization method in 2-dimensions

The proposed yield maximization method is presented now. As explained earlier, the design constraints define a feasible region in the parameter space, within which the nominal design should be chosen. The problem is depicted graphically in two-dimensions (2-D) in Fig.3.10 (for illustration) for an arbitrary feasible region, defined by arbitrary constraints. If the design parameters W_{drv} and W_{ax} are exactly realizable, then it is a deterministic optimization problem. Solving such a problem, would provide what is called the “nominal design”. However, due to inter-die manufacturing imperfections, it is not possible to realize the nominal design value exactly. The nominal design can only be specified with a tolerance. For example, if the nominal W_{drv} is 200nm, then because of the inter-die W_{drv} variations, the W_{drv} in the fabricated bit-cells can range from 197nm to 203nm. These lower and upper bounds can be referred to as W_{drv}^{lb} and W_{drv}^{ub}, respectively. The total spread or tolerance of the design parameter W_{drv}, is therefore, 6nm. The same applies to W_{ax}. The design parameters must be assumed to be random variables whose probability distributions are known. Knowing only the tolerances would imply uniform distribution for the design parameters, which is not accurate or realistic. E.g., for a nominal W_{drv} of 200nm, assuming that the probability of fabricated W_{drv} being 200nm and 197nm, is the same, is incorrect. Therefore, the design parameters W_{drv} and W_{ax} in 2-D, in Fig.3.10 (and $W_{ld}, L_{drv}, L_{ax}, L_{ld}$, in 6-D) are assumed to have a gaussian distribution around their respective chosen nominal value. The spread or tolerance of the gaussian distribution is assumed to be 6nm, as explained in Section 3.3.2.

For a chosen nominal design, a tolerance box can be specified around it, such that the dimensions of the tolerance box are defined by the spread of the design parameters, i.e., $\pm 3\sigma$ value of the normally distributed widths and lengths. This is shown in Fig.3.10. The smaller dots within the tolerance box represent all the possible design realizations due to variations in the design parameters, around their respective nominal values. This is analogous to throwing darts on a dart board. The dart board corresponds to the feasible region. The smaller dots represent all the darts thrown at the board. Yield is the ratio of the number of dots within the feasible region (acceptable realizations) over the total number of dots (all realizations). Therefore, the overlap between the tolerance box and the feasible region represents the yield. The nominal design should be moved (the tolerance box moves with it) and located in the feasible region, in a way, which ensures the maximum overlap between the tolerance box and the feasible region, thus maximizing the yield.

However, the overlap can assume any shape and a way to estimate the area of the overlap region for measurement of yield is needed. The inner box in Fig.3.10 is the maximum orthogonal overlap that is attained between the feasible region and the tolerance box, and can be used very well, to estimate the yield directly [73]. In six dimensions (which is the case of bit-cell design problem), maximizing the six-dimensional volume of the inner box or yield box (defined by coordinates, x^l and x^h), maximizes yield.

Qualitatively, the problem is reduced to finding x^l and x^h, the coordinates of the yield box in Fig.3.10 such that the following two conditions are satisfied. The first condition is that the yield box should lie within the tolerance box, which implies that the maximum difference between x^l and x^h should not be more than the maximum spread in the design parameters. The second condition is that the yield box should lie within the feasible region, which implies that all the points, lying within the yield box, should satisfy all the design constraints.

If the above mentioned conditions are met, then for the nominal design placed within the yield box, the probability (yield) that the design constraints are satisfied in the presence of parameter variations, can be estimated in two-dimensions (for illustrative purpose) as follows:

$$\begin{aligned}
P_{2-D} &= P\Big((W_{drv}^l \leq W_{drv} \leq W_{drv}^h)\, and\, (W_{ax}^l \leq W_{ax} \leq W_{ax}^h)\Big) \\
&= P(W_{drv}^l \leq W_{drv} \leq W_{drv}^h) \times P(W_{ax}^l \leq W_{ax} \leq W_{ax}^h) \\
&= (CDF(W_{drv}^h) - CDF(W_{drv}^l)) \times (CDF(W_{ax}^h) - CDF(W_{ax}^l)).
\end{aligned}$$

where W_{drv}^l , W_{drv}^h, W_{ax}^l and W_{ax}^h form the coordinates of the yield box (in 2-D as in Fig. 3.10) and $CDF(x)$ represents the cumulative distribution function of x [71]. Extending this to six dimensions (which is the case of our problem),

$$Yield(x^l, x^h) = \prod_{i=1}^{6} P(x_i^l \leq x_i \leq x_i^h) \; = \; \prod_{i=1}^{6} (CDF(x_i^h) - CDF(x_i^l)), \tag{3.9}$$

where x_i is the i^{th} design parameter. x_i^l and x_i^h are the coordinates of the yield box in the i^{th} dimension. The distribution of the physical parameters such as the transistor widths and lengths is gaussian, which does not have a closed form CDF. Therefore, the solution of (3.9) requires solution of a multi-dimensional probability integral by quadrature or Monte-Carlo based methods, which is computationally expensive [72]. The problem is simplified if a closed form expression for CDF can be used. Therefore, a double-bounded probability density function (DB-PDF), proposed by Kumaraswamy [74], is employed. With this model, the pdf (probability density function) of z - $f(z)$ is given by

$$\begin{aligned} f(z) &= abz^{a-1}(1 - z^a)^{b-1}, \\ where\, z &= \frac{x - x^{lb}}{x^{ub} - x^{lb}}\;,\; x^{lb} \leq x \leq x^{ub}. \end{aligned} \tag{3.10}$$

In equation 3.10, z is the normalized value of x, x^{lb} and x^{ub} are the lower and upper bounds, respectively, of the double-bounded random variable x. Assuming that the statistical distributions of the design parameters are independent, the joint probability density function is given by the product of the individual DB-PDF. a and b are the shape parameters and distributions such as uniform, triangular ,gaussian can be obtained by using different values of a and b. The joint PDF of two variables $z1$ and $z2$, computed using the DB-PDF, for different values of a and b, are demonstrated in Fig. 3.11. In this work, $a = 3$ and $b = 8$, are used to obtain a truncated gaussian shape. Assuming a truncated gaussian distribution is appropriate for physically bounded dimensions like the transistor widths and lengths. The $\pm 3\sigma$ is taken as the spread of the gaussian distribution around the nominal design x^n, as explained in Section 3.3.2.

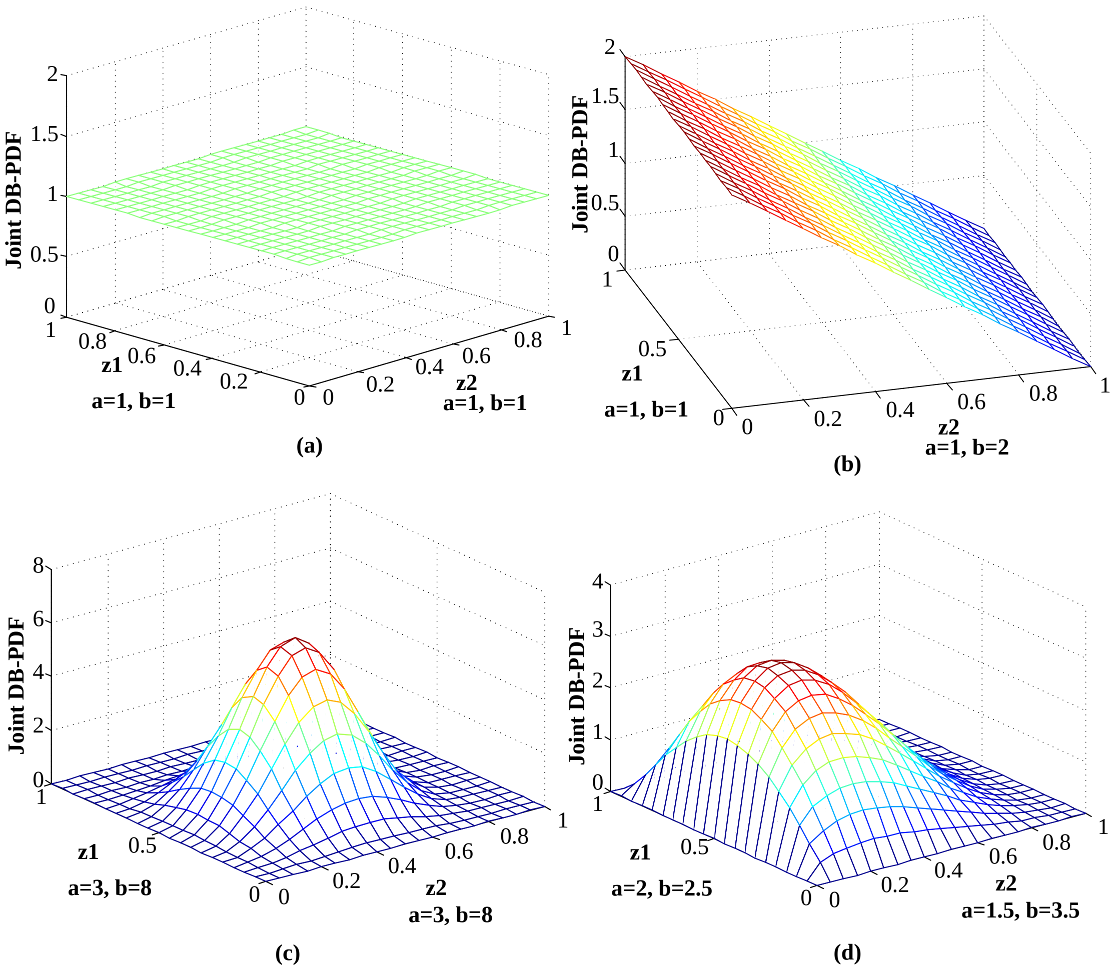


Figure 3.11: Joint PDF for variables $z1$ and $z2$ (a) Uniform (b) Triangular (c) Gaussian (d) Skewed

Therefore,

$$x^{ub} = x^n + 3\sigma_x\ ,\ x^{lb} = x^n - 3\sigma_x\ ,\ t = x^{ub} - x^{lb} = 6\sigma_x. \tag{3.11}$$

In equation 3.11, t represents the maximum spread of the design parameter x. The closed form DB-CDF can now be obtained by integrating $f(z)$ and is given by:

$$F(z) = 1 - (1 - z^a)^b. \tag{3.12}$$

In summary, DB-PDF has been chosen to approximate the distributions of the design parameters, because it provides a simple closed form analytical expression for the CDF. This expression can be used in equation 3.9, to compute the yield directly.

Due to the symmetrical nature of the distribution of the design parameters, the final optimized design solution is the center of the yield box and is computed as

$$x^n = \frac{x^l + x^h}{2}. \tag{3.13}$$

By using equations 3.9, 3.10, 3.11, 3.12 and 3.13, the yield in six-dimensions can be computed as follows:

$$\begin{aligned} Yield(x^l, x^h) &= \prod_{i=1}^{6} \left(F(z_i^h) - F(z_i^l) \right), \\ &= \prod_{i=1}^{6} \left(F(\frac{x_i^h - x_i^{lb}}{x_i^{ub} - x_i^{lb}}) - F(\frac{x_i^l - x_i^{lb}}{x_i^{ub} - x_i^{lb}}) \right) \\ &= \prod_{i=1}^{6} \left(F(\frac{x_i^h - (x_i^n - 0.5t)}{t}) - F(\frac{x_i^l - (x_i^n - 0.5t)}{t}) \right) \\ &= \prod_{i=1}^{6} \left(F(\frac{x_i^h - x_i^l + t}{2t}) - F(\frac{x_i^l - x_i^h + t}{2t}) \right). \end{aligned} \tag{3.14}$$

Equation (3.14) gives the probability of finding a design solution in the six-dimensional yield box, given the probability distributions of the widths and lengths of the transistors. Maximizing this, would maximize the yield.

Constraint Verification Approach

Equation (3.14) expresses the yield as a function of x^l and x^h, the coordinates of the yield box. Our intent is to widen the dimensions of the yield box, to approach the tolerance box, to maximize yield. While doing so, the yield box should lie in the feasible region, the entire time. In other words, all the points within the yield box, should meet the specifications of the performance constraints. To achieve this, as a first order condition, it is sufficient to check the constraint violation at the extreme corners of the yield box, which are given by $\{x^l, x^h\}$. For example, in two-dimensions there are $2^2 = 4$ corner points for the yield box, as illustrated in Fig.3.12 (b).

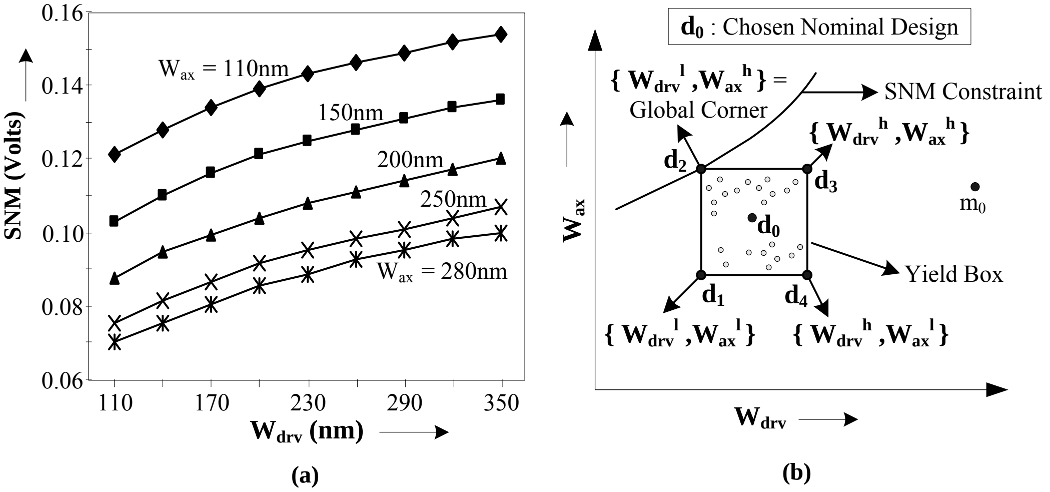


Figure 3.12: (a) SNM variation with W_{drv} and W_{ax} (b) Constraint Minimization

With this principle, in six dimensions, for each choice of the nominal design, it is required to simulate at $2^6 = 64$ corners, for each design constraint, to ensure that the yield box lies in the feasible region. However, this number can be reduced significantly by the application of the design understanding. For instance, as shown in Fig. 3.12 (a) for a 45nm design, the SNM degrades with reducing driver width. The SNM also degrades with an increase in the strength of the access transistor. Therefore, in 2 -D, the SNM constraint can be evaluated at only $\{W_{drv}^l, W_{ax}^h\}$, as depicted in Fig. 3.12 (b). For the other three corners, the SNM is only going to be better. Using this approach, the number of constraint evaluation corners can be minimized. The evaluation corners for all the design constraints are mentioned in Table 3.2.

Some of the entries in Table 3.2 are intuitive. E.g., the leakage constraint would

Table 3.2: Evaluation Corners for the Design Constraints

Design Constraint	W_{drv}	W_{ax}	W_{ld}	L_{drv}	L_{ax}	L_{ld}
SNM	L	H	L	L/H	L	H
Vtrip	H	L	H	L	H	L
Read Current	L	L	L	H	H	H
Leakage	H	H	H	L	L	L

be the worst at larger widths and smaller lengths and therefore, should be evaluated at that corner of the yield box, which is defined by $W^h_{drv}, W^h_{ax}, W^h_{ld}, L^l_{drv}, L^l_{ax}, L^l_{ld}$. The entries in Table 3.2 reflect trends, and should remain the same for all technologies. Some of the trends are depicted in the following figures. The variation of SNM with L_{drv} and L_{ax} is depicted in Fig. 3.13(a). Fig. 3.13 (b) shows the SNM variation with W_{ld} and L_{ld}.

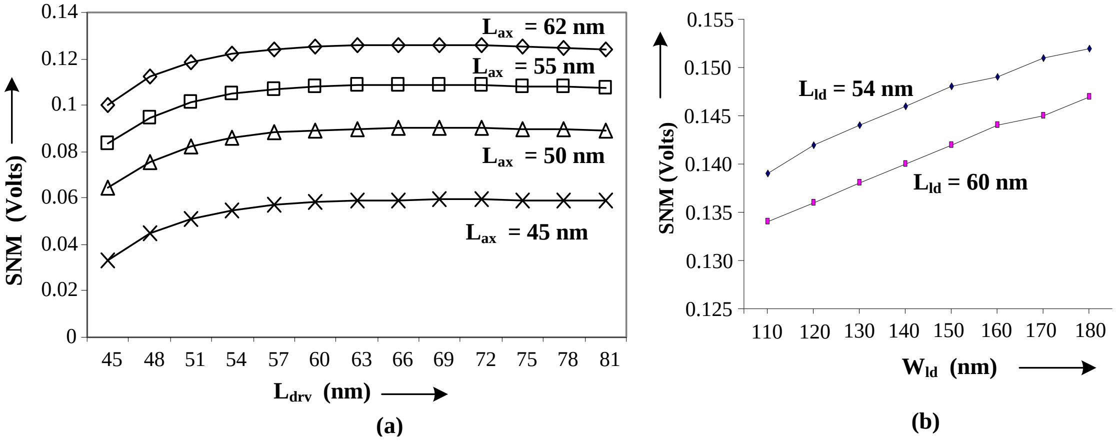


Figure 3.13: SNM variation with (a) L_{drv} and L_{ax} (b) L_{ld} and W_{ld}

Fig.3.14(a) shows the Vtrip variation with W_{drv} and W_{ax}. Some trends are not intuitive. For example, Fig.3.14(b) shows that the read current does not vary significantly with W_{ld}. However, a close inspection (inset) reveals that the current improves slightly for a larger load. A larger PMOS generates a stronger "1" at VR, improving the gate to source voltage of M1 through which the bit line discharges during *read*. Therefore, the read current should be evaluated only for W_{ld}^l. For all those corners of the 6-D cube, which involve W_{ld}^h, the read current need not be checked. By using this concept of constraint minimization in six dimensions, the total number of constraint evaluations, for every choice of the nominal design, is reduced to five. The area should be calculated by equation 3.1, at the nominal design parameters. This is because the allotted bit-cell boundary on silicon remains the same, even when the fabricated transistor sizes vary.

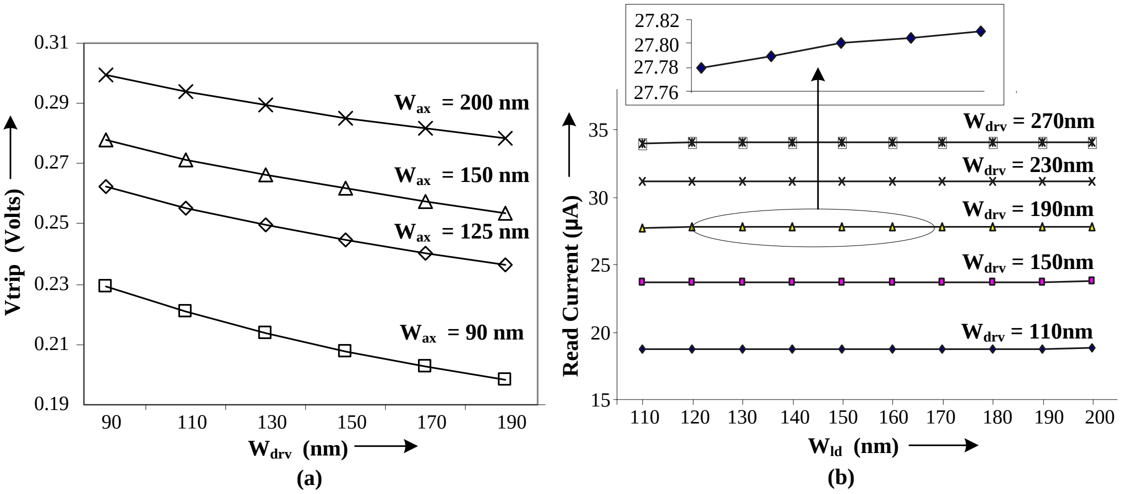


Figure 3.14: For a 45nm design -(a) Vtrip variation with W_{drv} and W_{ax} (b) Read current variation with W_{drv} and W_{ld}

Finally, the following constraints are added to the optimization problem.

$$x^l < x^h \tag{3.15}$$

$$\text{and } x^h - x^l \leq t. \tag{3.16}$$

Impact of Global Variations on Local Variations

In this section, the impact of global (inter-die) variations on the local (intra-die or within-die) variations is discussed. It can be observed from equation 3.6, that the standard deviation of the intra-die distribution of SNM - σ_{SNM}, is a function of the σ_{Vth} of the driver, access and load transistors. In turn, σ_{Vth} of each transistor is a function of the transistor width and length, as expressed in equation 3.2. Therefore, σ_{SNM}, which is a measure of the local variations, is influenced by the global or inter-die variations in the transistor dimensions.

To assess the impact of the global variations in the transistor sizes on the σ_{SNM}, the following experiment is conducted. A population of 15000 points is generated, with independent, gaussian variations in the widths and lengths of the bit-cell transistors. Therefore, each of the 15000 global design variants, has different σ_{Vth} values for the driver, access and load transistors (from equation 3.2). Around ten of the global design variants, ten populations of 5000 points each, with random V_{th} variations in all six transistors, are generated. SNM simulations for these 10 sets of 5000 points are used to calculate the statistical σ_{SNM}. These match very well with the σ_{SNM} computed using (3.2) and (3.6). Therefore, instead of running 5000 SNM simulations for each of the 15000 global design variants, equations (3.2) and (3.6) can be reliably used to evaluate σ_{SNM} at each of the 15000 global design variants.

The results are demonstrated in Fig. 3.15 (a). The figure indicates that 99.3% of the 15000 global variants have a σ_{SNM} (due to local variations), which is within $\pm 1mV$ of the nominal value (σ_{SNM} at nominal transistor widths and lengths). Therefore, intra-die σ_{SNM} can be assumed to remain approximately the same across all the dies (global variants). With respect to Fig.3.12 (b), this implies that for the inter-die variants around d_0, such as d_1, d_2 and so on, σ_{SNM} can be assumed to be the same. But if a different nominal design such as m_0 is chosen (Fig.3.12 (b)), then the σ_{SNM} can change appreciably and should be re-evaluated.

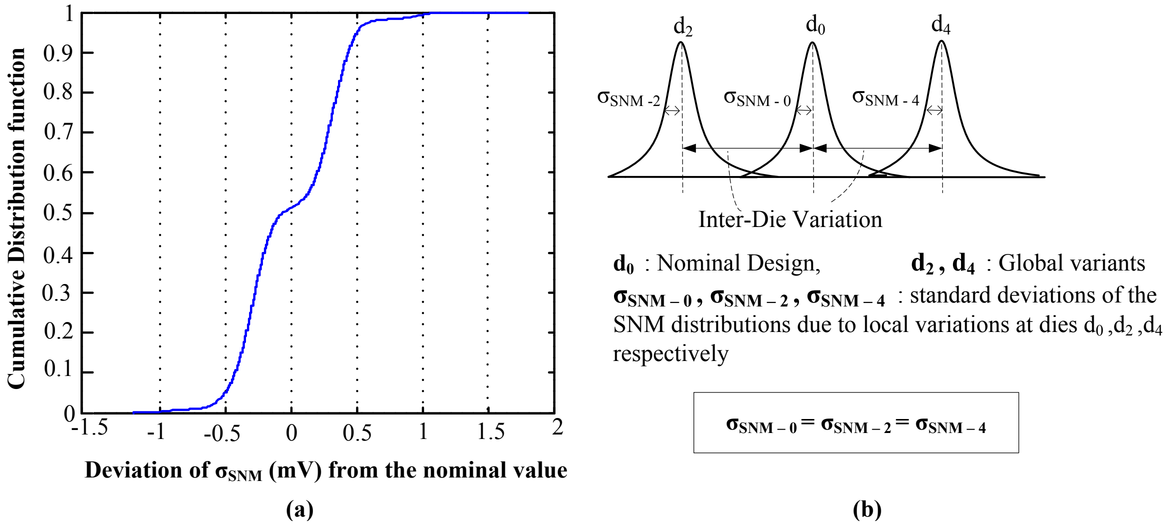


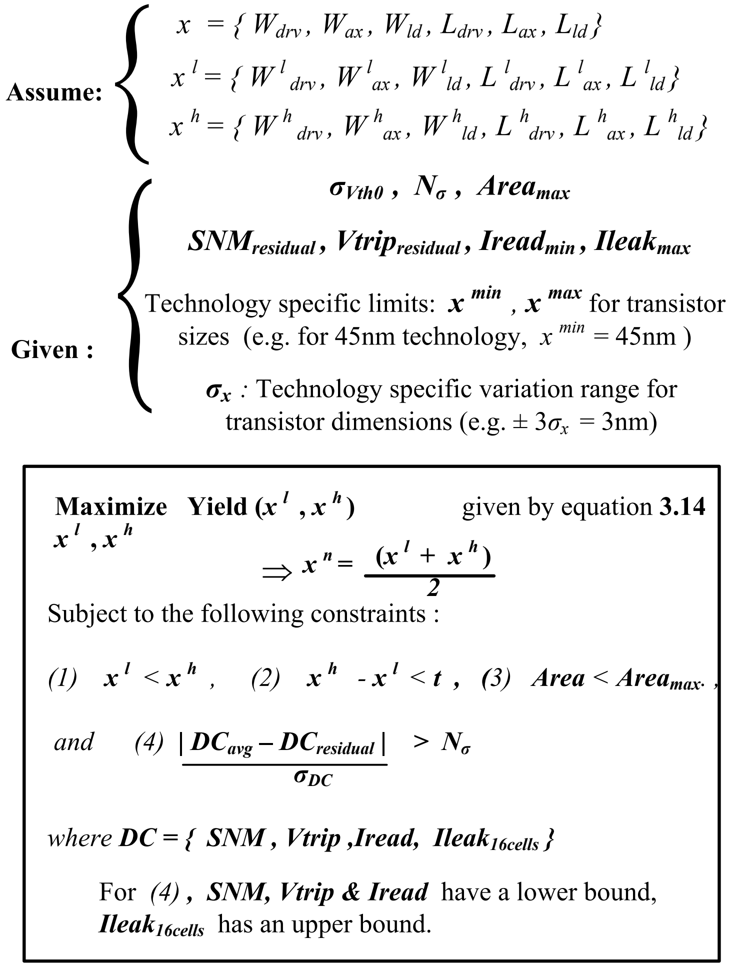
Figure 3.15: (a) CDF plot of Deviation in σ_{SNM} (b) SNM distribution at different dies

As technology progresses, the global variations have a greater impact on the local variation [52]. However, the local σ_{Vth} and σ_{SNM} would also increase with technology scaling and the percentage change in the local σ_{SNM} due to global variations is not expected to increase. Hence, it is reasonably accurate and practical to assume the same within-die variation for all the dies, as shown in Fig. 3.15 (b).

Till now, we have formulated the problem for simple constraints of the form - $SNM \geq SNM_{min}$. It needs to be verified that the constraint minimization proposed in Table 3.2, is applicable to the original constraints defined by equation (3.7) - $(SNM_{avg} - SNM_{residual})/\sigma_{SNM} \geq N_\sigma$. Here, $SNM_{residual}$ and N_σ are fixed, user-defined inputs. In Fig.3.12 (b), d_0 is the chosen nominal design. d_1, d_2 and all other design points on or within the yield box occur because of the inter-die variations in the transistor dimensions. It has been discussed in the previous paragraph, that for all these inter-die variants, σ_{SNM} can be assumed to remain fairly constant. Therefore, $(SNM_{avg} - SNM_{residual})/\sigma_{SNM} \geq N_\sigma$ can be evaluated at the corner, which has the worst SNM_{avg}, i.e., the die which constitutes the global worst-case corner, such as the point d_2 in Fig.3.12 (b), in 2-D. The same has been observed in [75]. Therefore, the proposed constraint minimization strategy in Table 3.2, is applicable to the constraints, defined by equation (3.7).

3.4.3 Final Optimization problem

To summarize, the final optimization problem is as follows:



Assume:

$$\begin{cases} x = \{ W_{drv}, W_{ax}, W_{ld}, L_{drv}, L_{ax}, L_{ld} \} \\ x^{l} = \{ W^{l}{}_{drv}, W^{l}{}_{ax}, W^{l}{}_{ld}, L^{l}{}_{drv}, L^{l}{}_{ax}, L^{l}{}_{ld} \} \\ x^{h} = \{ W^{h}{}_{drv}, W^{h}{}_{ax}, W^{h}{}_{ld}, L^{h}{}_{drv}, L^{h}{}_{ax}, L^{h}{}_{ld} \} \end{cases}$$

Given :

$\boldsymbol{\sigma_{Vth0}}$ **,** $\boldsymbol{N_{\sigma}}$ **,** $\boldsymbol{Area_{max}}$

$\boldsymbol{SNM_{residual}}$**,** $\boldsymbol{Vtrip_{residual}}$**,** $\boldsymbol{Iread_{min}}$**,** $\boldsymbol{Ileak_{max}}$

Technology specific limits: $\boldsymbol{x}^{min}$, $\boldsymbol{x}^{max}$ for transistor sizes (e.g. for 45nm technology, x^{min} = 45nm)

$\boldsymbol{\sigma_x}$ *:* Technology specific variation range for transistor dimensions (e.g. ± $3\sigma_x$ = 3nm)

Maximize Yield $(\boldsymbol{x^{l}}, \boldsymbol{x^{h}})$ given by equation **3.14**
$\boldsymbol{x^{l}}, \boldsymbol{x^{h}}$

$$\Rightarrow x^{n} = \frac{(x^{l} + x^{h})}{2}$$

Subject to the following constraints :

(1) $\boldsymbol{x^{l} < x^{h}}$, *(2)* $\boldsymbol{x^{h} - x^{l} < t}$, *(3)* $\boldsymbol{Area < Area_{max}}$. ,

and *(4)* $\dfrac{|\boldsymbol{DC_{avg} - DC_{residual}}|}{\boldsymbol{\sigma_{DC}}} > \boldsymbol{N_{\sigma}}$

where $\boldsymbol{DC = \{ SNM, Vtrip, Iread, Ileak_{16cells} \}}$

For *(4)* **,** ***SNM, Vtrip & Iread*** have a lower bound, $\boldsymbol{Ileak_{16cells}}$ has an upper bound.

Figure 3.16: Final Optimization Problem

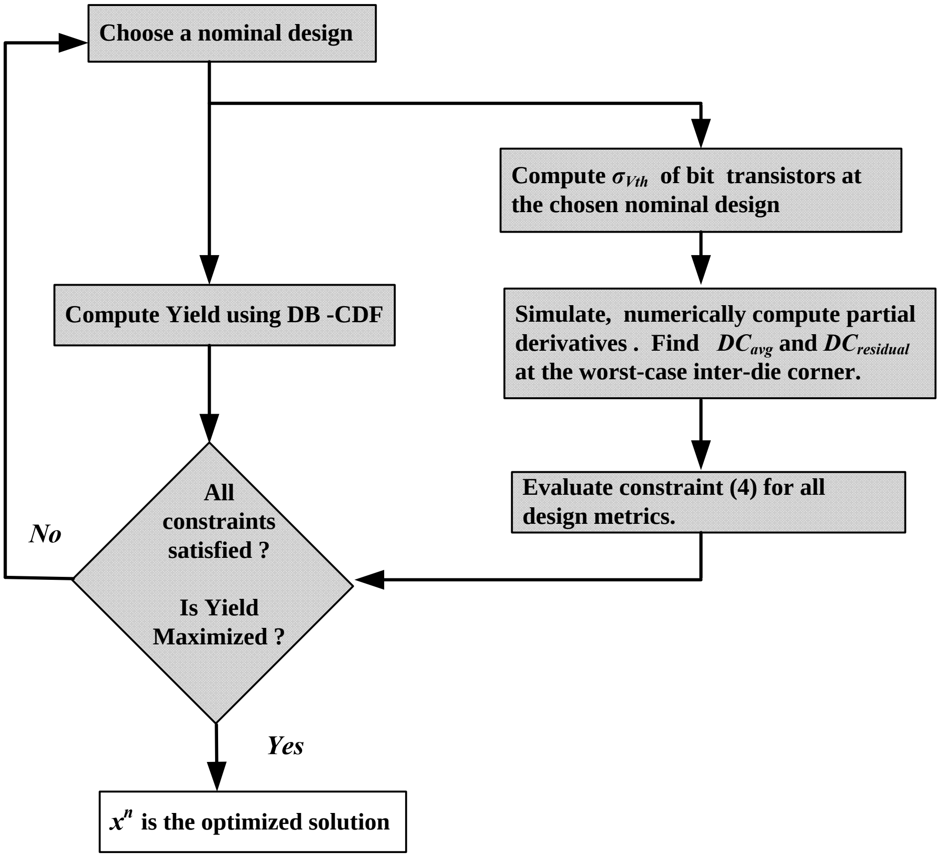

Choose a nominal design
Compute σ_{Vth} of bit transistors at the chosen nominal design
Compute Yield using DB -CDF
Simulate, numerically compute partial derivatives . Find DC_{avg} and $DC_{residual}$ at the worst-case inter-die corner.
Evaluate constraint (4) for all design metrics.
All constraints satisfied ?
Is Yield Maximized ?
No
Yes
x^n is the optimized solution

Figure 3.17: Optimization Method

A Sequential Quadratic Programming based optimization engine [76] is used to solve the constrained optimization problem in six dimensions. The optimization engine dynamically provides the lengths and widths of the bit-cell transistors to HSPICE [77] template files for simulation, to obtain the updated values of the design constraints. This continues till the optimizer arrives at an optimal set of the sizes for the bit-cell transistors. BSIM4 based simulator [78] has been used. The optimization procedure is presented in Fig. 3.17. Predictive Technology models are used [79].

3.5 Results and Discussion

To obtain an optimal 45nm bit-cell design, the following set-up has been used:

1. $Area_{max} = 0.68\mu m^2$, and $\sigma_{Vth0} = 55mV$ [51].

2. $N_\sigma = 4.763$, which mathematically, corresponds to only a single cell failure in an array of 1024×1024 cells. Typically, the largest size of an embedded SRAM block is 256 Kbits [12]; for larger blocks, performance begins to degrade. Therefore, the chosen value of N_σ covers a wide range of embedded memory sizes. With the use of page architecture and/or redundancy, the designer can reduce the required N_σ.

3. The transistor dimension is altered in steps of 1nm. This is the step size available for altering the layout geometries in 45nm technology.

4. Usually, the required residuals for the SNM and Vtrip are set to 0 and the designers attempt to obtain a value of 4 to 5 for N_σ [51]. However, in this work, the SNM and Vtrip residuals are selected as $15mV$ and $25mV$, respectively to set a higher reliability margin. Note that these are the desired bounds for the worst-case voltage-temperature corners, as in Table 3.1.

3.5.1 General purpose-high performance design

With the previous fixed inputs, different performance (read current) and leakage targets are used to design a general purpose bit-cell on the high performance-moderate leakage side. The results, documented in Table 3.3, and the associated trends are analysed to understand how the proposed method works. For a maximum cell leakage target of $60nA$ (the conventional leakage values mentioned in [80]), the required read current residual (the minimum required read current for the statistically weakest bit-cell) is increased from $30\mu A$ to $45\mu A$. This is documented in the first column of Table 3.3.

For every read current residual requirement, the optimal bit-cell design solutions are derived with the proposed statistical design method. The corresponding transistor sizes and the bit-cell areas are shown in columns 2 and 3 of Table 3.3, respectively. As explained in Section 3.4.1, N_σ is a measure of the maximum intra-die variation (within-die) that the design can tolerate. This number, measured at the worst-case inter-die variant (global corner) is used as a typical figure of merit in the industry for the electrical yield (e.g. ,SNM yield, Vtrip yield). The output N_σ, for all the design constraints, are shown in columns 4 to 8. Finally the actual yield, mentioned in the last column, is obtained by running Monte-Carlo simulations with 10 000 points at the resultant optimal designs. Here, gaussian variation is applied to the widths, lengths and threshold voltages of all the six transistors. The yield is equal to the percentage of bit-cells that satisfy the desired bounds for all the design constraints.

Table 3.3: Design of high performance - moderate leakage bit-cell - Optimization Results

Iread	W_{drv}, W_{ax}, W_{ld},	Area	N_σ	N_σ	N_σ	N_σ	N_σ	Yield
Res.	L_{drv}, L_{ax}, L_{ld}		Iread	Ileak	SNM	SNM	Vtrip	
(μA)	(nm)	(μm^2)			Low V	High V		(%)
30	419, 309, 103, 65, 73, 56	0.5660	5.432	14.536	5.044	4.903	10.921	100
35	419, 347, 106, 63, 77, 51	0.5704	4.976	6.482	5.049	5.117	11.062	100
40	437, 390, 117, 62, 77, 53	0.5905	4.970	5.581	4.820	4.808	12.047	100
45	463, 389, 145, 62, 71, 55	0.6218	4.768	5.232	4.771	4.795	11.637	99.8

Several interesting observations can be made from these results. Fig.3.18(a) shows the nominal read current and bit-cell leakage values at the resultant optimal designs. It can be observed that any gain in the read performance is accompanied by an increase in the nominal leakage. $Iread_{min} = 30\mu A$ is a relaxed performance

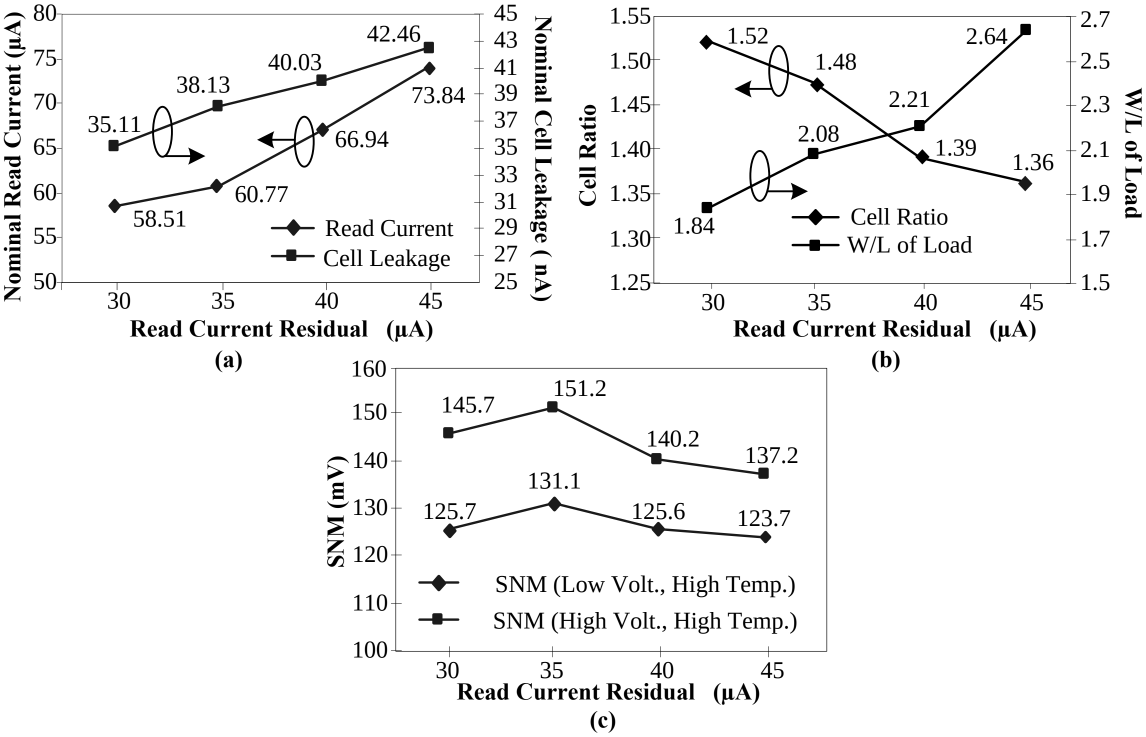


Figure 3.18: For varying read current residuals: (a) nominal read current (μA) and cell leakage (nA) (b) cell ratio and W_{ld}/L_{ld}, and (c) nominal SNM (mV)

requirement, and can be achieved with a small area, as can be see in Table 3.3. Smaller transistor widths at $Iread_{min} = 30\mu A$, allow for a low nominal leakage value and the $Ileak_{max} = 60nA$ bound is not violated for up to 14.536 sigmas of the within die variation due to RDF, at the worst-case global corner (column 5 of Table 3.3). However, as $Iread_{min}$ target is increased, the nominal cell leakage also increases (approaches $Ileak_{max}$) and fewer sigmas of within-die leakage variation can be tolerated by the design.

A similar trend is observed for the read current N_σ values (column 4). For a small $Iread_{min}$, the chosen nominal design can afford to have both, $Iread_{nominal}$ and $Ileak_{nominal}$, at a safe N_σ distance from their respective bounds. However, with an increase in the $Iread_{min}$, the $Iread_{nominal}$ has to be designed to be closer to $Iread_{min}$, so that the $Ileak_{nominal}$ value does not increase too much. Therefore, the N_σ value for the read current also reduces progressively, with an increase in $Iread_{min}$.

Fig.3.18 (b) shows that the cell ratio β reduces with the increasing read current residuals. A higher read current can be achieved by sizing up the driver and/or the access transistor. But driver transistors contribute heavily to total leakage. As a result, the optimizer increases the strength of the access transistor, more than that of the driver, which reduces β progressively. Another reason for this, is the chosen layout topology, in which driver and access transistors are placed parallel to each other (see Fig. 3.3). Therefore, for a chosen large driver, the width of the access transistor can be increased to some extent without impacting the bit-cell area (see (3.1) for $x1$ and $x2$).

It is expected that reducing the cell ratio would have a detrimental effect on the nominal SNM value. However, Fig.3.18 (c) indicates that the SNM, for the resultant optimal design solutions, degrades by only a few mV with the falling β. The plot of W_{ld}/L_{ld} in Fig.3.18 (b) explains this observation. A stronger PMOS not only results in a stronger "1" at VR, but also increases the gate drive of M1 for a stronger "0". This improves SNM as well as the read current.

Fig.3.19 depicts the variation in σ_{Vth} of the bit cell transistors. Because of the smaller channel area of the load transistor, σ_{Vth} of the load has a much higher absolute value than that of the driver or access transistors. Therefore, even though the SNM sensitivity to the V_{th} variation in the load transistor is relatively small (Fig.3.5), a reduction in the σ_{Vth} of the load helps to reduce σ_{SNM} (from equation 3.6). Hence, increasing the strength of the load transistor helps to mitigate the decline in the average SNM value as well as the SNM N_σ (column 6). It is evident, that the proposed problem formulation works well to provide robust design solutions.

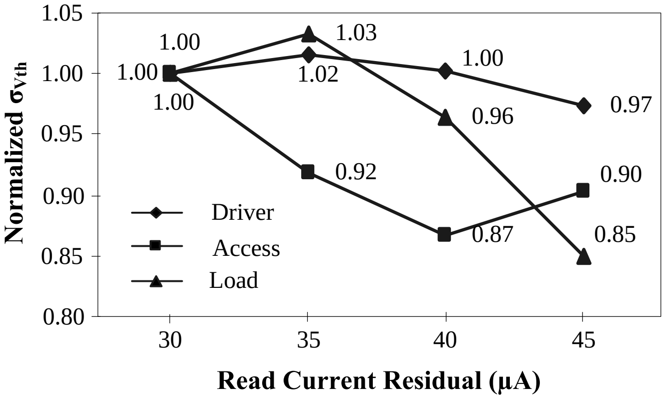


Figure 3.19: Variation of σ_{Vth} of driver, access and load transistors (normalized) with increasing read current requirement

3.5.2 General purpose-low leakage design

Table 3.4: Design of low leakage - moderate performance bit-cell, $Iread_{min} = 10\mu A$, $Ileak_{max} = 25nA$.

W_{drv}, W_{ax}, W_{ld} L_{drv}, L_{ax}, L_{ld} **(nm)**	149, 174, 175, 75, 79, 54,
Area (μm^2)	0.4405
Yield (%)	96.245

A general purpose bit-cell, on the moderate performance-low leakage side, is also designed. Tables 3.4 and 3.5 summarize the results. The resultant transistor sizes defy conventional sizing strategy for the bit-cell. The low leakage requirement ($Ileak_{max} = 25nA$) drives the optimizer to reduce the size of the driver significantly. This, however, considerably increases σ_{Vth} of the driver transistor and thus, σ_{SNM}, which can potentially degrade the SNM yield unless the average SNM value is also raised. Consequently, the load transistor is sized to be the largest to achieve a nominal SNM value of 152.2mV (Table 3.5), which is larger than those depicted in

Table 3.5: Design of low leakage - moderate performance bit-cell, $Iread_{min} = 10\mu A$, $Ileak_{max} = 25nA$.

Design Constr.	**Nominal Value**	N_σ
Ileak	15nA	4.7773
Iread	22.021 μA	4.9708
SNM(LV)	152.2mV	4.8280
SNM(HV)	177.2mV	4.9351
Vtrip	226.7mV	6.6629

Fig.3.18 (c).

It is difficult to analyse all the trade-offs and manually arrive at these transistor sizes. These results establish the benefits of the proposed method. A generic formulation of the bit-cell design optimization problem is presented here. To take care of the specific foundry guidelines, to minimize the layout-induced variations, different variables and additional constraints can be introduced. For example, for the horizontal poly-silicon gate of the driver and load transistor (Fig. 3.3), the designer can keep $L_{drv} = L_{ld} = L_{inv}$ in order to get rid of the awkward poly-shape (L-shape) which will result because of different lengths for the driver and the load. These modifications in the problem formulation, to a great extent, would depend on the chosen bit-cell layout topology and the extent of available advanced lithography correction mechanisms.

Fig.3.20 shows the variation of the Monte-Carlo yield in the neighborhood of the resultant optimal low leakage design solution. *snml* and *snmh* refer to SNM at low voltage, high temperature and high voltage, high temperature conditions, respectively as per Table 3.1. Running Monte-Carlo simulations to explore the entire search space to find the globally maximum yield point is infeasible. Therefore, in Fig.3.20, it is verified that the proposed optimization approach leads to,

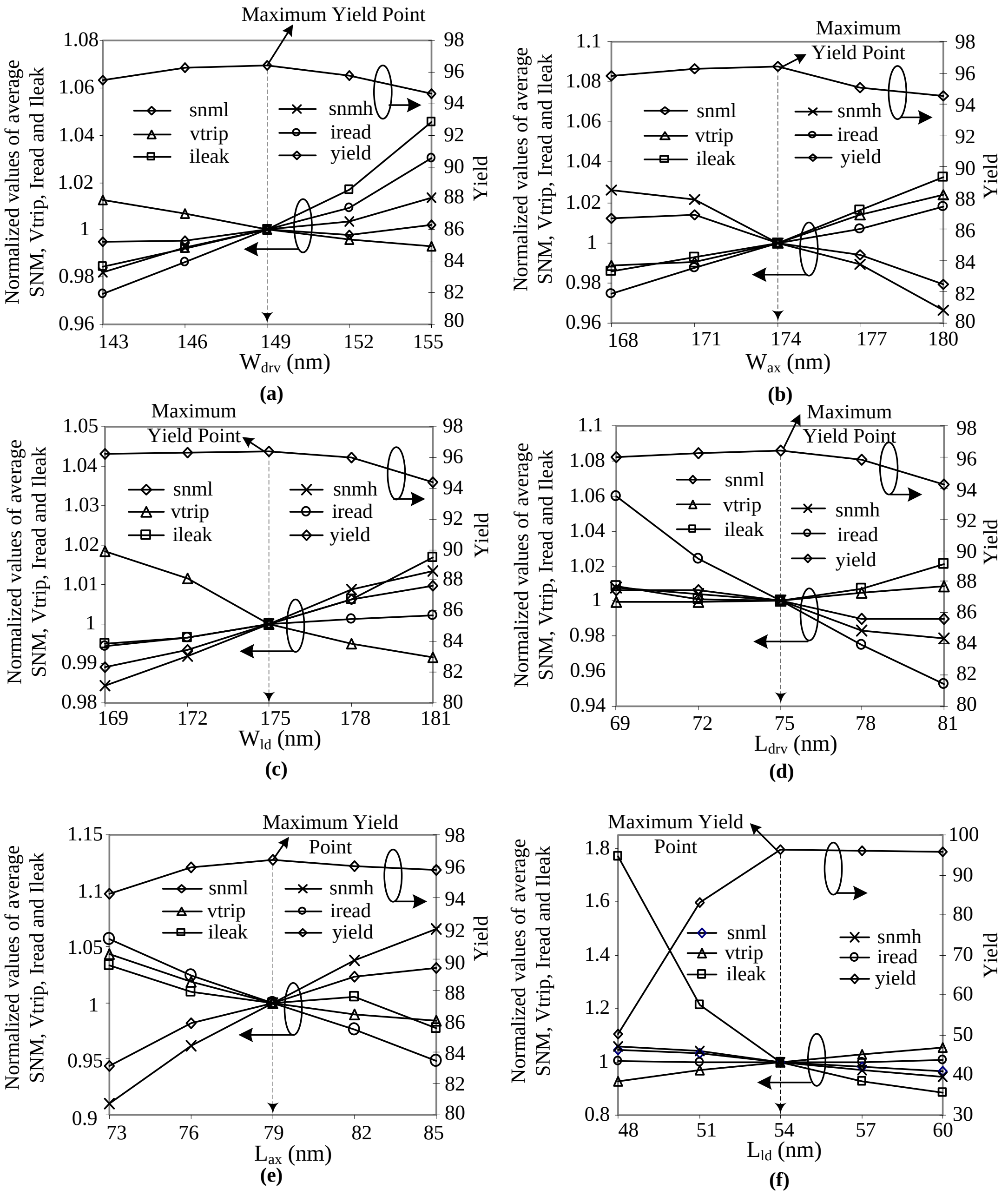


Figure 3.20: Yield and average SNM, Vtrip, I_{read} and leakage obtained by MC sims., in the neighborhood of the optimal low leakage design. In every figure, only one of the design parameters is varied.

at least, a locally maximum yield. In each sub-figure in Fig.3.20, one design parameter is varied, while the others are kept constant at the values mentioned in Table 3.4. For example, Fig.3.20 (a) shows the Monte Carlo yield and the average values of the design constraints when W_{drv} is varied from 143nm to 155nm, while $W_{ax} = 174nm, W_{ld} = 175nm, L_{drv} = 75nm, L_{ax} = 79nm$ and $L_{ld} = 54nm$. It is evident that the yield degrades as the design point is moved away from the obtained optimum.

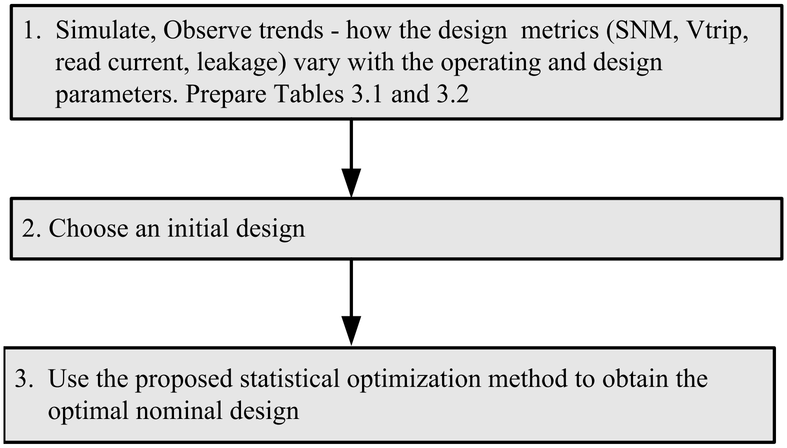

1. Simulate, Observe trends - how the design metrics (SNM, Vtrip, read current, leakage) vary with the operating and design parameters. Prepare Tables 3.1 and 3.2
2. Choose an initial design
3. Use the proposed statistical optimization method to obtain the optimal nominal design

Figure 3.21: Proposed Bit-cell Design Method

The flowchart in Fig. 3.21 outlines the proposed methodology. Each step is compared with the corresponding step in Fig. 3.1. In Step 1 of the proposed method, we need only trends, which derive Tables 3.1 and 3.2. We do not need an exhaustive database of actual values as in Step 1 of Fig. 3.1. Moreover, most of the trends are known and are expected to remain the same with technologies, e.g. SNM improves with reduction in access transistor width.

Step 2 in Fig. 3.21 involves selection of an initial nominal design. Again, note that this is only an initial guess for the proposed optimization framework and the final solution would be much different from this. This is unlike Step 2 in Fig. 3.1, where the chosen design would also be the final design, and therefore, must satisfy the specifications for all the design metrics. Step 3 of the proposed method runs the optimization routine to derive the optimized sizes for the bit-cell transistors.

The benefits of using the proposed method can now be seen. The entire exercise of iterative guessing and transistor size selection is eliminated. Steps 1, 2 and 3, in Fig. 3.1, are all time consuming and today, the SRAM bit-cell design takes about 3-4 weeks or more because of the iterations. Also, the chosen sizes are not optimal; there may be an over-design with larger bit-cell area. In the proposed method, steps 1 and 2 do not take much time and one can have the optimal bit-cell design in a day or two. Please note that Monte-Carlo simulations are done in the results section to verify the modeling procedure. This is done to prove the goodness of our method to the reader. The proposed method incorporates the impact of variability upfront, during the design phase by suitably modeling variability; therefore, no iterative Monte Carlo simulations are needed.

3.6 Summary

In this chapter, a statistical method to design the SRAM bit-cell is proposed. The method accounts for the manufacturing variability in the transistor dimensions, as well as the intrinsic V_{th} variations due to RDF. In addition, the widths and lengths of the transistors are chosen so as to satisfy the constraints of static noise margin, write trip voltage, read current, cell leakage and area. The developed method is flexible, involves a small initial infrastructure in terms of mathematical computations, and uses readily available models and tools in the industry, so that the extent of approximation in the proposed method is small. Robust bit-cell designs for high performance-moderate leakage and moderate performance-low leakage have been developed and analysed to demonstrate the working of the proposed method. It is concluded that conventional bit-cell sizing is not sufficient to ensure a low leakage optimal bit-cell design. The optimality of the resultant design is also demonstrated.

Chapter 4

Impact of Technology Scaling on SRAMs

For last three decades, achieving a 50% reduction in the SRAM bit-cell area has been the most important specification for the SRAM design, at every technology node. An aggressive scaling of the bit-cell area enables a higher density of the embedded SRAM. This, in turn, increases the data storage and manipulation capacity on the chip, which enables integration of more and more functional blocks, for a system-on-chip (SoC) design. This trend of aggressive scaling of the SRAM bit-cell has kept pace with, and sometimes exceeded, the expectations. However, in the nanometer regime, the bit-cell noise margins and leakage worsen due to lower supply and threshold voltages. Moreover, increasing variability makes it more difficult to ensure a satisfactory yield. In this chapter, the impact of technology scaling on the SRAM bit-cell is explored. In particular, the following are derived/analysed:

- The problem formulation of the statistical design method, proposed in Chapter 3, is improved. The definition of the performance constraint is revised, to better estimate the read speed of the SRAM. A high-yielding and optimal bit-cell design in the 65nm technology is derived and analysed.

- Optimal bit-cell designs in the 45nm and 32nm technologies are derived to study the impact of technology scaling on the bit-cell area and yield. It is explored, if it is possible to meet the common industry expectation of 50% scaling of the bit-cell area, and still ensure a high yield.

- Reasons for less than expected scaling of the bit-cell area are investigated. Two ways to improve the bit-cell scaling are proposed: memory partitioning and longer transistor lengths. It is demonstrated that progressively longer than nominal (e.g., nominal transistor length in the 45nm technology is 45nm) transistor lengths improve the bit-cell stability and allow narrower widths; therefore, a falling cell ratio can provide close to a 50% area scaling.

- Another design concern is investigated - With technology scaling, which failure mechanism, read stability (SNM), writability (Vtrip), performance (read current) or leakage; becomes the dominant cause of SRAM yield degradation? It is observed from the results that the 65nm optimal design is governed by leakage, the 45nm optimal design is governed by performance and the 32nm optimal design is governed by the SNM failure mechanism.

- The impact of voltage scaling on the bit-cell designs, in different technologies, is investigated. It is demonstrated that if the required performance is not relaxed with voltage scaling, then the area of the resultant optimal designs increases for all technologies. If the performance requirement is relaxed, then for the 65nm and the 45nm, reasonable bit-cell area scaling can be obtained. But for the 32nm optimal design, the SNM degrades severely with voltage scaling and necessitates a larger area, even for relaxed performance.

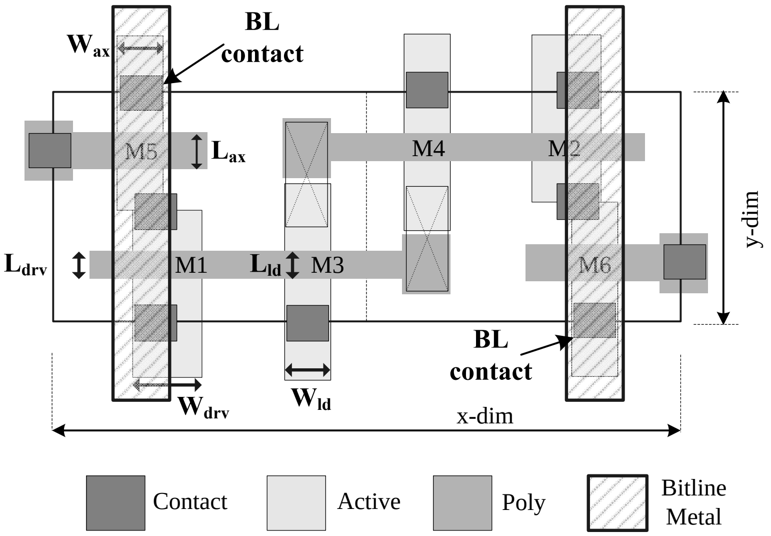


Figure 4.1: Bit-cell layout topology with bit line contacts and metal lines.

4.1 Improved Statistical Bit-Cell Design Method

4.1.1 Performance Constraint

In large SRAMs, with a large number of rows, such as 512 or 1024, the capacitance of the bit line is quite large, because it runs across all the rows of a memory block. Before any read or write operation, the bit lines are precharged to logic "1". During the read operation, the bit line discharges through the driver and the access transistor stack , to read "0". In other words, the bit line with a huge capacitance, is driven by the small bit-cell. The bit line capacitance is determined by the interconnect capacitance of the metal lines and the diffusion capacitances at the bit line contact with the access transistors, as depicted in Fig. 4.1. The interconnect capacitance is a function of the y-dimension of the bit-cell area. The diffusion capacitances are a function of the width of the access transistor. An improvement in the read current by increasing the strength of the access transistor, and other adjustments in the transistor sizes can cause an increase in the bit line capacitance. Therefore, the read current normalized by the bit line capacitance is a more effective means of measuring the read performance of the bit-cell.

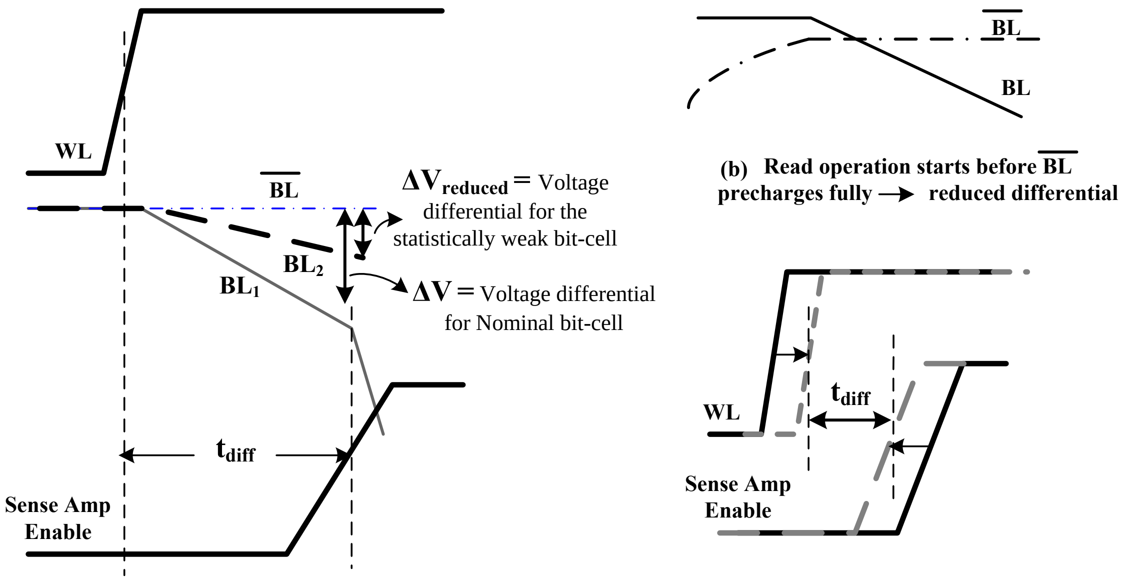
As explained in Chapter 2, large memories employ sense amplifiers to achieve a faster read access time. To a large extent, the read access time is governed by t_{diff} - the time required to develop a certain differential voltage between the bit-lines, as one of them discharges during the *read* operation [21].

$$t_{diff} = C_{BL} \ \times \Delta V / I_{read} \Rightarrow I_{read}/C_{BL} = \Delta V / t_{diff} \tag{4.1}$$

In equation 4.1, C_{BL} is the bit-line capacitance, I_{read} is the read current through the access M5 and driver M1 in Fig.3.2(a), and ΔV is the differential voltage required for read sensing. This differential voltage is amplified by the sense amplifier to achieve a full-logic swing on the read output. The nominal ΔV can degrade because of many reasons, e.g. sense amplifier V_{th} mismatch, degradation of I_{read} because of the intrinsic V_{th} fluctuations in the bit-cell transistors, incomplete precharge of the bit lines and on-chip variation in the arrival times of the word-line and the sense amplifier enable signals. Some of these are shown pictorially in Fig. 4.2. The minimum ΔV, required for a correct read sensing, should compensate for all these effects, within the differential build-up time t_{diff}, so as to ensure a sufficient input differential at the sense amplifier. Therefore, the minimum rate of differential build-up - $\Delta V / t_{diff}$, e.g. 10mV/100ps, can be used as the performance metric for the SRAM bit-cell [21].

It can be observed from equation 4.1, that the normalized read current (I_{read}/C_{BL}) is equal to the rate of differential build-up. Hence, the new read performance constraint, in the problem formulation for the statistical bit-cell design, is

$$\frac{Iread_{min}}{C_{BL}} = \frac{(Iread_{avg} - N_{\sigma} \times \sigma_{Iread})}{C_{BL}} \geq S_{differential} \tag{4.2}$$


WL
$\overline{BL}$
BL_1
BL_2
$\Delta V_{reduced}$ = Voltage differential for the statistically weak bit-cell
ΔV = Voltage differential for Nominal bit-cell
t_{diff}
Sense Amp Enable
$\overline{BL}$
BL
WL
t_{diff}
Sense Amp Enable

(a) Read current of the statistically weak cell degrades because of intrinsic Vth variations → Reduced BL slope → reduced differential

(b) Read operation starts before $\overline{BL}$ precharges fully → reduced differential

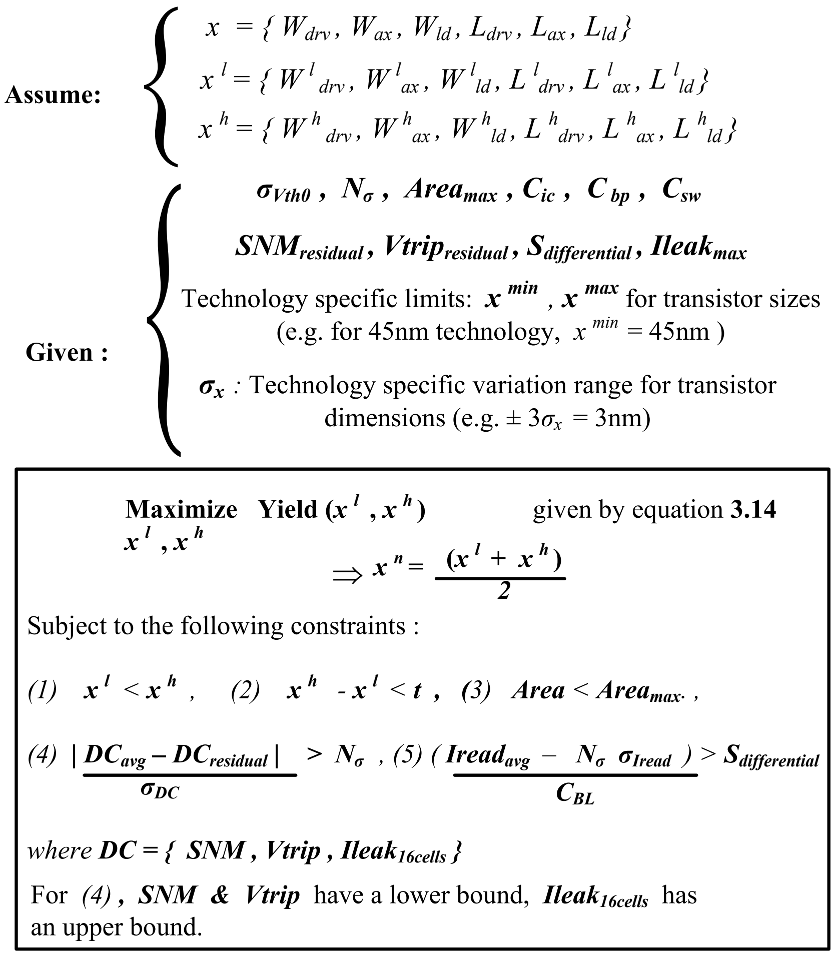
(c) Variation in the arrival times of WL and Sense Amp enable signals → Reduced t_{diff} → reduced differential

Figure 4.2: Reduction in the read differential voltage due to (a) degradation in the read current (b) incomplete precharge (c) variation in the signal arrival times.

$S_{differential}$ is the desired slope of the read differential build-up, for the statistically weakest cell, in mV/ps. The constraint formulation for all other design metrics - SNM, Vtrip, leakage and area, remains unchanged.

4.1.2 Revised Optimization Problem

The revised optimization problem is depicted in Fig. 4.3. To calculate C_{BL}, additional input parameters - interconnect capacitance per unit length (C_{ic}), diffusion bottom plate capacitance per unit area (C_{bp}) and sidewall capacitance per unit length (C_{sw}), are also required.



Assume:

$$\begin{cases} x = \{ W_{drv}, W_{ax}, W_{ld}, L_{drv}, L_{ax}, L_{ld} \} \\ x^{l} = \{ W^{l}{}_{drv}, W^{l}{}_{ax}, W^{l}{}_{ld}, L^{l}{}_{drv}, L^{l}{}_{ax}, L^{l}{}_{ld} \} \\ x^{h} = \{ W^{h}{}_{drv}, W^{h}{}_{ax}, W^{h}{}_{ld}, L^{h}{}_{drv}, L^{h}{}_{ax}, L^{h}{}_{ld} \} \end{cases}$$

Given :

σ_{Vth0} , N_σ , $Area_{max}$, C_{ic} , C_{bp} , C_{sw}

$SNM_{residual}$, $Vtrip_{residual}$, $S_{differential}$, $Ileak_{max}$

Technology specific limits: x^{min} , x^{max} for transistor sizes (e.g. for 45nm technology, x^{min} = 45nm)

σ_x : Technology specific variation range for transistor dimensions (e.g. ± $3\sigma_x$ = 3nm)

Maximize **Yield (x^l , x^h)** given by equation **3.14**
x^l , x^h

$$\Rightarrow x^n = \frac{(x^l + x^h)}{2}$$

Subject to the following constraints :

(1) $x^l < x^h$, *(2)* $x^h - x^l < t$, *(3)* $Area < Area_{max\cdot}$,

(4) $\dfrac{|DC_{avg} - DC_{residual}|}{\sigma_{DC}} > N_\sigma$, *(5)* $\dfrac{(Iread_{avg} - N_\sigma \ \sigma_{Iread})}{C_{BL}} > S_{differential}$

where $DC = \{ SNM, Vtrip, Ileak_{16cells} \}$

For *(4)* , ***SNM*** **&** ***Vtrip*** have a lower bound, $Ileak_{16cells}$ has an upper bound.

Figure 4.3: Revised Optimization problem

4.2 Results and Discussion

With the improved optimization method, optimal bit-cell designs are derived with the following inputs:

- $N_\sigma = 4.763$, which corresponds to only a single cell failure in an array of 1024×1024 cells - a yield of 99.9999%.
- $Iread_{min}/C_{BL,1024cells} = Iread_{residual}/(1024 \times C_{BL,single-bit-cell}) \geq 15mV/100ps$

, for a bit line spanning 1024 rows. With this, the read current of the statistically weakest cell (for 1024 rows) generates the read differential voltage of 60mV in 400ps [21].

- The maximum leakage bound of 16 cells is set to 640nA. The maximum leakage of a single cell can be more than 640/16 = 40nA, because of the exponential variation of leakage with V_{th}. But the average maximum leakage of a single cell is 640/16 = 40nA, as achieved in [80].
- The nominal voltage for the 65nm design is taken to be 1.1V [6].
- The 3σ variation in the transistor width and length is taken as 3nm.
- By using a Pelgrom coefficient of $3mV\mu m$ [81], σ_{Vth0} -standard deviation of the V_{th} distribution due to RDF for the smallest transistor (input parameter available in the vendor process kits) for the 65nm design is taken as 35mV.

All the constraints are evaluated at the respective worst-case voltage-temperature corners, as explained in Table 3.1.

4.2.1 Statistical Bit-cell Design (65nm technology)

With the afore-mentioned performance and leakage requirements, and by varying the maximum allowable area -$Area_{max}$ constraint, the optimized solutions for the 65nm technology are obtained. The actual areas of the resultant optimal bit-cell designs and the corresponding yield are displayed in Fig. 4.4 (a). For each of the resultant designs, the corresponding N_{σ} - the maximum tolerable sigma variation, in SNM, Vtrip and leakage are displayed in Fig. 4.4(b). The secondary axis of this figure shows the read differential build-up slope of the statistically weakest bit-cell. The yield in Fig. 4.4(a) and all the quantities in Fig.4.4(b) are obtained by MC

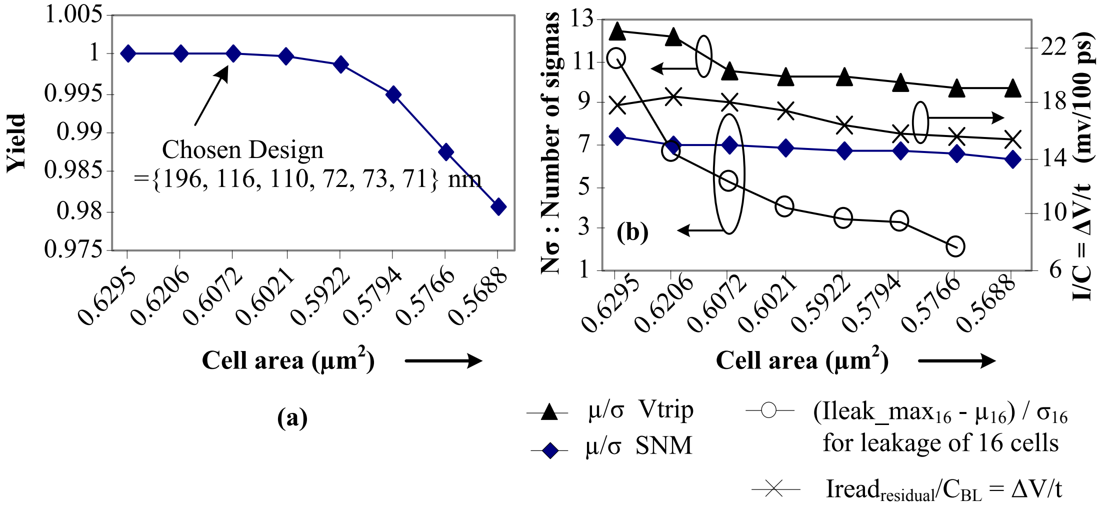


Figure 4.4: Simulation results for the 65nm design. With varying cell area, variation of (a) yield (b)$Iread_{residual}/C_{BL}$ and N_σ for SNM, Vtrip and leakage

simulations with 20000 points at the resultant design solutions (transistor width, length and V_{th}, all varied simultaneously).

Fig. 4.4(a) indicates that a relaxed area constraint allows larger transistors, which reduces σ_{Vth}, and the intra-die variations in the design metrics (e.g. σ_{SNM} in case of SNM), thereby giving a good yield. The chosen nominal design in Fig. 4.4(a), is the optimal design of minimum area, for which all the design constraints are satisfied ($N_\sigma \geq 4.763$). Fig. 4.4(b) shows that, for the chosen nominal design, N_σ of the leakage constraint is 5.17, but falls below 4.763, if $Area_{max}$ is further reduced. The N_σ of the design metrics -SNM and Vtrip remains above the required 4.763 and the read slope remains above $15mV/100ps$, even when the bit-cell area is reduced beyond the chosen point. Therefore, the 65nm optimal design is governed by the leakage yield.

The trends reflected in Fig.4.4 (b) and Fig.4.5 provide interesting insights. Reducing $Area_{max}$ forces a drastic reduction in the width and length of the driver transistor, as shown in Fig. 4.5(a) and (b). The width of the access transistor also

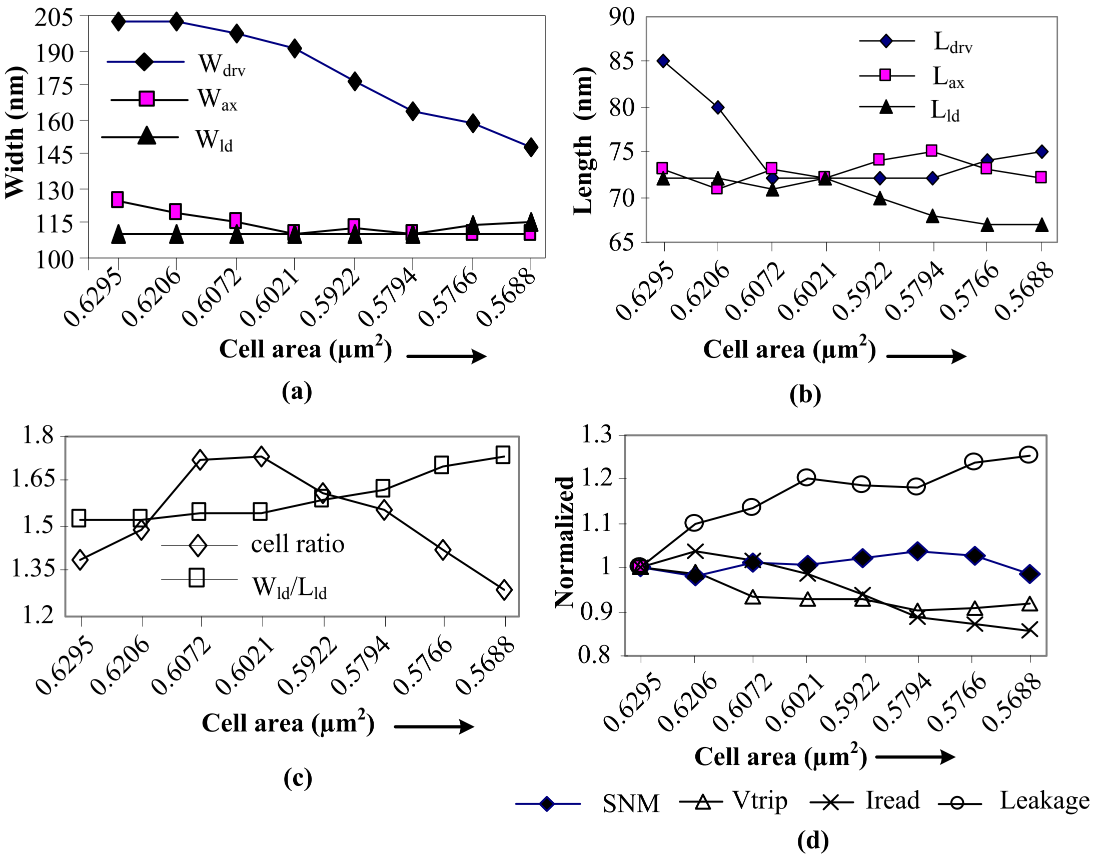


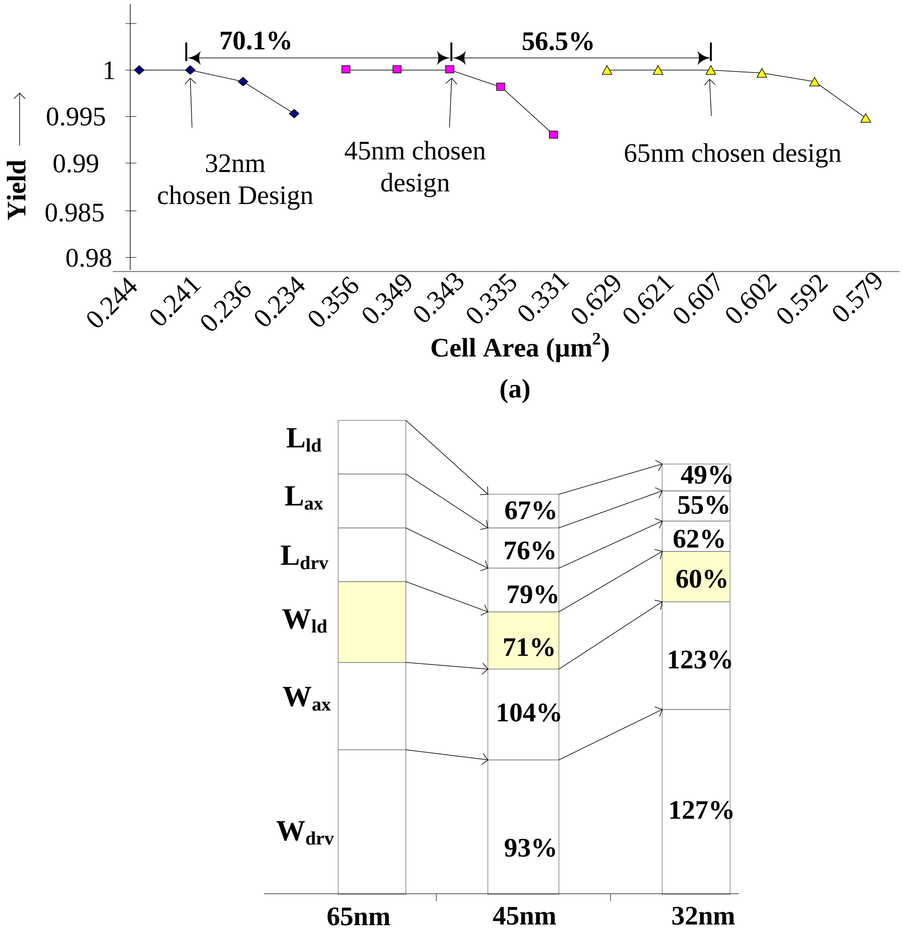
Figure 4.5: Simulation results for the 65nm design. With varying cell area, (a) transistor widths (b) transistor lengths (c) cell ratio and W_{ld}/L_{ld} (d) normalized average values of design constraints

becomes smaller. However, the reduction in the driver length, more than offsets the reduction in the widths of the driver and access transistor, which increases average leakage, as observed in Fig.4.5 (d). The observation that the decrease in the driver length, more than offsets the decrease in the widths, can be deduced from the initial increase in the cell ratio (=$(W_{drv}/L_{drv})/(W_{ax}/L_{ax})$), seen in Fig. 4.5 (c). The increase in the average leakage is accompanied by a sharp degradation in the N_σ for leakage (Fig. 4.4(b)), because of reducing transistor dimensions. This explains why the 65nm optimal design is governed by the leakage yield.

As seen in Fig. 4.5 (c), the cell ratio starts falling after increasing initially. This occurs when the driver length can no longer be reduced lest it would increase the σ_{Vth} too much. However, the driver width is continuously decreased to meet the smaller area specification. As a result, the cell ratio drops, as shown in Fig.4.5(c). This is accompanied by the rise in the W_{ld}/L_{ld} ratio. A stronger load mitigates the decline in the average SNM with the falling cell ratio, as depicted in Fig.4.5 (d). This prevents the SNM- N_σ from reducing too sharply, as portrayed in Fig.4.4 (b). The figure also demonstrates that the Vtrip-N_σ and the read current slope also show a steady decline with reducing area. Clearly, it is difficult to analyze all the trade-offs and manually arrive at the chosen design point - optimal design of minimum area, which satisfies the specifications for all the design metrics and is tolerant to variations.

4.2.2 Impact of Technology Scaling

The bit-cell leakage is expected to increase with technology scaling [6]. However, it is attempted to obtain optimal designs in the 45nm and 32nm, with the same maximum leakage requirement as that for the 65nm design, to impose a stricter design constraint. As a result, the required performance is also kept the same. The nominal voltage for the 45nm and 32nm designs is taken as 1.0V and 0.9V, respectively. The 3σ variation in the transistor width and length is taken as 3nm and 2nm, respectively. By using a Pelgrom coefficient of $3mV\mu m$ [81], σ_{Vth0} for the 45nm and 32nm designs are calculated to be 50mV and 70mV, respectively. For the bit line, constant per unit interconnect and diffusion capacitances are assumed for all technology nodes. A scaling ratio of 0.7 is assumed for all layout design rules.



(a)

(b)

Figure 4.6: (a) Area vs. yield trade-off for 65nm, 45nm and 32nm (b) Transistor sizes of the 45nm and 32nm optimal designs, compared to those in the 65nm design

The resultant designs for the 45nm and 32nm are compared with the 65nm design in Fig.4.6 (a). The area of the optimal bit-cell design in the 45nm technology is 56.5% of that in the 65nm technology. The area of the optimal bit-cell design in the 32nm technology is 70.1% of that in the 45nm technology. This does not meet the industry expectation of a 50% area scaling.

The reasons for the less than expected area scaling can be detected from Fig.4.6 (b), which depicts the transistor dimensions of the 45nm and 32nm optimal designs, relative to 65nm sizes. For instance, the width of the driver transistor in the 45nm and 32nm optimal designs, is 93% and 127%, respectively, of the corresponding 65nm value. To achieve an overall 50% area scaling, both the x-dimension and the y-dimension of the bit-cell should scale to 70% of the size in the previous technology ($0.7 \times 0.7 = 0.49 \approx 0.5$). Therefore, the expected ratio for all the dimensions in the 45nm design , with respect to the 65nm, is 70%. The expected ratio for all the transistor dimensions in the 32nm, with respect to 65nm, is around 50% ($0.7 \times 0.7 = 0.49 \approx 0.5$). However, it can be observed from Fig.4.6 (b), that the widths of the driver and the access transistors exhibit very poor scaling. Further analysis reveals that a wider access transistor is needed for higher read current, to meet the read speed specification. This, in turn, necessitates a wider driver (to maintain the cell ratio) to satisfy the SNM constraint. It can be concluded for the 45nm and 32nm designs that most of the scaling is obtained by scaling of the design rules, as the transistor dimensions do not scale as expected.

4.2.3 Achieving 50% Area Scaling - Longer Transistors and Partitioning

It has been observed in the previous section, that the resultant optimal bit-cell designs in the 45nm and 32nm technologies do not meet the industry expectation of 50% area scaling with respect to the 65nm optimal design. In this section, two approaches - longer transistor lengths and partitioning, are proposed, to improve the bit-cell area of the 45nm and 32nm optimal designs. The goal is to significantly improve the scaling of the driver and access transistor widths, to enable an overall better area scaling.

Longer transistors

Longer L_{drv} and L_{ax} are proposed, e.g. the maximum length for the 32nm design, which was previously set to 45nm, is increased to 52nm, which is 60% more than the nominal length of 32nm. Use of longer transistor lengths, to achieve better scaling of the bit-cell area, seems to be counter-intuitive. To understand this, the following observations from Fig.4.7 are made.

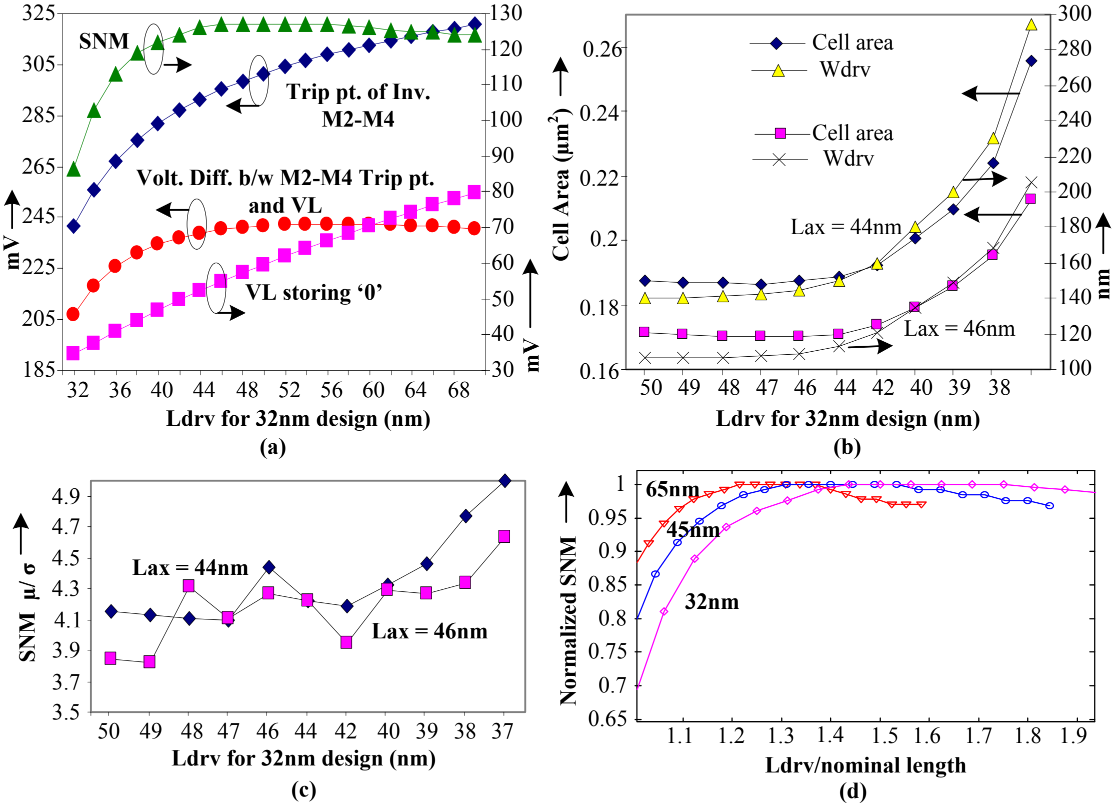


Figure 4.7: Average SNM and N_{σ} analysis with varying driver transistor length

- During the read cycle, the node VL rises to an intermediate voltage (weak "0", about 200-300mV) due to the voltage divider action between M1 and M5. If this voltage at node VL exceeds the trip point of the M2-M4 inverter, the inverter output at node VR can flip to logic “0” from logic “1”. This amounts to an unintentional change in the state of the bit-cell or a destructive read operation. Hence, the voltage difference between the trip point of M2-M4 inverter and the voltage at node VL, is an indicator of the bit-cell SNM [21]. The intermediate voltage at node VL during the read operation and the trip point of the inverter M2-M4, both depend on the strength of the driver transistor.

Fig.4.7(a) plots the variation of voltage at node VL, trip point of inverter M2-M4, the difference between these two quantities, and the SNM, with the variation in the length of the driver transistor for a 32nm design. A short L_{drv} strengthens both the driver transistors M1 and M2. However, as depicted in Fig.4.7(a), a very short L_{drv} reduces the M2-M4 trip point much more than the VL node voltage. This reduces the voltage difference between these two quantities (for $L_{drv} \leq 44nm$) and degrades the bit-cell SNM, even though $\beta = (W_{drv}/L_{drv})/(W_{ax}/L_{ax})$ improves with smaller L_{drv}. Therefore, as shown in Fig.4.7(b), very short L_{drv} necessitates a much larger W_{drv} to achieve the same SNM. E.g., the driver width, required to achieve 120mV SNM, increases from 150nm to 295nm as the driver length is reduced from 44nm to 37nm (for L_{ax} = 44nm). This increases the bit-cell area significantly. Hence, it can be concluded that the same bit-cell SNM can be achieved with a smaller driver width, and a smaller bit-cell area, if longer than the nominal driver lengths are employed. The area benefit can be enhanced further, by using longer access transistors, as evident from Fig.4.7(b).

- However, Fig.4.7(b) also indicates that the gain in the bit-cell area diminishes as the driver length is made too large. Additionally, Fig.4.7 (c), which plots the SNM- N_σ corresponding to the W_{drv} and L_{drv} in Fig.4.7 (b), shows that the SNM- N_σ degrades with increasing L_{drv}. This establishes that the W_{drv} cannot be made too small, even for longer L_{drv}, because this reduces N_σ

despite maintaining the same nominal SNM (120mV). These two observations imply that an optimal L_{drv}, longer than the nominal, would reduce the W_{drv} required to meet the SNM constraint.

- Fig.4.7(d) plots the variation in the normalized SNM with the variation in the L_{drv}, for bit-cell designs in 65nm, 45nm and 32nm technologies. The L_{drv} is normalized by the nominal length. The nominal length in the 65nm, 45nm and 32nm technologies is 65nm, 45nm and 32nm, respectively. The figure depicts that the required L_{drv} (relative to nominal) for the same SNM increases from 65nm to 32nm technology. This is because of the sharper V_{th} roll-off with the transistor length in the 45nm and 32nm technologies, which leads to a sharper degradation in the trip point of the M2-M4 inverter. All the above discussion builds up the case for employing progressively longer than nominal transistor lengths, as the technology scales, to meet the SNM specification in a smaller area.

Partitioning

Secondly, memory partitioning is proposed. For the previously described read speed constraint -$Iread_{residual}/(1024 \times C_{BL,single-bit-cell}) = \Delta V/t_{diff} \geq 15mV/100ps$, if the memory is partitioned to restrict the maximum number of rows to 512 for the 45nm design, the $Iread_{residual}$ can be halved, but the required slope of the statistically weakest cell remains fixed at 15mV/100ps. Similarly, the number of rows for the 32nm design is limited to 256. Partitioning of the memory reduces the required current during the read operation, and hence the required strength of the access transistor. This allows a narrower driver (for the same SNM), which again helps in achieving a smaller cell area. For example, it is depicted in Fig.4.7(b), that the W_{drv} required to achieve the same SNM, is smaller when $L_{ax} = 46nm$ than when $L_{ax} = 44nm$.

Results with longer transistors and partitioning

With the two modifications explained above, optimal designs in the 45nm and 32nm are derived for the same set of specifications for the performance and leakage, i.e. a differential build up slope of 15mV/100ps for the statistically weakest cell, and a maximum leakage of 640nA for 16 cells. The optimization results are presented in Fig. 4.8.

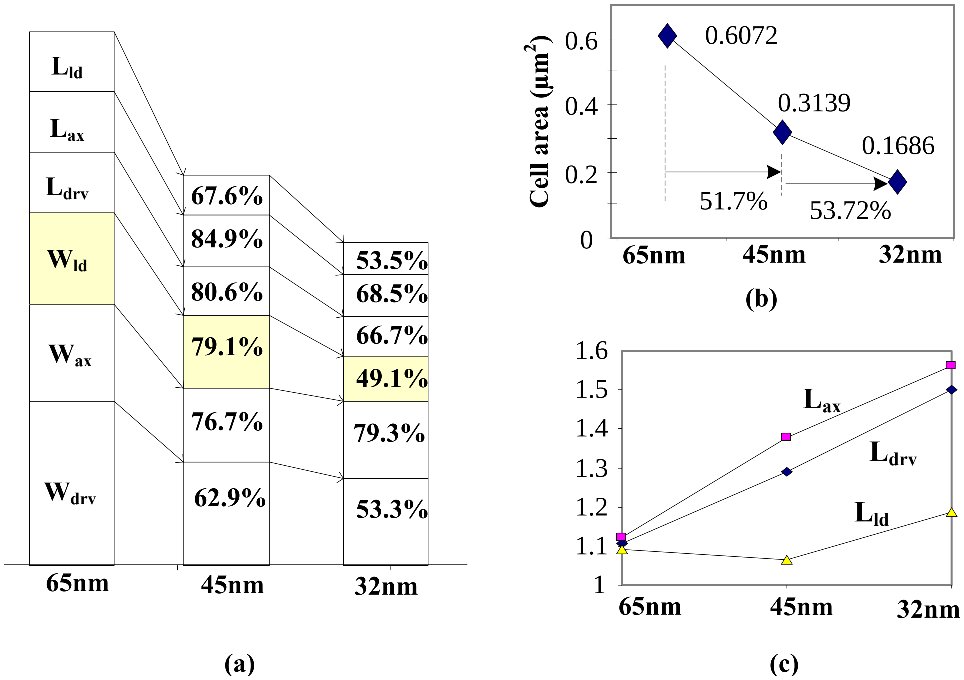


Figure 4.8: Optimization results with partitioning and longer lengths (a) transistor widths and lengths in the newly derived 45nm and 32nm optimal designs, relative to the optimal 65nm design obtained in Fig. 4.4 (b) bit-cell area (c) transistor lengths/nominal lengths.

Fig.4.8 (a) demonstrates the transistor dimensions in the newly derived 45nm and 32nm optimal designs, as a percentage of the corresponding dimensions in the 65nm optimal design (obtained in Fig. 4.4). Compared with Fig.4.6 (b), Fig.4.8 (a) shows that all transistor widths scale with respect to the corresponding widths in the 65nm optimal design. Therefore, the use of longer transistors and partitioning improves scaling. As expected, the lengths of the driver and access transistors do not scale as much. E.g., in Fig.4.8 (a), the lengths of the driver and access transistors in the 32nm design scale to 66.7% and 68.5% of their respective 65nm values. But in Fig.4.6 (b), the corresponding numbers are 62% and 55%, respectively.

Fig.4.8 (b) signifies that a much better bit-cell area scaling for the 45nm and 32nm designs (51.7% and 53.72%, respectively) is achieved compared to that in Fig.4.6 (a). The resultant scaling is satisfactorily close to expected scaling of 50%. Fig.4.8(c) plots the lengths of the transistors in the optimal bit-designs, relative to the respective nominal length. For example, the relative L_{drv} increases from 1.1 in the 65nm design to 1.29 in the 45nm design, and to 1.5 in the 32nm design. The computed cell ratio is **1.72, 1.49 and 1.19** in the 65nm, 45nm and 32nm designs, respectively. This is counter-intuitive and defies conventional belief that a progressively higher cell ratio (as high as 4) is needed [54] with technology scaling to ensure an acceptable SNM yield. The optimizer results confirm the initial assertion that progressively longer than nominal relative transistor lengths at scaled technology nodes, enable smaller widths and an overall better area scaling for high-yield designs. The use of longer transistors improves the leakage N_{σ}. The Vtrip N_{σ} degrades, but meets the minimum requirement.

Fig. 4.9 compares the area of the 45nm optimal designs in Fig. 4.6 (non-partitioned) and Fig. 4.8 (partitioned). The X and Y dimensions of the bit-cell layout, for both the cases, are also mentioned in Fig. 4.9. The area benefit per bit-cell, due to partitioning, for the 45nm design is $0.02919\mu m^2$. However, the column peripheral circuits are duplicated in the partitioned memory, and the area overhead due to the peripherals must be considered.

The column peripheral circuits are usually pitch-matched with four (or eight) bit-cells in a row, because the circuits like the sense amplifier employ big transistors,

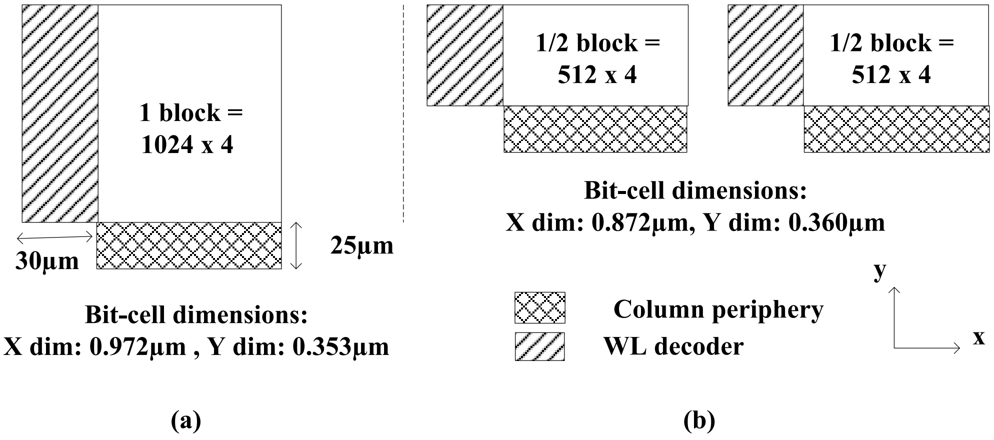


Figure 4.9: Area comparison of (a) non-partitioned, and (b) partitioned memory banks. The column periphery Y-dim and WL decoder X-dim are assumed to be $25\mu m$ and $30\mu m$, respectively.

which cannot be fitted in the pitch of a single cell. Considering a block of 1024 rows x 4 columns, as illustrated in Fig. 4.9, the array area gain per block in the partitioned memory, is $119.5\mu m^2$. Because the column periphery is duplicated in the partitioned memory, the column peripheral overhead per block would be $77.2\mu m^2$ (assuming a $25\mu m$ periphery height). Therefore, the net area benefit per block of 1024 rows x 4 columns, in the partitioned memory, is $42.3\mu m^2$.

The Y-dimension of the bit-cell increases in the partitioned case. Therefore, the word line decoder, which sits alongside the Y-dimension of the array, increases in size. Five array blocks (each providing an area benefit of $42.3\mu m^2$) should compensate for the increase of $215\mu m^2$ in the WL decoder area (assuming a $30\mu m$ decoder width). Moreover, the column periphery can be made shorter in the partitioned memory, because the write drivers can be smaller for the shorter bit lines. The assumptions for the periphery dimensions are industry estimates, based on laid out designs. A similar comparison for the 32nm design also shows that the area benefits of the partitioned design compensate for the duplicated peripheral overhead.

4.2.4 Which failure mechanism becomes more dominant with technology scaling ?

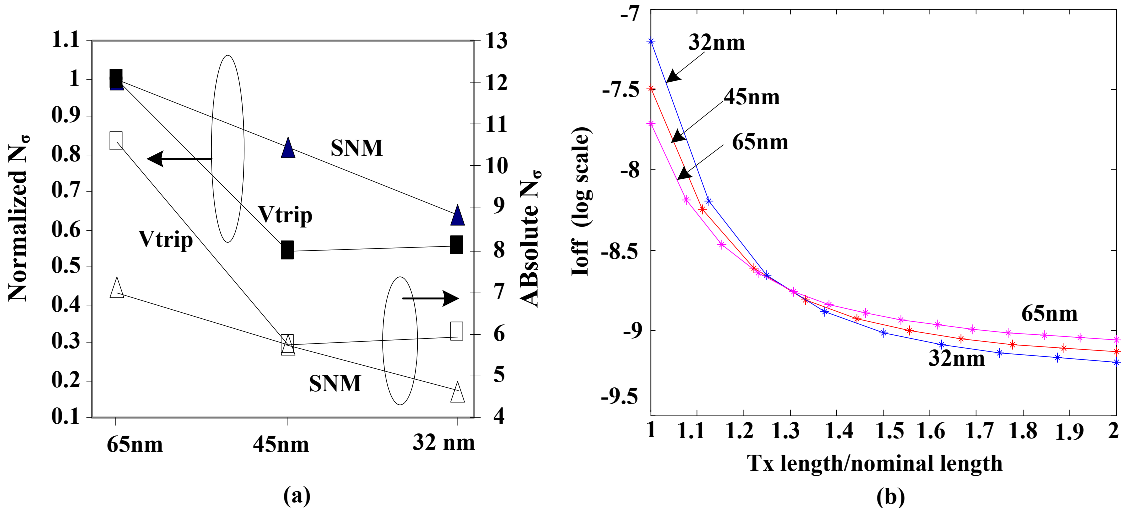


Figure 4.10: (a)Relative and absolute N_σ for SNM and Vtrip (with partitioning for 45nm and 32nm) (b) I_{off} of a transistor with min. width for varying lengths.

The scaling of the bit-cell area by 50% has been the most important SRAM design concern for over three decades. Therefore, the optimal designs in Fig. 4.8 that exhibit close to 50% scaling, are considered for the failure mechanism analysis. It has been explained in section 4.4 that the leakage N_σ determines the 65nm optimal design. Therefore, leakage is the most dominant failure mechanism observed in the 65nm optimal design. The leakage becomes less important for the 45nm and 32nm optimal designs, because these contain longer relative transistor lengths, which reduce the average leakage (Fig.4.10 (b)) and the leakage variation.

The differential build up slope of the statistically weakest cell, at the 45nm chosen design point, is 15.1 mV/100ps. With a further reduction in the allowable bit-cell area, this slope falls below the required value of 15mV/100ps, while the other design constraints are still satisfied. Therefore, the 45nm optimal design is determined by the read performance constraint.

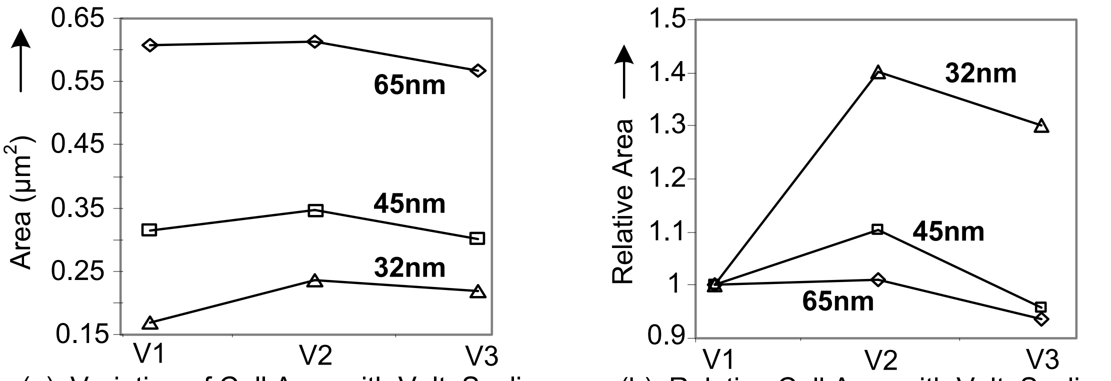
For the 32nm optimal design, the observed limiting constraint is neither performance, nor leakage. The relative N_σ of the functional design constraints (SNM and Vtrip) are plotted in Fig.4.10 (a). The Vtrip N_σ exhibits a sharp decline, but the SNM-N_σ is the lowest for the 32nm design . Therefore, read stability failure or SNM is the most dominant failure mechanism at the 32nm node and the 32nm optimal design is governed by the SNM constraint.

4.2.5 Impact of Voltage Scaling

Scaling the supply voltage reduces the dynamic power consumption at the chip level. Therefore, the IC industry has consistently reduced the chip supply voltages for successive technology generations. However, it is often not possible to operate memories at the same lower voltage as the rest of the chip, because of stability issues. In this section, the impact of voltage scaling on the bit-cell designs, at scaled technology nodes, is explored. Optimal bit-cell designs in the 65nm , 45nm and 32nm technologies are obtained with the proposed optimization framework and with the reduced nominal voltages of 1V, 0.9V and 0.8V, respectively [6]. Three cases are explored in the subsequent discussion:

- case V1: No voltage scaling, $I/C \geq 15mV/100ps$ (Designs in Fig 4.8).
- case V2: With voltage scaling, $I/C \geq 15mV/100ps$ (Maintain high speed at low voltage)
- case V3: With voltage scaling, $I/C \geq 10mV/100ps$ (Give up speed to achieve high density at low voltage)

Following are the results and observations:



(a) Variation of Cell Area with Volt. Scaling (b) Relative Cell Area with Volt. Scaling

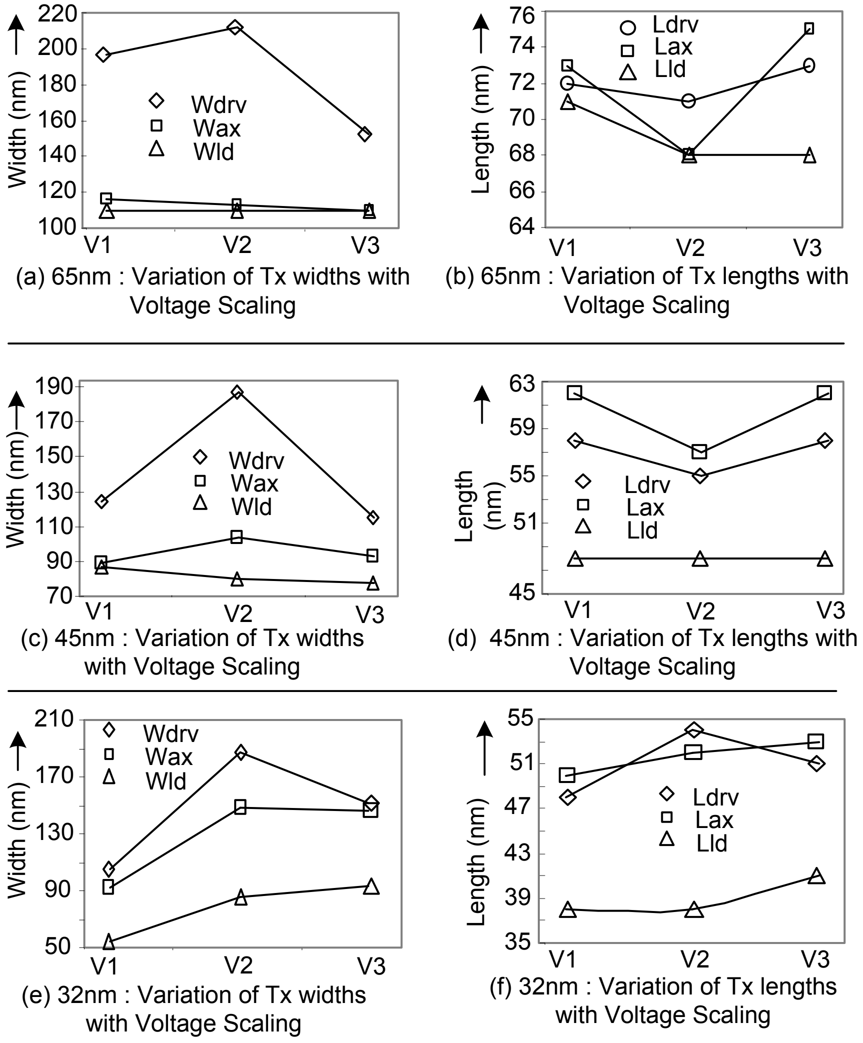
case V1 : Original Designs without Voltage Scaling, I/C >15mV/100ps
case V2 : Designs with Voltage Scaling , I/C >15mV/100ps
case V3 : Designs with Voltage Scaling , I/C >10mV/100ps

Figure 4.11: Area comparison between 65nm, 45nm and 32nm optimal bit-cell designs obtained at scaled voltages. Cases V2 and V3 are compared with case V1.

65nm

The 65nm design in case V1 is determined by the leakage N_σ, as explained in section 4.4. For case V2, a lower voltage reduces leakage, hence leakage no longer constrains the design. Therefore, the transistor lengths can be reduced (Fig 4.12(b)) and W_{drv} increased (Fig 4.12(a)), to achieve the same performance ($I/C \geq 15mV/100ps$) at lower voltage. The overall increase in area for the case V2, compared to case V1, in 65nm, is quite small, 1.1%, as displayed in Fig 4.11 (a) and (b).

For case V3, with a lower performance requirement of $10mV/100ps$, the W_{drv} can be reduced significantly, as shown in Fig 4.12 (a), while increasing the lengths of the driver and access transistors for SNM adjustment (Fig 4.12(b)). The area of the 65nm design -case V3 is 6.5% smaller than that of the original 65nm design -case V1 (Fig 4.11 (a)(b)). Therefore, for the 65nm technology, if the required performance requirement is relaxed, the supply voltage can be lowered, without sacrificing area. In fact, the bit-cell area can be improved to some extent.



(a) 65nm : Variation of Tx widths with Voltage Scaling

(b) 65nm : Variation of Tx lengths with Voltage Scaling

(c) 45nm : Variation of Tx widths with Voltage Scaling

(d) 45nm : Variation of Tx lengths with Voltage Scaling

(e) 32nm : Variation of Tx widths with Voltage Scaling

(f) 32nm : Variation of Tx lengths with Voltage Scaling

case V1 : Original Designs without Voltage Scaling, I/C >15mV/100ps
case V2 : Designs with Voltage Scaling , I/C >15mV/100ps
case V3 : Designs with Voltage Scaling , I/C >10mV/100ps

Figure 4.12: 65nm, 45nm, 32nm optimal bit-cell designs obtained at scaled voltages. Cases V2, V3 are compared with case V1.

45nm

The original 65nm design -case V1 is limited by the leakage constraint. The performance metric at the case V1-65nm design is 18.1 mV/100ps, when the specification is only 15mV/100ps. Therefore, there is extra read speed available in the original 65nm design. This explains why only 1.1% extra area is required in case V2, to maintain the same performance at a smaller supply voltage.

However, the original 45nm design -case V1 is limited by the performance constraint (15.13mV/100ps read slope), as mentioned in Section 4.2.4. Therefore, to achieve the same performance at a lower voltage in case V2, a larger percentage increase of 10.3% in the bit-cell area is required in the 45nm, as depicted in Fig. 4.11(b).

It is depicted in Fig. 4.10 that the value of the SNM-N_σ for the 65nm design (case V1) is sufficiently above the required value (4.763) and degrades for the 45nm design (case V1). Therefore, the SNM degradation due to a lower supply voltage would have a stronger impact on the 45nm design. Hence, even when the performance requirement is relaxed for a lower supply voltage, the bit-cell area is limited by the SNM requirement. As a result, the area benefit obtained in case V3, by reducing the performance requirement to $10mV/100ps$, is smaller (as compared to case V3 in 65nm), i.e. 4.2% as shown in Fig 4.11 (b).

32nm

The original 32nm design (case V1) is determined by the SNM constraint. It has been discussed in section 4.2.3 that the bit-cell area, required to meet the SNM requirement, can be kept reasonable, if longer transistors are used. Therefore, to meet the N_σ requirement for SNM at reduced voltage (case V2) in 32nm, the lengths of the driver and access transistors in the optimal design are longer as compared to those in case V1 (Fig 4.12(f)). Longer transistors necessitate much wider transistors to meet the read current requirement for case V2. This is shown in Fig 4.12 (e). The overall increase in the bit-cell area for case V2, is 40% (Fig 4.11 (b)). Because the 32nm designs are governed by the SNM (which degrades with reducing voltage), a reduction in the read current requirement in case V3, does not provide any area benefit compared to case V1. On the contrary, the cell area in case V3 is 30% higher than that in case V1.

4.3 Summary

In this chapter, optimal nominal bit-cells designs in the 65nm, 45nm and 32nm technologies are obtained with the revised statistical design method. The scaling of the bit-cell area is examined and following recommendations are made. Longer driver and access transistors should be used to reduce the variability; to improve the SNM and leakage and to enable better width scaling. Partitioning is an effective way to improve performance, achieve a better area scaling and to reduce the overall leakage (since unused banks can be disabled). It is demonstrated that memory partitioning and progressively longer than the nominal transistor lengths enable smaller transistor widths, therefore a falling cell ratio can provide close to a 50% bit-cell area scaling.

If the operating voltage is scaled and the performance requirement is not relaxed, the area of the optimal designs increases for all technologies. The increase in the cell area is the largest for the 32nm node and the smallest for the 65nm node. On the other hand, if the performance requirement is relaxed, the designer can lower the supply voltage for the memory without compromising on the bit-cell area for the 65nm and 45nm. The 32nm design is governed by the SNM, and therefore, voltage scaling inevitably requires a larger cell area, even with a relaxed performance requirement.

Chapter 5

Conclusions & Future Work

5.1 Conclusions

In this thesis, the focus is on the process variations and the design of SRAM array in nanometer technologies. The importance of a high-yielding SRAM, in today's integrated circuits, is demonstrated. At the same time, the challenges in the design of a robust SRAM array are described. The ever-increasing impact of variability on the characteristics of the SRAM bit-cell is explained. It is clearly established that statistical design methods are absolutely essential in the nanometer regime, to accurately account for the variability in the SRAM design metrics, in a systematic manner, right at the design stage. The SRAM has conflicting requirements such as the SNM, Vtrip, read speed and leakage, and even a deterministic design is quite involved. The design problem becomes even more challenging, with the impact of variability.

A novel statistical method to design the SRAM bit-cell, is proposed in this thesis. The method provides an optimization framework, which can be deployed to automatically derive the optimal bit-cell designs, for a given set of specifications for the design constraints such as the SNM, Vtrip, performance, leakage and area. The resultant designs are optimal, because they provide a high yield, in the smallest possible area. The method considers the inter-die manufacturing variations in the transistor geometrical parameters and the within-die intrinsic threshold voltage variations due to RDF. The benefits of using the proposed method vis-a-vis the current industrial SRAM design method are outlined. The proposed method is not only capable of reducing the design time drastically, but also provides optimal designs, which may not be possible with the current industrial approach. The benefits of the proposed method compared to other approaches in the literature, are also explained. The proposed framework is developed, with the needs of the circuit designer in mind. Therefore, standard models for the transistors and standard simulators are integrated in the proposed framework. This obviates the need for any analytical modeling for the transistor currents or SRAM design metrics. Therefore, the proposed method is also practical and readily usable. Analytical modeling is used only for variability and yield estimation. The accuracy of the modeling is also established in the relevant sections.

The proposed method is technology scalable. It can be adapted for additional design constraints, and is tuneable according to the specifications. Several bit-cell designs in the 65nm, 45nm and 32nm technology nodes are derived with the proposed method. The optimality of the design yield is verified by Monte Carlo simulations. The trends in the results are studied to understand the working of the proposed method. To achieve 50% scaling of the bit-cell area, two suggestions are made - progressively longer than nominal transistors and partitioning. The scaling benefits, with these modifications in the optimization framework, are shown to be satisfactory. Furthermore, the failure mechanisms of the bit-cell at different technology nodes, and the impact of voltage scaling with different performance requirements, are also analysed.

5.2 Future Work

In future, the proposed method can be enhanced to incorporate the impact of other second order sources of variability, such as the line-edge roughness and the oxide thickness variation. It needs to be understood, how these sources of variability impact the transistor electrical characteristics, and how this can be modeled accurately. As more sources of variability are incorporated, it needs to be kept in mind that not all of them occur independently of each other. In order to make sure that the designs are not pessimistic, the correlation between the various sources of variability also needs to be considered. In addition, the method can be used to design other newer topologies of the SRAM bit-cell, such as the 8T version. Making the supply voltage a design parameter can also be experimented with. It would also be informative to fabricate arrays with these optimal bit-cell designs and compare the simulated yield with the silicon yield.

Bibliography

[1] R.H. Dennard,“Evolution of the MOSFET Dynamic RAM - a Persoanl view,” *IEEE Transactions on Electron Devices*, vol. ed-31, no. 11, Nov. 1984. 1, 3

[2] R.H. Dennard,“Field-effect transistor memory,” US patent 3 387 286, June 4, 1968. 1

[3] H. G. Cragon,*Memory Systems and Pipelined Processors*, chapter 1, Jones and Barlett Publishers. 1

[4] J. L. Hennessy, D. A. Patterson, *Computer Architecture : A Quantitative Approach*, chapter 5, Morgan Kaufman, 2006. x, 1, 2

[5] J. Wuu, D. Weiss et al., “The asynchronous 24MB on-chip level-3 cache for a dual core Itanium-family processor,”*IEEE International Solid State circuits conference*, 2005. x, 5, 6

[6] International Technology Roadmap for Semiconductors, *http://www.itrs.net/*. 6, 49, 50, 91, 94, 104

[7] P. Gelsinger, “Microprocessors for the New millennium: Challenges, Opportunities and New Frontiers,”*IEEE International Solid State circuits conference*, 2001. x, 6, 7

[8] G. Moore, “Progress in Digital Integrated Electronics,”*IEDM*, 1975. 6

[9] R. Yung, S. Rusu, K. Shoemaker, "Future Trend of Microprocessor Design,"*European Solid State Circuits International Research Conference (ESSCIRC)*, 2002. x, 6, 8, 9

[10] J.D. Plummer, "Material and Process limits in Silicon VLSI Technology,"*Proceedings of the IEEE*, vol. 89, March 2001. 7, 8

[11] D.J. Frank, E. Nowak, H.P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies,"*Proceedings of the IEEE*, vol. 89, March 2001.

[12] J.M. Rabaey, A. Chandrakasan, B. Nikolic *Digital Integrated Circuits: A Design perspective*, Prentice Hall, pp. 125-130. 8, 13, 15, 76

[13] M. Kanda, E. Morifuji et al., "Highly Stable 65nm Node(CMOS5) 0.56 μm^2 SRAM Cell Design for Very Low Operation Voltage,"*Symposium on VLSI Technology Digest of Technical papers*, 2003. 10

[14] B. Prince, *High Performance Memories: New Architecture DRAMs and SRAMs*, Jon Wiley & sons, 1999, chapter 1. 13

[15] Y. Tsiatouhas et al., "New Memory sense amplifier designs in CMOS technology,"*Proceedings of the IEEE*, 2000. 15

[16] Yi-Ming-Sheng et al., "A Measurement Unit for Input Signal Analysis of SRAM Sense Amplifiers,"*Proceedings of the Asian Test Symposium, IEEE*, 2004.

[17] Bernhard Wicht et al., "A Yield Optimized Latch-type SRAM Sense Amplifier,"*ESSCIRC*, 2003. 15

[18] E. Seevinck et al.,"Static Noise Margin Analysis of MOS SRAM Cells," *IEEE JSSC*, Oct. 1987, pp. 748-754. 19

[19] Jan Lohstroh, E. Seevinck,"Worst-Case Static Noise Margin Criteria for Logic Circuits and their Mathematical Equivalence," *IEEE JSSC*, Dec. 1983, pp. 803-806. 19, 20, 22

[20] A. Bhavnagarwala et al., "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability," *IEEE JSSC*, April 2001, pp. 658-665. xi, 22, 36, 45, 50

[21] R. Heald, P. Wang, "Variability in Sub-100nm SRAM Designs," *IEEE/ACM ICCAD*, 2004, pp. 347-351. 22, 24, 43, 58, 88, 91, 98

[22] D. Redwine, "SRAM cell with Independent Static Noise Margin, Trip voltage and Read current Optimization," *U.S. Patents*, 2005. 22

[23] K.Zhang, U. Bhattacharya, et.al, "SRAM Design on 65-nm CMOS technology with Dynamic Sleep Transistor for Leakage Reduction," *IEEE JSSC*, April 2005, pp. 895-900. 24, 43

[24] H.Qin, Y. Cao, et.al, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," *Proceedings of IEEE*, 2004. 24

[25] S. Borkar, "Parameter Variations and Impact on Circuits and Microarchitecture," *DAC*,pp. 338-342, 2003. 26

[26] C.E. Blat et al., "Mechanism of Negative Bias Temperature Instability," *J. Applied Phys.*, pp. 1712-1720, 1991. 27

[27] E. Takeda, "Hot-Carrier Effects in Submicrometer MOS VLSIs," *IEEE Proc.*, pp. 153-162, 1984. 27

[28] D. Young and A. Christou, "Failure Mechanism models for Electromigration," *IEEE Trans. Reliability*, pp. 186-192, 1994.

[29] H. Su et al., “Full Chip Leakage Estimation Considering Power supply and Leakage Variations,” *ISLPED*, pp. 78-83, 2003. 27

[30] B.E. Stine et al., “Analysis and Decomposition of Spatial Variation in Integrated Circuit Processes and Devices,” *IEEE Trans. Semicond. Manuf.*, pp. 24-41, 1997. 27

[31] A. Srivastava, D. Sylvester et al.,*Statistical Analysis and Optimization for VLSI: Timing and Power*, Springer, 2005, chapter 1. 28

[32] S. Nassif, “Within-Chip Variability Analysis,” *IEDM Tech. Digest*, pp. 283-286, 1998. 28, 29, 31

[33] A. B. Kahng, Y.C. Pati, “Sub wavelength Lithography and its potential impact on Design and EDA,” *DAC*, pp. 799-804, 1999.

[34] Mukong Choi, Linda Milor, “Impact on Circuit Performance of Deterministic Within-Die Variation in Nanoscale Semiconductor Manufacturing,” *TCAD*, pp. 1350-1366, 2005. 29

[35] V. Mehrotra, S. Nasif et al., “Modeling the Effects of Manufacturing Variation on High-Speed Microprocessor Interconnect Performance,” *IEDM Tech. Digest*, pp. 767-770, 1998. 30

[36] K.Bernstein et al.,“High Performance CMOS variability in the 65nm regime and beyond,” *IBM Journal of Research and Development*, Sept. 2006, pp. 433-449. 31

[37] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge, U.K.: CUP, 1998.

[38] Asen Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub 50nm MOSFETs: A statistical 3D atomistic simulation study,” *Nanotechnology*, 1999, pp. 153-158. xi, 30, 31, 32, 34, 51

[39] R. W. Keyes,“The effect of randomness in the distribution of impurity atoms on FET threshold,” *Appl. Phys.*, vol.8, pp. 251-259, 1975. 32, 33

[40] X. Tang, V. De, J. Meindl,“Intrinsic MOSFET Parameter Fluctuations due to Random Dopant Placement,” *IEEE Trans. VLSI*, Dec. 1997, pp. 369-375. xi, 32, 33

[41] H.-S.P. Wong, Y. Taur et al.,“Discrete Random Dopant Distribution Effects in Nanometer-scale MOSFETs,” *Microelec. Reliab.*, 1998, pp. 1447-1456. 33

[42] D. J. Frank, Y. Taur et al.,“Monte-Carlo Modeling of Threshold Variation Due to Dopant Fluctuations,” *Symp. VLSI Tech.*, 1999, pp. 169-170.

[43] A. Asenov, A. R. Brown et al.,“Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer Scale MOSFETs,” *IEEE Trans. Electron Devices*, 2003, pp. 1837.

[44] G. Roy, F. Adamu-Lema et al.,“Intrinsic Parameter Fluctuations in conventional MOSFETs until the end of ITRS: A statistical simulation study,” *Journal Phys.*, 2006, pp. 188-191.

[45] J. Yin, X. Shi et al.,“A new method to simulate random dopant induced threshold voltage fluctuations in sub-50nm MOSFETs with non-uniform channel doping,” *Solid State Electronics*, 2006, pp. 1551-1556. 32

[46] A. Keshwarzi, G.Schrom,“Measurements and Modeling of Intrinsic Fluctuations in MOSFET Threshold Voltage,” *ISLPED*, 2005, pp. 26-29. 32, 33

[47] K. Agarwal, F.Liu et al.,“A Test Structure for Characterizing Local Device Mismatches,” *Symp. VLSI Circuits Dig. tech. Papers*, 2006, pp. 67-68. 33

[48] D. Burnett, K.Erington et al., “Implications of Fundamental Threshold Voltage Variations for High Density SRAM and Logic Circuits,” *VLSI Symp. Dig.*, 1994, pp. 15-16. 33

[49] A. Bhavnagarwala et al., “Fluctuation Limits and Scaling Opportunities for CMOS SRAM cells,” *IEEE Proc.*, 2005, pp. 15-16. 33

35

[50] B. Cheng, S. Roy, A. Asenov, “The Impact of Random Doping Effects on CMOS SRAM Cell,” *ESSCIRC*, 2004, pp. 219-222. xi, 36

[51] F.Tachibana, T. Hiramoto,“Re-examination of Impact of Intrinsic Dopant Fluctuations on SRAM Static Noise Margin,” *Japanese Journal of Applied Physics*, 2005, pp. 2147-2151. 36

[52] M. Yamaoka et al., “A Detailed Vth-variation Analysis for Sub-100nm Embedded SRAM Design”, *IEEE Proc.*, 2006, pp. 315-318. 58, 76

[53] C.K. Tsai et.al, “Analysis of Process Variation’s Effect on SRAM Read Stability,” *IDQED*, 2006. 73

[54] B. Cheng, S. Roy et al.,“Impact of Random Dopant Fluctuation on Bulk CMOS 6-T SRAM Scaling,” *IEEE Proc.*, 2006, pp. 258-261.

[55] B. Mohammad et al.,“Cache Design for Low Power and High Yield,” *IEEE Proc.*, 2008, pp. 103-107. 36, 101

[56] K. Zhang et al.,“A 3-GHz 70-Mb SRAM in 65nm CMOS Technology with Integrated Column-based Dynamic Power Supply,” *IEEE JSSC*, 2006, pp. 146. 41

[57] A. Agarwal, B.C. Paul, et al.,“A Process Tolerant Cache Architecture for Improved Yield in Nanoscale Technologies,” *IEEE Trans. VLSI*, 2005, pp. 27-38. 42

[58] S. Mukhopadhyay et al.,“Design of a Process Variant Tolerant Self-Repairing SRAM for Yield Enhancement in Nanoscaled CMOS ,” *IEEE JSSC*, 2007, pp. 1370. 42

[59] B. Cheng, S. Roy et al.,"CMOS 6-T SRAM cell design subject to atomistic fluctuations ," *Solid State Electronics*, 2007. 42

[60] H. Pilo ,J. Barwin, et al. "An SRAM Design in 65nm and 45nm Technology nodes featuring read and write assist circuits to expand operating voltage" *Symp. VLSI Circuits Digest Tech. papers*, 2006. 42

[61] R. Venkatraman and R. Castagnetti, "The Design, Analysis and Development of Highly Manufacturable 6-T SRAM Bitcells for SoC applications" *IEEE Trans. Electron Devices*, 2005, pp. 218-224. 42, 43

[62] M. Craig and J. Petersen, "Robust Methodology for State of the Art Embedded SRAM BitCell Design" *Proc. SPIE*, vol. 4692, 2002, pp. 380-389. 42, 46, 47

[63] S. Mukhopadhyay, H. Mahmoodi, Kaushik Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS," *TCAD*, 2005, pp. 1859-1879. 42, 47

[64] R.W. Mann et al.,"Ultralow-power SRAM technology", *IBM Journal of Research and Development*, Sept. 2003, pp. 553-563. 43, 44, 50, 51, 60

[65] F. Boeuf et al.,"0.248um2 and 0.334um2 Conventional Bulk 6T-SRAM bitcells for 45nm node Low Cost - general Purpose Applications", *Sym. On VLSI Tech. Digest of Tech. Papers*, 2005, pp. 130-131. 46

[66] M. Lavin et al.,"Backend CAD Flows for Restrictive Design Rules", *IEEE*, 2004, pp. 739-746. 46

[67] L. Liebmann, A. Barish et al.,"High-performance circuit design for the RET-enabled 65-nm technology node", *Proc. SPIE Vol. 5379*, May 2004, pp. 20. 46

[68] P. Gelsinger, Keynote address to 41st DAC, 2004. 46

[69] Online reference, *http://www.pimrc2006.org/Buss_sLocosto.pdf.* 46

[70] A. Srivastava, D. Sylvester,"Statistical Optimization of Leakage Power considering process variations using Dual-Vth and sizing", *DAC*, Jun. 2004, pp. 773-778. 46

[71] A. Papoulis,*Probability, Random Variables and Stochastic Process*, New York, McGraw Hill, Third Edition. 50

[72] S. Director, P. Feldmann,"Optimization of Parametric Yield: A tutorial," *IEEE Custom Integrated Circuits Conference*, 1992, pp. 3.1.1-3.1.8. 51, 60, 65

[73] J. Jaffari and M. Anis, "Variability Aware Device Optimization under Ion and leakage current constraints,"*ISLPED*, 2006, pp. 119-122. 62, 63, 66

[74] P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *Journal of Hydrology(46)*, 1980, pp. 79-88. 65

66

[75] B. H. Calhoun and A. Chandrakasan, "Analysing Static Noise Margin for Subthreshold SRAM in 65nm CMOS,"*ESSCIRC*, 2005. 74

[76] T. F. Coleman , Y. Zhang, Optimization Toolbox for use with Matlab, The Math works Inc. , 2005. 75

[77] HSPICE Manual. 75

[78] M.Dunga, et al, BSIM4.6 MOSFET model, University of California, Berkeley, *http://www-device.eecs.berkeley.edu/ bsim3/bsim4.html*. 75

[79] Predictive Technology Models for 45nm, *http://www.eas.asu.edu/ ptm*. 75

[80] A. Goel and B. Mazhari, "Gate Leakage and its Reduction in Deep-Submicron SRAM,"*Intl. Conf. VLSI Design, VLSID*, 2005, pp. 606-611. 76, 91

[81] Y. Hirano, M. Tsujiuchi et al, “A Robust SOI SRAM Architecture by using Advanced ABC technology for 32nm node and beyond LSTP devices,” *Symp. on VLSI Technology Digest of Technical Papers*, 2007, pp. 78-79. 91, 94

Appendix A

Publications from this work

This work resulted in the following publications/submissions under review.

1. V. Gupta and M. Anis,“Area/Yield Trade-offs in Scaled CMOS SRAM cells”, *European Solid State Circuits Conference, Edinburgh, U.K.*, 2008.