LOW-POWER AND HIGH-PERFORMANCE SRAM DESIGN IN HIGH VARIABILITY ADVANCED CMOS TECHNOLOGY

By

FATEMEH ATAEI

Bachelor of Science in Electrical Engineering

Shahid Beheshti University

Tehran, Iran

2007

Master of Science in Electrical Engineering

Amirkabir University of Technology

Tehran, Iran

2011

Submitted to the Faculty of the

Graduate Collage of the

Oklahoma State University

in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

May, 2017

LOW-POWER AND HIGH-PERFORMANCE SRAM DESIGN IN HIGH VARIABILITY ADVANCED CMOS TECHNOLOGY

Dissertation Approved by:

Professor James E. Stine

Dissertation Advisor

Professor Matthew R. Guthaus

Commitee Member

Professor Chriswell Hutchens

Commitee Member

Professor Jingtong Hu

Commitee Member

Professor John W. Mintmire

Outside Commitee Member

Acknowledgments [1]

First and foremost, I would like to thank my outstanding advisor Professor James E. Stine for his invaluable instructions throughout the course of my PhD. His insights, encouragement, support, and friendship have been a great source of inspiration for this work. His immense knowledge on VLSI had helped me a lot in generating new ideas and implementing them. By challenging me and expecting more from me, James shaped my professional identity and built confidence in me. He has supported me through every research endeavor, and he has taught me to be a critical, sincere, cooperative, and respectful researcher. He has given me something to strive for technically and personally. Thank you James for your generous support and guidance.

I am also grateful to Professor Matthew R. Guthaus, Professor Chriswell Hutchens, Professor Jing-tong Hu and Professor John W. Mintmire for agreeing to be a part of reading committee. To be able to contribute alongside an outstanding researcher as Professor Guthaus is a great honor of my career. Thank you Matt for your feedbacks and supports. Because of your inputs, I am better at programming, which means a lot!

I am proud to have been a part of VLSIARCH group at OSU for its great research environment and wonderful colleagues who I could interact with. The many hours I have spent in this group have been very stimulating and enriching. I must thank past and present members of VLSIARCH not just for their technical feedback but also for their support and assistance. Dr. Mehedi Sarwar, Dr. Masoud Sadeghian, Dr. Son Bui, Matthew Gaalswyk, Robert Elliott, Manju Subbarayappa, Rabin Thapa, Mohammad Al Mestiraihi and Tuan Nguyen, good luck to each of you in your future endeavors.

Last but not the least, thanks to my family and my amazing parents for their love, encouragement, support, blessing and prayers to lift me over obstacles. You are far away, but I have always felt you here with me. I thank my amazing husband Dr. Siamak Dadras for his patience during the course of my PhD. Siamak, you are the reason behind this accomplishment, and your smiles are the only rewards I hope for every day. Thank you for your love and support. I love you with all my heart!

[1] Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: FATEMEH ATAEI

Date of Degree: MAY 2017

Title of Study: LOW-POWER AND HIGH-PERFORMANCE SRAM DESIGN IN HIGH VARIABILITY ADVANCED CMOS TECHNOLOGY

Major Field: ELECTRICAL ENGINEERING

Abstract: As process technologies shrink, the size and number of memories on a chip are exponentially in- creasing. Embedded SRAMs are a critical component in modern digital systems, and they strongly impact the overall power, performance, and area. To promote memory-related research in academia, this dissertation introduces OpenRAM, a flexible, portable and open-source memory compiler and characterization methodology for generating and verifying memory designs across different technologies.

In addition, SRAM designs, focusing on improving power consumption, access time and bitcell stability are explored in high variability advanced CMOS technologies. To have a stable read/write operation for SRAM in high variability process nodes, a differential-ended single-port 8T bitcell is proposed that improves the read noise margin, write noise margin and readout bitcell current by 45%, 48% and 21%, respectively, compared to a conventional 6T bitcell. Also, a differential-ended single-port 12T bitcell for subthreshold operation is proposed that solves the half-select disturbance and allows efficient bit-interleaving. 12T bitcell has a leakage control mechanism which helps to reduce the power consumption and provides operation down to 0.3 V. Both 8T and 12T bitcells are analyzed in a 64 kb SRAM array using 32 nm technology. Besides, to further improve the access time and power consumption, two tracking circuits (multi replica bitline delay and reconfigurable replica bitline delay techniques) are proposed to aid the generation of accurate and optimum sense amplifier set time.

An error tolerant SRAM architecture suitable for low voltage video application with dynamic power-quality management is also proposed in this dissertation. This memory uses three power supplies to improve the SRAM stability in low voltages. The proposed triple-supply approach achieves 63% improvement in image quality and 69% reduction in power consumption compared to a single-supply 64 kb SRAM array at 0.70 V.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Moore's law of scaling [6] directly or indirectly has been the most important driving force behind the semiconductor industry and the main cause of the tremendous capabilities of today's ICs. Scaling of CMOS transistors and push for better performance enabled embedding of millions of Static Random Access Memory (SRAM) cells into contemporary ICs. Although many aspects of CMOS scaling begin to saturate, density scaling remains a key objective of the semiconductor industry [7]. Density scaling enables circuit and architecture level parallelism and leads to energy-efficiency and performance improvements. Embedded SRAMs can occupy the majority of the chip area in some applications and due to their regular structures and broad applications, SRAMs are carefully designed with aggressive layout rules [8]. Therefore, SRAM occupy a dominating portion of the total die area and the total power consumption [9, 10].

There are several obstacles on the way of continuous scaling of SRAM. Power-delay-area product of SRAM is not scaling as efficiently as that of logic circuits. This phenomenon is known as the non-scaling problem of SRAM which presents one of the most difficult tasks in designing of nano-scaled SRAMs. The possible solutions helping to mitigate the SRAM non-scaling problem are driven by the target application of SRAM arrays with the high-performance applications on one end of the spectrum and the low- power applications on the other. SRAMs are strongly subjected to the power, performance, and density trade-offs; improvement in one of the dimensions strongly stresses the others and all three dimensions are important to some degree in all applications. SRAM design involves making wise compromises to support the most important requirements based on the application.

As the process technology continues to scale deeper into the nanometer region, the stability of

embedded SRAM cells is a growing concern as well. As a consequence, large SRAM arrays impact all aspects of chip design and manufacturing because they became the yield-limiters in modern high-performance ICs. Large arrays of fast SRAM help to boost the system performance. However, the area impact of incorporating large SRAM arrays into a chip directly translates into a higher chip cost. Balancing these requirements is driving the effort to minimize the footprint of SRAM cells. As a result, millions of minimum-size SRAM cells are tightly packed making SRAM arrays the densest circuitry on a chip. Such areas on the chip can be especially susceptible and sensitive to manufacturing defects and process variations. SRAM design is highly constrained, especially in the face of limitations from device-level process variability to system-level power consumption.

SRAM is a major yield limiter due to large number of transistor count, use of the smallest transistors in the bitcells and increasing leakage-power consumption in sub-100 nm technologies. SRAM is more vulnerable to process, voltage and temperature variations than other logic circuits since minimum size devices are used in its design. As the SRAM bitcell size is scaled down in advanced technologies, the device threshold voltage mismatch due to random dopant effect further irritates the problem and makes a fundamental limit on the SRAM bitcell size. Therefore, SRAMs must be designed based on application to support its important requirements.

The main motivation behind this research is the need to develop techniques for highly stable, power efficient and high performance SRAM which has a heightening importance in digital systems. SRAM design requires coordination with technology development due to its increasing sensitivity to processing and manufacturing factors. As a result, this dissertation focuses on low-leakage and high performance SRAM circuit techniques which are compatible with industry methodologies. It is the hope that this dissertation contributes to solve some of the most critical issues facing SRAMs and every effort is made to identify those as well. This dissertation contributes in the following areas:

1. Promoting memory-related research in academia by introducing OpenRAM memory Compiler.
2. SRAM stability and writability improvement by proposing a novel 8T bitcell.
3. SRAM performance improvement under process, temperature and voltage variations in deep nanometer technologies, by proposing multi replica bitline delay and reconfigurable replica bitline delay techniques to control the timing of SRAM sense amplifier in read operation.

4. SRAM leakage and dynamic power reduction by proposing a novel, subthreshold and half-select disturbance free 12T bitcell.

5. New approximate SRAM architecture for low-voltage video applications.

Chapter 2 describes the overall SRAM architecture, its operation principles and metrics needed to evaluate the SRAM quality. Process variation effects on SRAM stability and operation and also different types of process variations are described in this chapter. It is shown how process variation impacts the critical parameters of SRAM.

Chapter 3 introduces OpenRAM, an open-source memory compiler and characterization methodology. OpenRAM compiler is a flexible and portable platform for generating and verifying memory designs across different technologies. OpenRAM in written in Python and generates Spice netlist, GDSII layout and timing/power information for SRAMs in Freepdk45 and SCMOS technologies.

Chapter 4, 5, 6 and 7 explore circuit designs in order to improve the stability, access time and power consumption in SRAMs. Chapter 4, describes a differential-ended single-port 8T SRAM bitcell which is designed to improve the read stability and writability of SRAM in sub-100 nm technologies under process variations. Dual-threshold configurations of proposed 8T bitcell are exhaustively examined for read/write stability, leakage power, access time and other important metrics to find the best combination of low and high threshold voltage devices in 8T bitcell which leads to a low power and high performance SRAM.

In Chapter 5, Multi Replica Bitline Delay (MRBD) and Reconfigurable Replica Bitline Delay (RRBD) techniques for sense amplifier enable signal generation are introduced. MRBD helps to reduce the access time and power consumption of SRAM in scaled technology nodes. MRBD uses multiple replica bitlines and replica cells to generate an accurate timing signal in high variability technologies. In RRBD which has the same architecture as MRBD, number of replica columns and replica cells are controlled with a digital control code. MRBD decreases the deviation and RRBD adjust the timing of sense amplifier control signal. Performance of a 64 kb SRAM array with proposed 8T bitcell and MRBD technique is compared to 6T SRAM array with traditional replica bitline, demonstrating improvements in the read/write noise margin and the standard deviation reduction of the access time. Also, simulation results of a 64 kb SRAM array with proposed RRBD technique are discussed in this chapter.

Chapter 6 introduces a novel subthreshold 12T SRAM bitcell suitable for low voltage applications.

This 12T bitcell is half-select and read-disturb free, therefor provides efficient bit-interleaving structure to solve multi-bit soft errors. Besides, This bitcell has a leakage control mechanism leads to a low-voltage with less access time design for SRAM. Proposed 12T bitcell is demonstrated in a 64 kb (256×256 bit) SRAM array in 32 nm CMOS SOI technology which operates up to 2 GHz at 0.9 V while it shows a robust operation at 0.3 V and 50 MHz as well.

Chapter 7 introduces an approximate SRAM design for video applications. Energy can be traded with output signal quality in this SRAM. The proposed approximate 6T SRAM architectures uses three supply voltages to improve the static noise margin during read and write modes and also reduces leakage current in retention mode, hence, it allows aggressive supply voltage scaling for low power multimedia applications. Simulation results in IBM/Global Foundries cmos32soi 32 nm technology show a 69% power saving and a 63% improvement in image quality for the proposed array compared to a conventional single-supply 64 kb 6T SRAM at 0.70 V and 20 MHz. The proposed SRAM also allows a dynamic power-quality trade-off at run time and makes the 6T SRAM array a suitable power efficient memory for different video/image applications.

Chapter 2

SRAM Overview

2.1 Introduction

Static Random Access Memory (SRAM) is a volatile memory as the information or the instructions stored in the memory will be lost if the power is switched off. The word static means that the memory retains its contents as long as the power is turned on and no refresh is needed like dynamic RAMs. The word random refers to the fact that any piece of data can be returned at a constant time regardless of its physical location and whether or not it is related to the previous piece of data. Random access means that locations in the memory can be written to or read from in any order, regardless of the memory location that was last accessed [11]. SRAM is used as one of the primary storage because of its speed and consistency. SRAM is the main working area used for loading, displaying and manipulating applications and data. In this Chapter SRAM architecture and its modules are described. Effect of process variations on SRAM operation and common metrics to evaluate the stability of SRAM bitcell are also explained.

2.2 Overall SRAM Architecture

SRAM's typically consist of an array of memory bitcells with peripheral circuits and control logic. Figure 2.1 clearly depicts the memory array as well as the other main blocks. The SRAM bitcell array, address decoder, wordline driver, column multiplexer, precharge circuitry, write driver, sense amplifier, and control logic are the SRAM main modules. Although there are different configurations for each

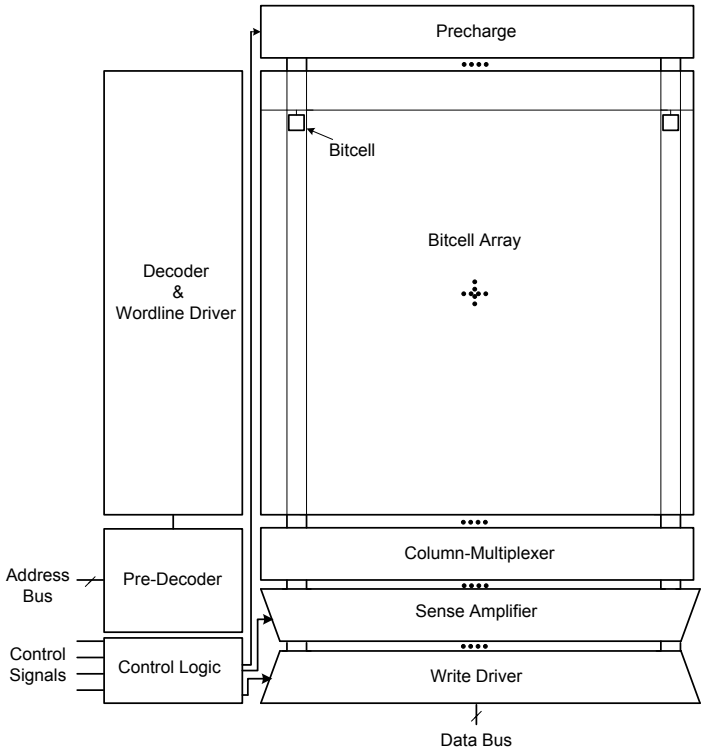


Figure 2.1: Overall SRAM architecture.

of these modules, in this dissertation only the schemes which are used in proposed SRAM designs are explained.

SRAM Bitcell: The 6T bitcell is the most commonly used memory bitcell in SRAM devices. It is named 6T bitcell because it consists of six transistors: two cross-coupled inverters and two access transistors (M5 and M6) as shown in Figure 2.2. The cross-coupled inverters hold a data bit, Q, and its inverted value, /Q. This bit can either be written into or read from by the bitlines (BL, /BL). The access transistors are used to isolate the bitcell from the bitlines so that data is not corrupted while a bitcell is idle. The transistors in the bitcell must be carefully sized to ensure proper operation. For a read operation, the M1 and M2 must be sized larger than access transistors M5 and M6. This is necessary because when the wordline is asserted, both Q and /Q are initially pulled up to the precharge value. Assuming that a 1 is stored at Q, /Q must remain 0 regardless of the voltage rise experienced when the wordline is asserted. During a write operation, the value stored in the bitcell is being overwritten. This means that M5 and M6 must be strong enough to overpower the feedback inverter and must be sized larger than M3 and M4 [11].

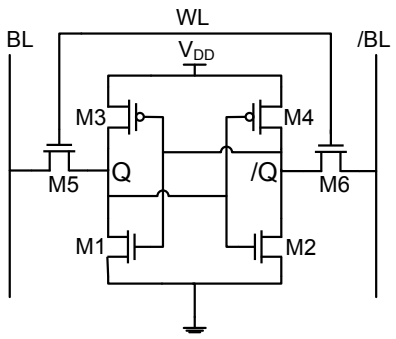


Figure 2.2: 6T SRAM bitcell (M1/M2 pull-down, M3/M4 pull-up and M5/M6 access transistors).

6T bitcells are tiled together in both the horizontal and vertical directions to make up the memory array. This means that the memory bitcell should be made as small as possible so that the array can be as dense as possible. The size of the memory array is directly related to the number of words and the size of the words that will need to be stored in the SRAM. For example, an 8 kB memory with a word size of 8 bits, will consist of 8 columns and 1024 rows. It is common practice to keep the aspect ratio of memory arrays as square as possible. This helps ensure that the bitlines do not become too long, which can increase the bitline capacitance, slow down the operation, and lead to increased leakage. To make the design more square, multiple words can share rows by interleaving the bits of each word. If the 8 kb memory was rearranged to allow two words per row, then the array would have 16 columns and 512 rows. Geometrically distributing the bits of each word (by allowing multiple words per row with interleaved bits) improves yield and soft-error robustness [11].

Other types of memory bitcells, such as 7T, 8T, 10T and even 12T bitcells, can be used as alternatives to the 6T bitcell. Each of these bitcells offer certain advantages. For example, the 8T bitcell provides higher read and write noise margins in comparison to the 6T bitcell [1, 2, 12]. The 10T [13, 14] and 12T [4, 5, 15] bitcells provide improved soft error tolerance and operation at lower supply voltages. More details on 8T, 10T and 12T bitcells are provided in Chapter 4 and 6.

Precharge Circuitry: The precharge circuit is used to precharge both bitlines during the first phase of the clock in read and write operations and is depicted in Figure 2.3(a). It is a fairly simple circuit that consists of three PMOS transistors. The input signal to the cell, $PCLK$, enables all three transistors. M1 and M2 charge BL and /BL to V_{DD} and M3 helps to equalize the voltages on the bitlines [11]. Equalizing the bitlines helps to avoid any errors that may occur while the sense amplifier is sensing the

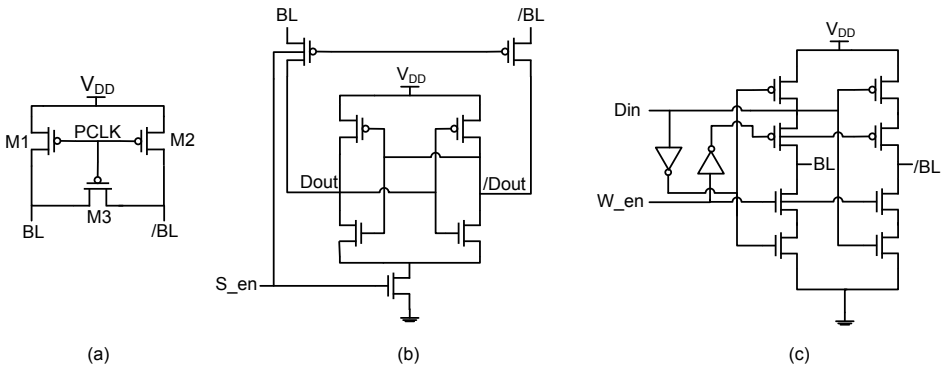


Figure 2.3: (a) Precharger, (b) sense amplifier with isolation transistors and (C) write driver.

voltage difference between the bitlines during a read operation.

Sense Amplifier: The sense amplifier is used to sense the difference between the bitlines while a read operation is performed. A sense amplifier is necessary to recover the signals from the bitlines because they do not experience full voltage swing. As the size of the memory array grows, the capacitive load of the bitlines increase and the voltage swing is limited by the small memory bitcells driving this large load. A differential sense amplifier is used to sense the small voltage difference between the bitlines and accelerates the read operation. The schematic for the sense amp is shown in Figure 2.3(b). The sense amplifier is enabled by the S_en signal, which initiates the read operation. Before the sense amplifier is enabled, the bitlines are precharged to V_{DD} by the precharge unit. When the sense amp is enabled, one of the bitlines experiences a voltage drop based on the value stored in the memory bitcell. If a zero is stored, the BL voltage drops and if a one is stored, the /BL voltage drops. The voltage difference between BL and /BL is sensed and the output signal is then taken to a true logic level and latched to the data bus [11].

Write Driver: The write driver is used to drive the input signal into the memory bitcell during a write operation. It can be seen in Figure 2.3(c) that the write driver consists of two tristate buffers. It takes in a data bit, from the data bus, and outputs that value on the BL, and its complement on /BL. Both tristates are enabled by the W_en signal. The bitlines always need to be complements to ensure that correct data is stored in the 6T bitcell. Also, the drivers need to be appropriately sized as the memory array grows and the bitline capacitance increases [11].

Address Decoder and Wordline Driver: The address decoder takes the row address bits from the

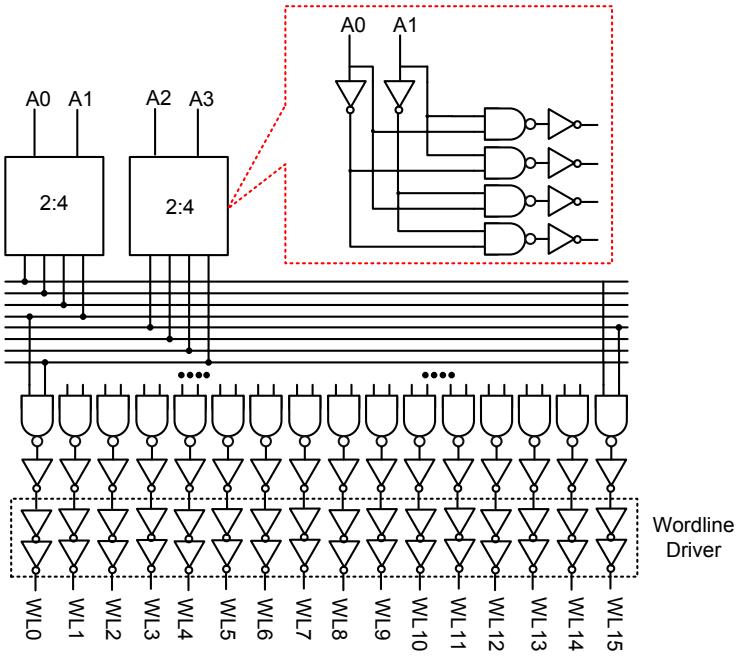


Figure 2.4: A 4:16 hierarchical decoder with wordline driver.

address bus as inputs, and asserts the appropriate wordline in the row that data is to be read from or written to. An n-bit input can control 2^n wordlines. Figure 2.4 illustrates a 4-to-16 hierarchical decoder made of two 2-to-4 decoders. Based on the input address, a specific wordlines is asserted [11]. Wordline drivers are inserted, as buffers, in-between the wordline output of the address decoder and the input of the 6T bitcell. The wordline drivers ensure that as the size of the memory array increases, and the wordline capacitance increases, the signal is still able to turn on the access transistors in all bitcells.

Column Multiplexer: The schematic for a 2-to-1 multiplexer is shown in Figure 2.5. This type of multiplexer is bi-directional and is used for both the read and write operations; it connects the bitlines of the memory array to both the sense amplifier and the write driver. If there is only one word per row in the array, then no column mux is needed. Relative to other column mux designs, such as tree mux, single pass-transistor mux uses significantly less devices and better performance [11]. tree mux design can provide poor performance if a large decoder with many levels is needed. The delay of a tree mux quadratically increases with each level [11]. Due to this fact, single pass-transistor mux should be considered for larger memory arrays.

Control Logic: The control circuitry ensures that the SRAM operates as intended during a read or write cycle by enabling the necessary structures in the SRAM. As shown in Figure 2.6, the control

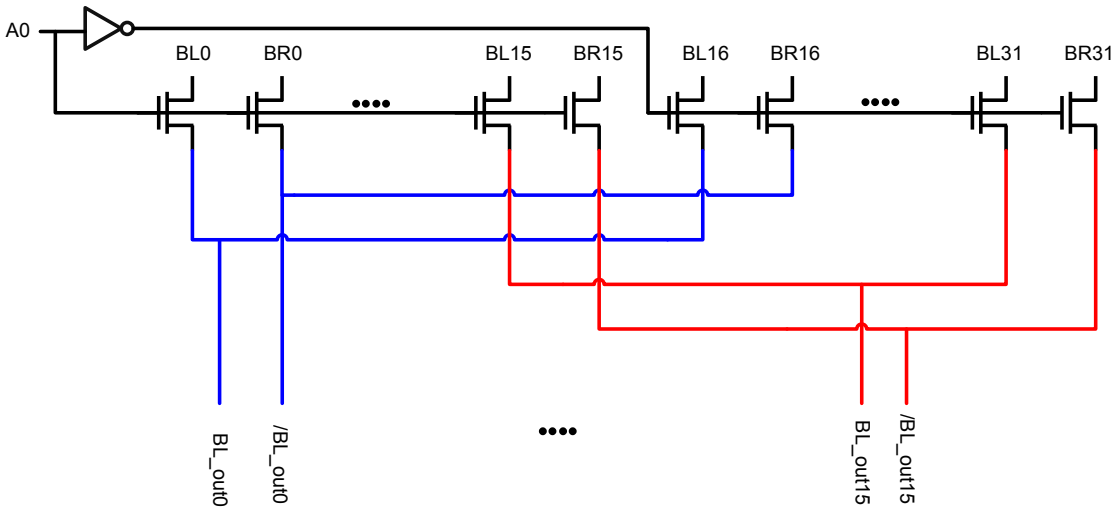


Figure 2.5: A 2:1 single-pass-transtistor column multiplexer.

Table 2.1: Generation of control signals in a synchronous SRAM design.

Operation	**Inputs**			**Outputs**		
	CSb	OEb	WEb	s_en	w_en	tri_en
READ	0	0	1	1	0	1
WRITE	0	1	0	0	1	0

logic takes three active low signals as inputs: chip select bar (CSb), output enable bar (OEb), and write enable bar (WEb). CSb enables the entire SRAM chip. When CSb is low, the appropriate control signals are generated and sent to the architecture blocks. Conversely, if CSb is high then no control signals are generated and SRAM is turned off or disabled. The OEb signal signifies a read operation; while it is low the value seen on the data bus will be an output from the memory. Similarly, the WEb signal signifies a write operation. All of the input control signals are latched with master-slave flip-flops, ensuring that the control signal stays valid for the entire operation cycle. The control signal flip-flops use the normal clock to generate local signals used to enable or disable structures based on the operation [16, 17, 18]. Address and input-data flip-flops are combined with global clock as well.

After all control signals are latched, they are ANDED with the clk_bar because the read/write circuitries should only be enabled after the precharging of the bitlines had ended on the negative edge of the clock. The w_en signal enables the write driver during a write to the memory.The s_en signal enables the sense amplifier during a read operation. Details on MRBD architecture are outlined in Chapter 5. tri_en and tri_en_bar enable the tristates during read in order to drive the outputs onto the

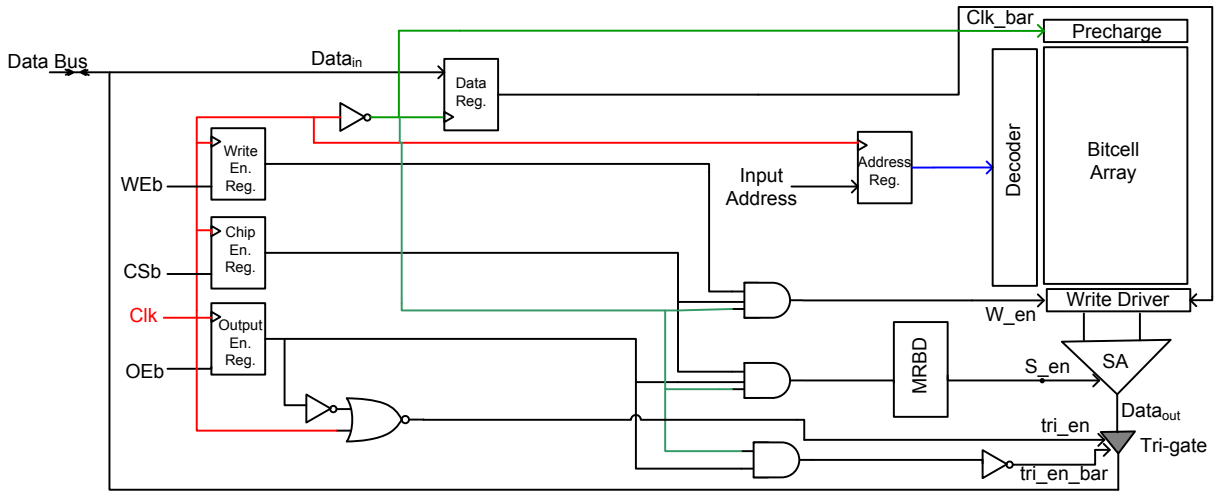


Figure 2.6: Control logic circuitry with SRAM array.

data bus. Table 2.1 shows the truth table for the control logic. The s_en signal to enable the sense amplifier is true when $(CSb\ .\ OEb\ .\ Clk\_bar)$ is true. Similarly, write driver enable signal, w_en, is true when $(CSb\ .\ WEb\ .\ Clk\_bar)$ is true. tri_en and tri_en_bar are true when $\sim (OEb\_bar|Clk)$ and $\sim (OEb\ .\ Clk\_bar)$ are true, respectively.

2.3 Process Variation Effects on SRAM Stability

In classic Dennard scaling [19], oxide thickness, transistor length and width were scaled by a constant factor $(1/k)$ in order to provide a delay improvement of $1/k$ at constant power density [19]. As a consequence of continued density scaling, features are moving ever closer to atomic dimensions and light wavelengths which means management of variation will play a major role in future technology scaling.

Aggressive scaling of CMOS transistors in sub-100 nm nodes has led to chips with billions of transistors in modern ICs that created huge design challenges. Manufacturing tolerances in process technology are not scaling at the same pace as transistor channel length due to process control limitations and variations due to fundamental physical limits are increasing significantly with technology scaling. Currently variability is one of the biggest challenges facing the semiconductor industry which dramatically impacting SRAM design at nanometer technology nodes. Process variations have a strong impact on SRAM because they increase bitcell failure probability especially at low voltage. In addition,

SRAM uses the minimum sized transistors to achieve the highest possible integration density, which makes SRAM the most sensitive circuit to process variations. SRAM yield has a strong impact on the overall product yield and SRAM yield loss due to variability is likely the dominant cause of yield loss in modern ICs. Therefore, development of variation-tolerant techniques to reduce SRAM sensitivity to variations is the main focus of SRAM design in sub-100 nm technologies.

Variation is the deviation from intended values for structure or a parameter of concern. The electrical performance of modern ICs are subject to different sources of variations that affect both the device and interconnects. The sources of variation can be categorized into two groups [20]:

1. **Die-to-Die Variations (global variations or inter-die variations):** affect all devices on the same chip in the same way. For example, Die-to-Die variations may cause all the transistors' gate lengths to be larger than a nominal value. global variations have been a longstanding design issue, and are typically accounted for during circuit design with using corner models. These corners are chosen to account for the circuit behavior under the worst possible variation, and were considered efficient in older technologies where the major sources of variation were global variations [20]. Global variation is the gradient variation across the wafer caused due to physical errors during manufacturing a device. It is caused due to misalignment in the lenses and change in properties of elements used in the lithographic process. Global variation is the difference between average parameter values of the die and can include the average NMOS/PMOS threshold voltage, dielectric thickness or poly width.

2. **Within-Die Variations (local variations or intra-die variations):** correspond to variability within a single chip, and may affect different devices differently on the same chip. For example, devices in close proximity may have different V_{TH} than the rest of the devices. Local variation is between the devices placed in close vicinity of each other and can include the number of NMOS/PMOS channel doping ions, poly line edge roughness and local layout dependent lithography effects. local variations can be subdivided into two classes [20]:

   Random variations: are sources that show random behavior, and can be characterized using their statistical distribution.

   Systematic variations: show variational trends across a chip which are caused by physical phenomena during manufacturing such as distortions in lenses and other elements of lithographic

systems.

The terms **random variation** and **systematic variation** do not have a unified definition in the semiconductor community. random variation can be defined as the variation measured between a pair of two closely spaced objects. Systematic variation can be defined as the variation measured from a number of widely separated objects, after the random variation has been removed [21]. If the SRAM is within the process defined area then it is dominated by random mismatch. If the SRAM is large enough then systematic variation will dominate over random variation. Devices fabricated at the center of the wafer will have different properties when compared to the devices fabricated at the edge of the wafer [22, 23]. Process variations impact device structure and change the electrical properties of the circuit. However, process variation is not a barrier to Moor's law of scaling [6]; it is a challenge to be overcome and design can play an important role in variability reduction.

2.4 SRAM Operation

2.4.1 Read and Write Operation

The 6T bitcell can be accessed to perform the two main operations associated with memory: reading and writing. When a read is to be performed, both bitlines are precharged to V_{DD}. This precharging is done during the first half of the read cycle and is handled by the precharge circuitry. In the second half of the read cycle the wordline is asserted, which enables the access transistors. If a 1 is stored in the bitcell then /BL is discharged to GND, and BL is pulled up to V_{DD}. Conversely, if a 0 is stored, then BL is discharged to GND and /BL is pulled up to V_{DD} [11]. While performing a write operation, both bitlines are also precharged to V_{DD} during the first half of the write cycle. Again, the world line is asserted, and the access transistors are enabled. The value that is to be written into the bitcell is applied to BL, and its complement is applied to /BL. The drivers that are applying the signals to the bitlines must be appropriately sized so that the previous value in the bitcell can be overwritten [11].

2.4.2 Read and Write Stability

Voltage Transfer Characteristic Curves: The read Static Noise Margin (SNM) quantifies the extent to which a 6T bitcell can reliably hold each of the two data states required while being subjected to a static

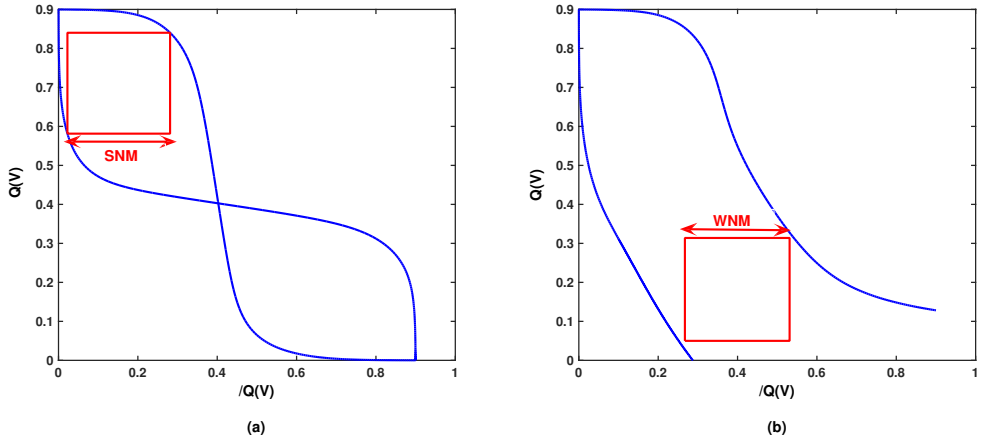


Figure 2.7: Voltage transfer characteristic curves for (a) SNM and (b) WNM calculation.

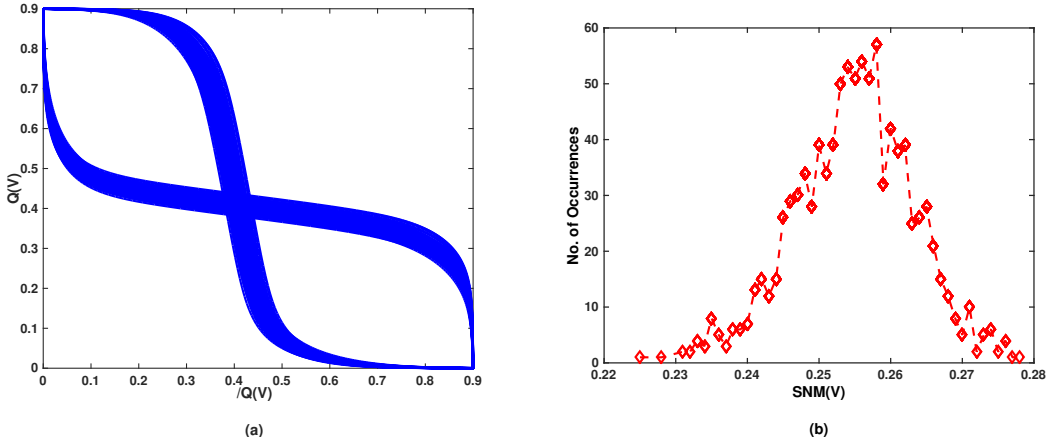


Figure 2.8: (a) Shifted butterfly curves under process variation and (b) Gaussian distribution for SNM.

read condition. The SNM is defined as the maximum possible noise available at the gates of the cross-coupled inverters or storage element that does not flip the bitcell value [24]. The read Voltage Transfer Characteristic (VTC) can be measured by sweeping the voltage at storage node Q with both BL and $/BL$ and WL biased at V_{DD} while monitoring the node voltage at $/Q$ [11]. The SNM can be quantified by the side of the largest square embedded between the read VTC curves. Figure 2.7(a) shows the read VTC curves for the 6T bitcell during read operation. Figure 2.8(a) shows how variation can shift the transfer-functions, easily leading to the loss of the read SNM. Figure 2.8(b) shows the distribution of the SNMs in presence of local and global variations. As shown, variation strongly limits the region where read SNM is preserved, specifically restricting operation at low V_{DD} and high V_{TH}, where sub-array energy tends to be optimized.

The Write Noise Margin (WNM) measures how easy or difficult it is to write into the bitcell; it is the highest BL potential that can flip the bitcell data [24]. The write VTC is measured by sweeping the voltage at the storage node Q with BL and WL biased at V_{DD} and /BL biased at GND while monitoring the node voltage at /Q. This VTC should be used in combination with the VTC measured by sweeping the voltage at the storage node /Q while monitoring the node voltage at Q [11]. WNM can be quantified by the side of the smallest square embedded between the lower-right half of the VTC curves. Figure 2.7(b) shows the WNM of 6T bitcell. During read or hold, three roots of intersection are desired, indicating bistability as shown in Figure 2.7(a). During write, only one root of intersection is desired, so that the cell will deterministically flip to one of the two data states, as set by the BL polarity as shown in Figure 2.7(b).

Variation strongly limits the functional region and decreases the SNM and WNM value at low supply voltages. Some read VTC curves will fail to preserve a 0 when device M5 becomes too strong, relative to M1 (see Figure 2.2), and the trip point of the opposite inverter shifts toward 0 V from a weakened M4 and a strengthened M2. Also, some write VTC curves will fail to successfully write from 1 to 0 under complementary conditions of a strong M3, weak M5, and strong M2. The main reason is related to the fact that the current through a transistor is proportional to (V_{GS} - V_{TH}) when $V_{GS} > V_{TH}$, and is approximately proportional to $10^{(V_{GS}-V_{TH})/100\ mv}$ when $V_{GS} < V_{TH}$. Thus, the impact of V_{TH} fluctuation reduces with increasing power supply but becomes intolerable at low supply voltages.

N-Curves: Figure 2.9(a) shows the N-curves of a 6T SRAM bitcell under V_{TH} variation. N-curves contains information on both read stability and writability of a SRAM bitcell, therefore allows a complete stability analysis with only one curve [25, 26]. For extracting the N-curve as shown in Figure 2.9(b), bitlines and wordline are at V_{DD}. A voltage (V_{in}) sweep node Q from 0 to V_{DD} and the corresponding current I_{in} is measured [25]. Following parameters define how stable is bitcell in read and write mode:

Static Voltage Noise Margin (SVNM): is the voltage difference between first two zero crossing points in Figure 2.9(a) and indicates the maximum tolerable DC noise voltage at the input of the inverter in bitcell before its content changes [25].

Static Current Noise Margin (SINM): is defined as the maximum value of the DC current that can be injected into the SRAM bitcell before its content changes [25] and is given by the peak value of the I_{in} that is between the first two zero crossing points in Figure 2.9(a).

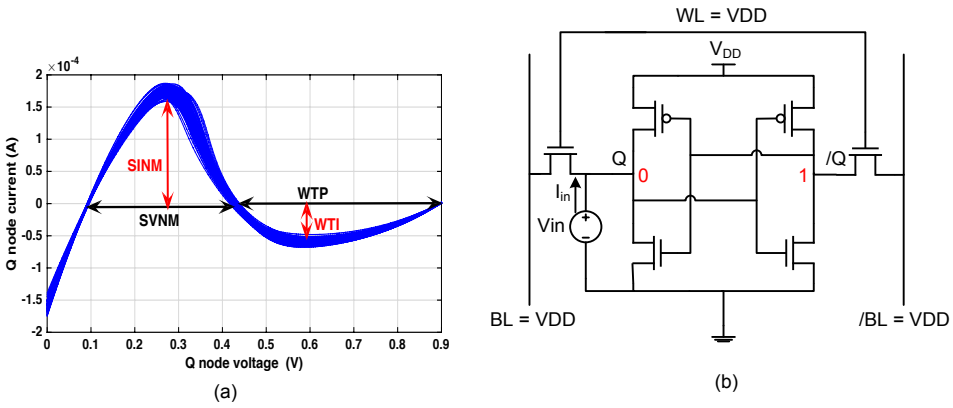


Figure 2.9: (a) Shifted N-Curves under process variations (b) setup to calculate the N-curve values.

Write-Trip Voltage (WTV): defines the maximum voltage on the bitline to flip the bitcell state [25]. In N-curve, the voltage difference between the second and last zero crossing points is the WTV.

Write-Trip Current (WTI): is the amount of current needed to write over the content of bitcell. The negative peak value of I_{in} after the second zero crossing of N-curves in Figure 2.9(a) gives the WTI [25].

In order to measure the read stability, the static power noise margin (SPNM), which is the area below the N-curves between the first two crossing points and has a unit of power, must be considered. The SPNM can be calculated by multiplying the SVNM and SINM values. Bitcell with larger SPNM have better read data stability. Similarly, for the write-ability, write-trip power (WTP) which is the area above N-curve between second and last crossing point must be considered. The WTP can be calculated by multiplying WTV and WTI and for a faster writability the value of WTP must be smaller.

According to the measurement and analysis in [27] values from N-curve are poorly correlated with write ability and N-curve is not a suitable method to calculate the write noise margin. Here in this dissertation, only VTC curves are used to calculate the stability in both read and write operation.

2.5 Summary and Conclusions

This Chapter Explained the SRAM modules and SRAM operation. Effect of process variation on SRAM operation and stability are discussed and it is shown that how increased process variation in sub-100 nm technologies leads to less stability in SRAM operation. Several metrics which are going to be use in the rest of this dissertation are introduced and explained how stability of SRAM bitcell can be evaluated using these metrics.

Chapter 3

OpenRAM: A Portable Open-Source Memory Compiler and Characterizer

3.1 Introduction

SRAMs have become a standard component embedded in all System-on-Chip (SoC), Application-Specific Integrated Circuit (ASIC), and micro-processor designs. Their wide application leads to a variety of requirements in circuit design and memory configuration. However, manual design is too time consuming. The regular structure of memories leads well to automation that produces size and configuration variations quickly, but developing this with multiple technologies and tool methodologies is challenging. Thus, a memory compiler is a critical tool. Most academic ICs design methodologies are limited by the availability of memories. Many standard-cell Process Design Kits (PDKs) are available from foundries and vendors, but these PDKs frequently do not come with memory arrays or memory compilers. If a memory compiler is freely available, it often only supports a generic process technology that is not fabricable. Due to academic funding restrictions, commercial industry solutions are often not feasible for researchers. In addition, these commercial solutions are limited in customization of the memory sizes and specific components of the memory. PDKs may have the options to request black box memory models, but these are also not modifiable and have limited available configurations. These restrictions and licensing issues make comparison and experimentation with real world memories impossible.

Academic researchers are able to design their own custom memories, but this can be a tedious and time-consuming task and may not be the intended purpose of the research. Frequently, the memory design is the bare minimum that the research project requires, and, because of this, the memory designs are often inferior and are not optimized. In memory research, peripheral circuits are often not considered when comparing memory performance and density. The lack of a customizable compiler makes it difficult for researchers to prototype and verify circuits and methodologies beyond a single row or column of memory bitcells.

The OpenRAM compiler aims to provide an open-source memory compiler development framework for memories. It provides reference circuit and physical implementations in a generic 45 nm technology and fabricable Scalable CMOS (SCMOS), but it has also been ported to several commercial technology nodes using a simple technology file. OpenRAM also includes a characterization methodology so that it can generate the timing/power characterization results in addition to circuits and layout while remaining independent of specific commercial tools. Most importantly, OpenRAM is completely user-modifiable since all source code is open source at:

`https://openram.soe.ucsc.edu`

This Chapter provides a background on previous memory compilers and presents the reference memory architecture in OpenRAM. Implementation and main features of the OpenRAM memory compiler are introduced and an analysis of the area, timing and power is shown for different sizes and technologies of memory.

3.2 Background

Memory compilers have been used in design flows to reduce the design time long before contemporary compilers [28, 29]. However, these compilers were generally not portable as they were nothing more than quick scripts to aid designers. Porting to a new technology essentially required rewriting the scripts. However, the increase in design productivity when porting designs between technologies has led to more research on memory array compilers [30, 31, 32, 33].

As technology entered the Deep Sub-Micron (DSM) era, memory designs started becoming one of the most challenging parts of circuit design due to decreasing SNM, increasing fabrication variability,

and increasing leakage power consumption. This increased the complexity of memory compilers dramatically as they had to adapt to the ever-changing technologies. Simultaneously, design methodologies shifted from silicon compilers to standard cell place and route methods which required large optimized libraries. During this time, industry began using third-party suppliers of standard cell libraries and memory compilers that allowed their reuse to amortize development costs. These next-generation memory compilers provided silicon-verification that allowed designers to focus on their new design contribution rather than time-consuming tasks such as memory generation.

Contemporary memory compilers have been widely used by industry, but the internal operation is typically hidden. Several prominent companies and foundries have provided memory compilers to their customers. These memory compilers usually allow customers to view front-end simulation, timing/power values, and pin locations after a license agreement is signed. Back-end features such as layout are normally supplied directly to the fab and are only given to the user for a licensing fee. Specifically, Global Foundries [34] offers front-end PDKs for free, but not back-end detailed views. Virage Logic [35] provides a dashboard control compiler that selects from a pre-designed set of memory configurations. Faraday Technologies [36] provides a black box design kit for UMC technologies, where users do not know the details of the internal memory design. Dolphin Technology [37] offers closed-source compilers which can create RAMs, ROMs, and CAMs for TSMC, UMC, and IBM technologies. The majority of these commercial compilers do not allow the customer to alter the base design, are restricted by the company's license, and usually require a fee. This makes them virtually unavailable and not useful for many academic research projects.

In addition to memory compilers provided by industry, various research groups have released scripts to generate memories. However, these designs are not silicon verified and are usually only composed of simple structures. For example, FabMem [38] is able to create small arrays, but it is highly dependent on the Cadence design tools. The scripts do not provide any characterization capability and cannot easily integrate with commercial place and route tools. Another recent, promising solution for academia is the Synopsys Generic Memory Compiler (GMC) [39]. The software is provided with sample generic libraries such as Synopsys' 32/28 nm and 90 nm abstract technologies and can generate the whole SRAM for these technologies. GMC generates GDSII layout data, SPICE netlists, Verilog and VHDL models, timing/power libraries, and DRC/LVS verification reports. GMC, however, is not recommended for fabrication since the technologies it supports are not real. Its sole purpose is to aid students in Very

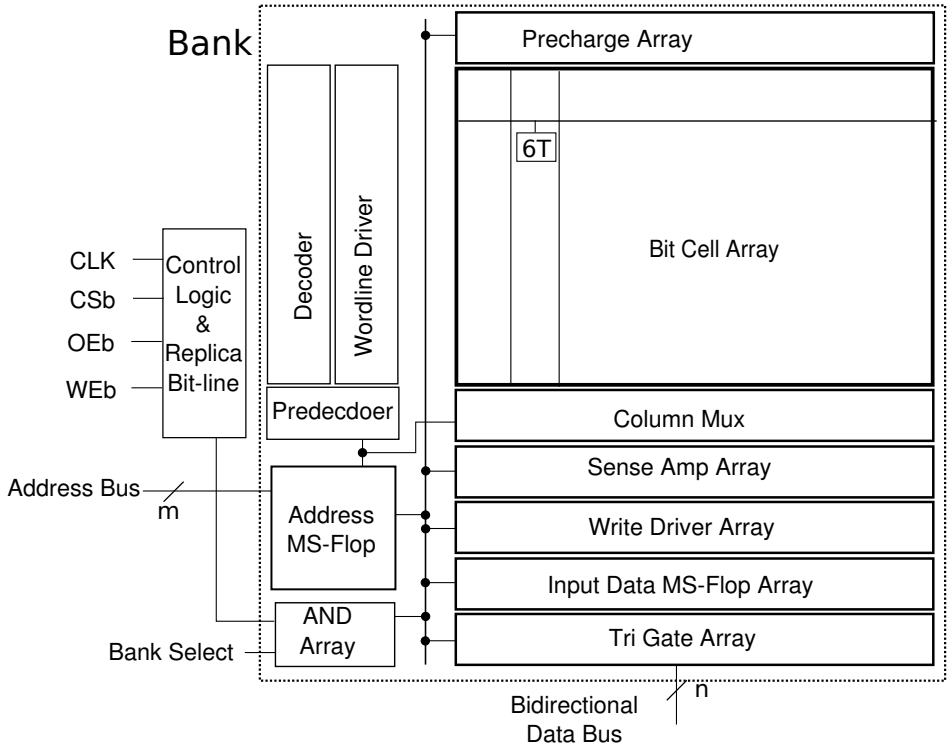


Figure 3.1: OpenRAM SRAM architectures.

Large Scale Integration (VLSI) courses to learn about using memories in design flows.

There have been multiple attempts by academia to implement a memory compiler that is not restricted: the Institute of Microelectronics' SRAM IP Compiler [33], School of Electronic Science and Engineering at Southeast University's Memory IP Compiler [40], and Tsinghua University's Low Power SRAM Compiler [41]. These are all methodologies and design flows for a memory compiler, but there were no public releases to view or to use.

3.3 Architecture

The OpenRAM SRAM architecture is based on a bank of memory bitcells with peripheral circuits and control logic as illustrated in Figure 3.1. These are further refined into eight major blocks: the bitcell array, the address decoder, the wordline drivers, the column multiplexer, the precharge circuitry, the sense amplifier, the write drivers, and the control logic.

Bitcell Array: In the initial release of OpenRAM, the 6T bitcell is the default memory bitcell because it is the most commonly used bitcell in SRAM devices. 6T bitcells are tiled together with

abutting wordline and bitlines to make up the memory array. The bitcell array's aspect ratio is made as square as possible using multiple columns of data words. The memory bitcell is a custom designed library cell for each technology. Other types of memory bitcells, such as 7T, 8T, and 10T bitcells [1, 2, 13, 14], can be used as alternatives to the 6T bitcell.

Address Decoder: The address decoder takes the row address bits as inputs and asserts the appropriate wordline so that the correct memory bitcells can be read from or written to. The address decoder is placed to the left of the memory array and spans the array's vertical length. Different types of decoders can be used such as an included dynamic NAND decoder, but OpenRAM's default option is a hierarchical CMOS decoder.

WordLine Driver: Wordline drivers are inserted between the address decoder and the memory array as buffers. The wordline drivers are sized based on the width of the memory array so that they can drive the row select signal across the bitcell array.

Column Multiplexer: The column multiplexer is an optional block that uses the lower address bits to select the associated word in a row. The column mux is dynamically generated and can be omitted or can have 2 or 4 inputs. Larger column muxes are possible, but are not frequently used in memories. There are options for a single pass transistor mux or a multi-level tree mux.

Bitline Precharge: This circuitry precharges the bitlines during the first phase of the clock for read operations. The precharge circuit is placed on top of every column in the memory array and equalizes the bitline voltages so that the sense amplifier can sense the voltage difference between the two bitlines. Precharge schematic is shown in Figure 2.3(a).

Sense Amplifier: A differential sense amplifier is used to sense the voltage difference between the bitlines of a memory bitcell while a read operation is performed. The sense amplifier uses a bitline isolation technique to increase performance. The sense amplifier circuitry is placed below the column multiplexer or the memory array if no column multiplexer is used. There is a sense amplifier for each output bit. The sense amplifier schematic is shown in Figure 2.3(b).

Write Driver: The write drivers send the input data signals onto the bitlines for a write operation. The write drivers are tristated so that they can be placed between the column multiplexer/memory array and the sense amplifiers. There is a write driver for each input data bit. The write driver schematic is shown in Figure 2.3(c).

Control Logic: The OpenRAM SRAM architecture incorporates a standard synchronous memory

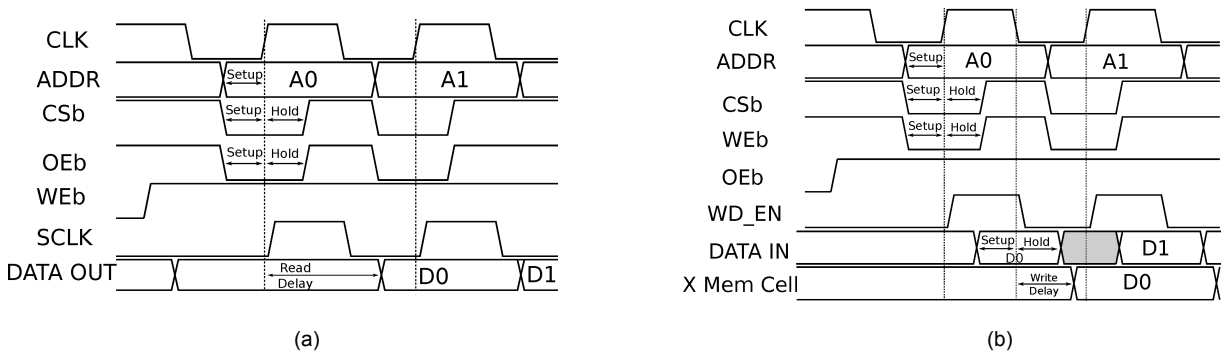


Figure 3.2: Synchronous SRAM interface of OpenRAM.

interface using a system clock (clk). The control logic uses an externally provided output enable (OEb), chip select (CSb), and write enable (WEb) signals to enable the combination of multiple SRAMs into a larger structure. Internally, the OpenRAM compiler can have 1, 2, or 4 memory banks to amortize the control logic and peripheral circuitry. All of the input control signals are stored using master-slave (MS) flip-flops (FF) to ensure that the signals are valid for the entire clock cycle. During a read operation, data is available after the negative clock edge (second half of cycle) as shown in Figure 3.2(a). To avoid dead cycles which degrade performance, a Zero Bus Turn-around (ZBT) technique is used in OpenRAM timing [42]. The ZBT enables higher memory throughput since there are no wait states. During ZBT writes, data is set up before the negative clock edge and is captured on the negative edge. Figure 3.2(b) shows the timing for input signals during the write operation. The internal control signals are generated using a replica bitline (RBL) structure for the timing of the sense amplifier enable and output data storage [43]. The RBL turns on the sense amplifiers at the exact time in presence of process variability in sub-100 nm technologies.

3.4 Implementation

OpenRAM's methodology is implemented using an object-oriented approach in the Python programming language. Python is a simple, yet powerful language that is easy to learn and very human-readable. Moreover, Python enables portability to most operating systems. OpenRAM has no additional dependencies except a DRC/LVS tool, but that is disabled with a warning if the tools are unavailable.

In addition to system portability, OpenRAM is also translatable across numerous process technologies. This is accomplished by using generalized routines to generate the memory based on common

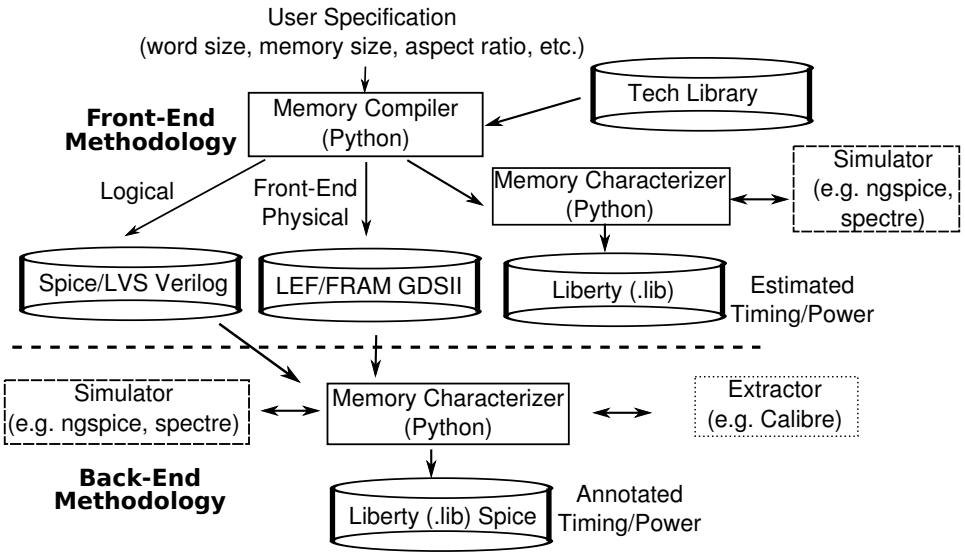


Figure 3.3: Overall compilation and characterization methodology.

features across all technologies. To facilitate user modification and technology interoperability, OpenRAM provides a reference implementation in 45 nm FreePDK45 [44] and a fabricable option using the MOSIS Scalable CMOS (SCMOS) design rules [45]. FreePDK45 uses many design rules found in modern technologies, but is non-fabricable, while SCMOS enables fabrication of designs using the MOSIS foundry services. SCMOS is not confidential and an implementation using it is included, however, it does not include many advanced DSM design rules. OpenRAM has also been ported to other commercial technologies, but these are not directly included due to licensing issues.

OpenRAM's framework is divided into front-end and back-end methodologies as shown in Figure 3.3. The front-end has the compiler and the characterizer. The compiler generates Spice models and its GDSII layouts based on user inputs. The characterizer calls a Spice simulator to produce timing/power results. The back-end uses the generated spice netlists and GDSII layouts to generate annotated timing and power models using back-annotated characterizations.

3.4.1 Base Data Structures

The design modules in OpenRAM are derived from the *design* class (design.py). The design class has a name, a Spice model (netlist), and a layout. Both the Spice model and the layout inherit their capabilities from a hierarchical class. The design class also provides inherited functions to perform DRC and LVS verification of any design. The design class derives from the *spice* class (hierarchy_spice.py) which has

a data structure to maintain the circuit hierarchy. This class maintains the design instances, their pins, and their connections as well as helper functions to maintain the structure and connectivity of the circuit hierarchy.

The design class also derives from a *layout* class (hierarchy_layout.py). This class has list of physical instances of sub-modules in the layout and a structure for simple objects such as shapes and labels in the current hierarchy level. In addition, there are helper functions that maintain the physical layout structures. OpenRAM has an integrated, custom GDSII library to read, write, and manipulate GDSII files. The library, originally called GdsMill [46], has been modified, debugged, and extended for OpenRAM. Full rights were given to include the GdsMill source with OpenRAM, but to make the interfacing easier and porting to other physical layout databases possible, OpenRAM implements a *geometry* wrapper class (geometry.py) that abstracts the GdsMill library.

3.4.2 Technology and Tool Portability

OpenRAM is technology-independent by using a technology directory that includes the technology's specific information, rules, and library cells. Technology parameters such as the design rule check (DRC) rules and the GDS layer map are required to ensure that the dynamically generated designs are DRC clean. Custom designed library cells such as the memory bitcell and the sense amplifier are also placed in this directory. A very simple design rule parameter file has the most important design rules for constructing basic interconnect and transistor devices. FreePDK45 and SCMOS reference libraries are provided.

OpenRAM uses some custom-designed library primitives as technology input. Since density is extremely important, the following cells are pre-designed in each technology: 6T bitcell, sense amplifier, master-slave flip-flop, tri-state gate, and write driver. All other cells are generated on-the-fly using a parameterizable gate primitive.

OpenRAM can be used for various technologies since it creates the basic components of memory designs that are common over these technologies. For technologies that have specific design requirements, such as specialized well contacts, the user can include helper functions in the technology directory. This is done so that the main compiler remains free of dependencies to specific technologies.

OpenRAM has two functions that provide a wrapper interface with DRC and LVS tools. These two functions perform DRC and LVS using the GDSII layout and Spice netlist files. Since each DRC and

LVS tool has different output, this routine is customized per tool to parse DRC/LVS reports and return the number of errors while also outputting debug information. These routines allow flexibility of any DRC/LVS tool, but the default implementation calls Calibre nmDRC and nmLVS. In OpenRAM, both DRC and LVS are performed at all levels of the design hierarchy to enhance bug tracking. DRC and LVS can be disabled for improved run-time or if tool licenses are not available.

3.4.3 Class Hierarchy

1. **Low-Level Classes** : OpenRAM provides parameterized transistor and logic gate classes that help with technology portability. These classes generate a technology-specific transistor and simple logic gate layouts so that many modules do not rely on library cells. It is also used when a module such as the write driver needs transistor sizing to optimize performance. The parameterized transistor (ptx.py) generates a basic transistor of specified type and size. The parameterized transistor class is used to provide several parameterized gates including pinv.py, nand2.py, nand3.py, and nor2.py.

2. **Top-Level Classes** : The *openram* class (openram.py) organizes execution of the program and instantiates a single memory design in the *sram* class while saving the resulting design files. It accepts user-provided parameters to generate the design. The *sram* class (sram.py) decides the appropriate internal parameter dependencies shown in Table 3.1. They are dependent on the user-desired data word size, number of words, and number of banks. It is responsible for instantiation of the single control logic module which controls the SRAM banks. The control logic ensures that only one bank is active in a given address range. The *bank* class (bank.py) does the bulk of the non-control memory layout. It instantiates 1, 2, or 4 bitcell arrays and coordinates the row and column address decoders along with their precharge, sense amplifiers, and input/output data flops. Every other block in the memory design has a class for its base cell (e.g., sense_amplifier.py) and an array class (e.g., sense_amplifier_array.py) that is responsible for tiling the base cell. Each class is responsible for physically placing and logically connecting its own sub-circuits while passing its dimensions and port locations up to higher-level modules.

Table 3.1: Dependencies required for sub-modules.

Variable	Equation
Total Bits	$word\_size * num\_words$
Words Per Row	$\sqrt{(num\_words)/word\_size}$
Num of Rows	$num\_words/words\_per\_row$
Num of Cols	$words\_per\_row * word\_size$
Col Addr Size	$\log_2(words\_per\_row)$
Row Addr Size	$\log_2(num\_of\_rows)$
Total Addr Size	$row\_addr\_size + col\_addr\_size$
Data Size	$word\_size$
Num of Bank	num_banks

3.4.4 Characterization

OpenRAM includes a memory characterizer that measures the timing and power characteristics through Spice simulation. The characterizer has four main stages: generating the Spice stimulus, running the circuit simulations, parsing the simulator's output, and producing the characteristics in a Liberty (.lib) file.

The stimulus is written in standard Spice format and can be used with any simulator that supports this. The stimulus only uses the interface of the memory (e.g., bi-directional data bus, address bus, and control signals) to perform black box timing measurements. Results from simulations are used to produce the average power, setup/hold times, and timing delay of the memory design. Setup and hold times are obtained by analyzing the flip-flop library cell in the technology. These setup and hold times are equivalent because OpenRAM has a completely synchronous input interface. The setup time, hold time, and delay are found using a bidirectional search technique.

3.4.5 Unit Tests

Probably the most important feature of OpenRAM is the set of thorough regression tests implemented with the Python unit test framework. These unit tests allow users to add features without worrying about breaking functionality. They also guide users when porting to new technologies. Every sub-module has its own regression test and there are also regression tests for memory functionality, library cell verification, timing verification, and technology verification.

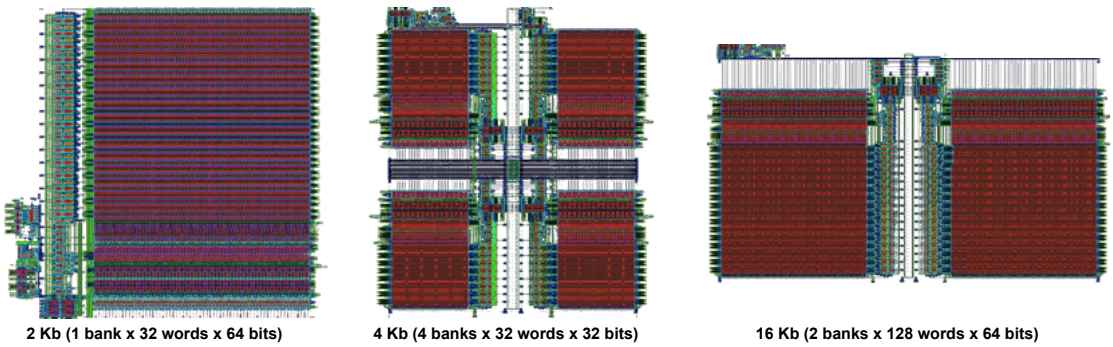


Figure 3.4: Symmetrical placement of single and multi-bank SRAMs in OpenRAM.

3.5 Results

Figure 3.4 shows several different SRAM layouts generated by OpenRAM in FreePDK45. OpenRAM can generate single bank and multi-bank SRAM arrays. Banks are symmetrically placed to have the same delay for data, address and control signals.

Figure 3.5 shows the memory area of different total size and data word width in FreePDK45 and SCMOS, respectively. As expected, the smaller process technology (45 nm) has lower total area overall but the trends are similar in both technologies. Figure 3.5 also shows the access time of different size and data word width in FreePDK45 and SCMOS, respectively. Increasing the memory size generally increases the access time; long bitline and wordline increase the access time by adding more parasitic capacitance and resistance, therefore having shorter bitline and wordline helps to speed up the read operation. Since OpenRAM uses multiple banks and column muxing, it is possible to have a smaller access time for larger memory designs, but this will sacrifice density.

Table 3.2 compares the bit-density of OpenRAM against published designs using similar technology nodes. The results show the benefit of technology scaling and that OpenRAM has very good density in both technologies. Comparison of power consumption and read access time, however, are a bit more complicated to make a conclusion, because there are many trade-offs. Power and performance are highly dependent on circuit style (CMOS, ECL, etc.), memory organization (more banks be faster but sacrifices density), and the optimization goal: low-power or high-performance. In general, OpenRAM has reasonable trade-off between the two and can be customized by using an alternate sense amplifier or organization. As a comparison, a 76 ns SRAM consumes 3.9 mW [47] while OpenRAM is much faster at 44.9 ns but consumes 115 mW for the same size.

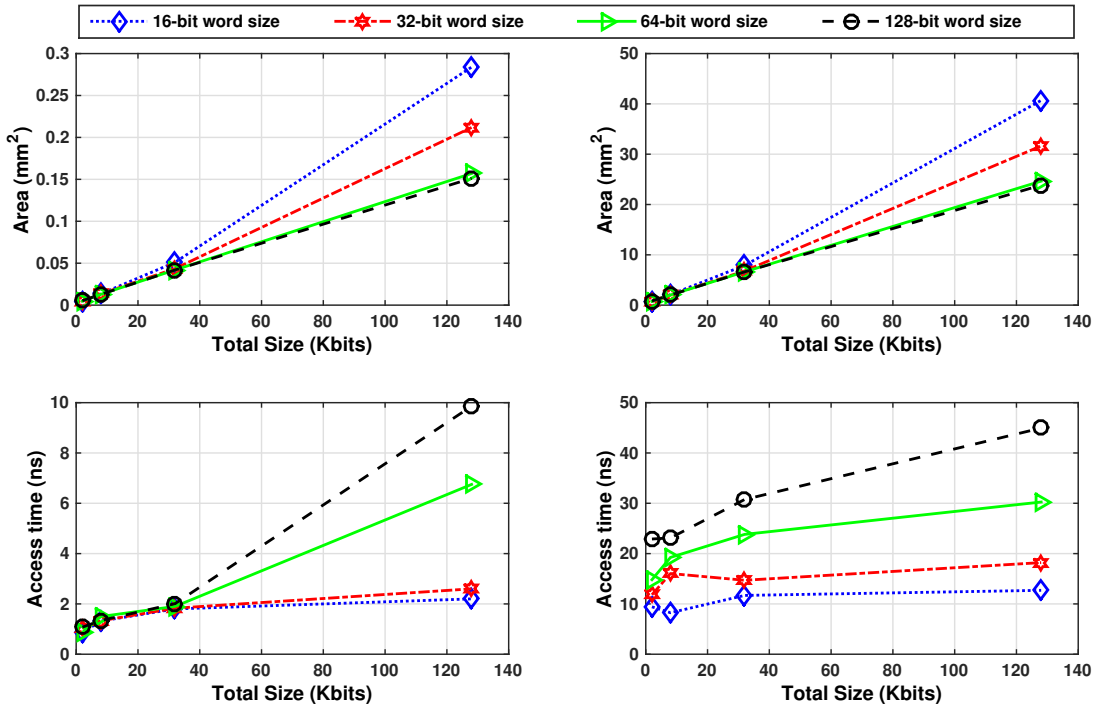


Figure 3.5: High-density and fast memories generated by OpenRAM.

Table 3.2: OpenRAM has high density compared to published memories in similar technologies.

Ref.	**Feature Size**	**Technology**	**Density [Mb/mm^2]**
[**48**]	65 nm	CMOS	0.7700
[**49**]	45 nm	CMOS	0.3300
[**50**]	40 nm	CMOS	0.9400
OpenRAM	45 nm	FreePDK45	0.8260
[**51**]	0.5 um	CMOS	0.0036
[**52**]	0.5 um	BiCMOS	0.0020
[**47**]	0.5 um	CMOS	0.0050
OpenRAM	0.5 um	SCMOS	0.0050

3.6 Summary and Conclusions

This Chapter introduced OpenRAM, an open-source and portable memory compiler. OpenRAM generates the circuit, functional model, and layout of variable-sized SRAMs in a generic 45 nm technology and fabricable Scalable CMOS (SCMOS), but it has also been ported to several commercial technology nodes using a simple technology file. In addition, a memory characterizer provides synthesis timing/power models. The main motivation behind OpenRAM is to promote and simplify memory-related research in academia. Since OpenRAM is open-sourced, flexible, and portable, this memory compiler can be adapted to various technologies and is easily modified to address specific design requirements. Therefore, OpenRAM provides a platform to implement and test new memory designs. In this Chapter implementation and main features of the OpenRAM memory compiler are introduced and an analysis of the area, timing and power is shown for different sizes and technologies of memory.

Chapter 4

A Single-Port 8T SRAM Bitcell for Noise-Margin Improvement

4.1 Introduction

SRAM scaling is one of the major bottlenecks for the reduction of supply voltages in current and future CMOS technology nodes [49]. Threshold voltage (V_{TH}) sensitivity to process variation reduces SRAM's stability during read and write operations [53]. Unfortunately, V_{TH} variation has a large impact on the stability of small size transistors inside a SRAM bitcell, because transistor current is sensitive to V_{TH} variation [54]. And, maintaining sufficient Static Noise Margin (SNM) becomes difficult in scaled SRAMs due to the dramatic mismatch between the reflected SNMs of the two halves of a SRAM bitcell. Therefore, with an increase in process variation for lower supply voltages, it is becoming difficult to balance the read and write stability for a 6T bitcell. This is because the read stability and writability have conflicting design requirements; that is, for stability during a read, the storage inverters must be stronger than the access-transistors inside SRAM architectures. And, the opposite is desired for a bitcell's writability: a weak storage inverter and strong access-transistors. These two conditions cannot be simultaneously optimized in low voltages due to an increase in process variations.

To avoid this problem, researchers have considered different configurations for SRAM bitcells and several 8T [55, 56, 57, 58, 59] to 10T [60, 61, 62] SRAM bitcells that can be categorized into single-ended and differential-ended sensing bitcells [2], [13] have been proposed. Generally, single-ended

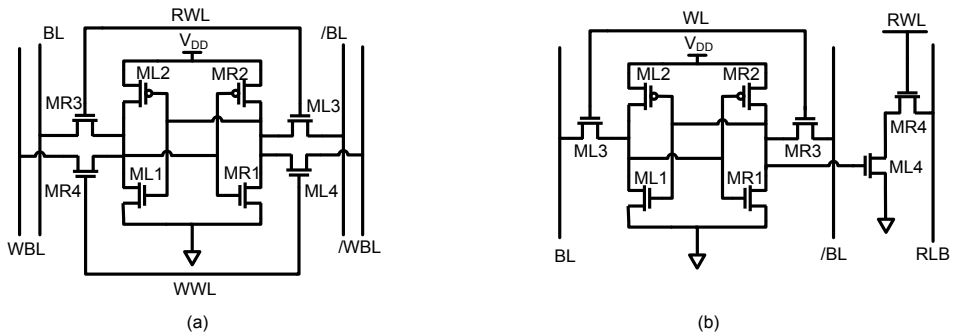


Figure 4.1: (a) Dual port 8T bitcell [1], (b) single-ended 8T bitcell [2].

sensing is not as robust as a differential-ended one. On the other hand, differential-ended sensing utilizes extra ports and more than one wordline, therefore, this consists of additional space within its layout which could consume more area compared to a traditional 6T bitcell. Besides, these bitcells require extra circuitry to be controlled that potentially could lead to an increase in power and as well as a larger area penalty.

In this Chapter, a differential single port 8T SRAM bitcell is proposed that provides greater improvements in stability during read and write access, thus, possibly better facilitating technology scaling in SRAM design. The proposed 8T bitcell enhances the writability and read stability by improving the strength of pull down transistors during read operation and the access transistors during a write operation.

4.2 Background

Figure 4.1 shows the schematics of commonly-used 8T SRAM bitcells in previous implementations ([2] and [1]). In Figure 4.1(a), a differential-ended 8T bitcell makes use of the voltage difference between BL pair during a read operation that is suitable for high-speed applications [1]. This SRAM bitcell consumes approximately two times the area of a 6T bitcell, because it requires extra ports and has dead-space within its layout. For this dual-port bitcell, the read and write ports can be operated in parallel, however, this causes a read-disturbance issue, in which the bitcell current is degraded when the read and write ports access the same row simultaneously [63].

To prevent read disturbance issues, a single-ended read-port 8T bitcell architecture is shown in Figure 4.1(b) that eliminates charge sharing between the bitline and internal storage nodes [2]. While this

bitcell significantly improves the SRAM stability in low voltages, it suffers from a reduced swing on bitlines due to leakage as well as poor noise immunity due to its single-ended structure. In addition, the read operation in the single-ended bitcell utilizes a full swing of the single read bitline, so an improvement of the access time is not expected and read operation is slow because of its full rail sensing.

Generally, single-ended bitcells are not as robust as the differential-ended ones, therefore, they often requires some extra circuitry to maintain their reliability. In both differential-ended and single-ended 8T bitcells, read and write separation allows each to be independently optimized. However, since both these designs have more than one wordline and one bitline pair for a bitcell, they require extra circuitry to be controlled which leads to an enhanced power and area penalty compared to the 6T bitcell. As both these bitcells have limitations, there is a need for new SRAM bitcell designs that enhance writability and read stability while achieving smaller access time, good noise immunity and less area overhead.

4.3 Proposed 8T SRAM Bitcell

Figure 4.2 shows the schematic and one possible layout of the proposed 8T bitcell. As shown in Figure 4.2(a) for the proposed 8T SRAM bitcell, two additional transistors (ML4, MR4) are added between the pull down (ML1, MR1) and access transistors (ML3, MR3). The gates of ML4 and MR4 are connected to /Q and Q and their source are tied to $CTRL1$ and $CTRL2$ signals; these two transistors change their operation according to $CTRL1$ and $CTRL2$ voltages. That is, $CTRL1$ and $CTRL2$ are changed in accordance with operations; they are connected to GND during the read mode to improve the read current and enlarge the SNM. In the write mode, $CTRL1$ and $CTRL2$ are connected to BL and /BL, respectively, which assists the write operation and WNM. Therefore, ML4 and MR4 are considered as pull down transistors during the read operation and the access transistors during the write operation. The basis behind this circuit comes from idea of the 10T SRAM in [14], however, the design presented here is significant enhanced to an 8T structure by allowing the ML4 and MR4 sources to track the $CTRL1$ and $CTRL2$ signals based on the operation.

Figure 4.2(b) shows one possible layout for the proposed 8T bitcell. The proposed 8T bitcell is compact and it is 1.3x the area of a conventional 6T bitcell. V_{DD}, GND, BL, /BL, WL, $CTRL$ and /$CTRL$ contacts between adjacent bitcells are all pitch-matched accordingly. It is worth mentioning that although the single-ended 8T bitcell [2] is an area efficient bitcell, its asymmetric device placement

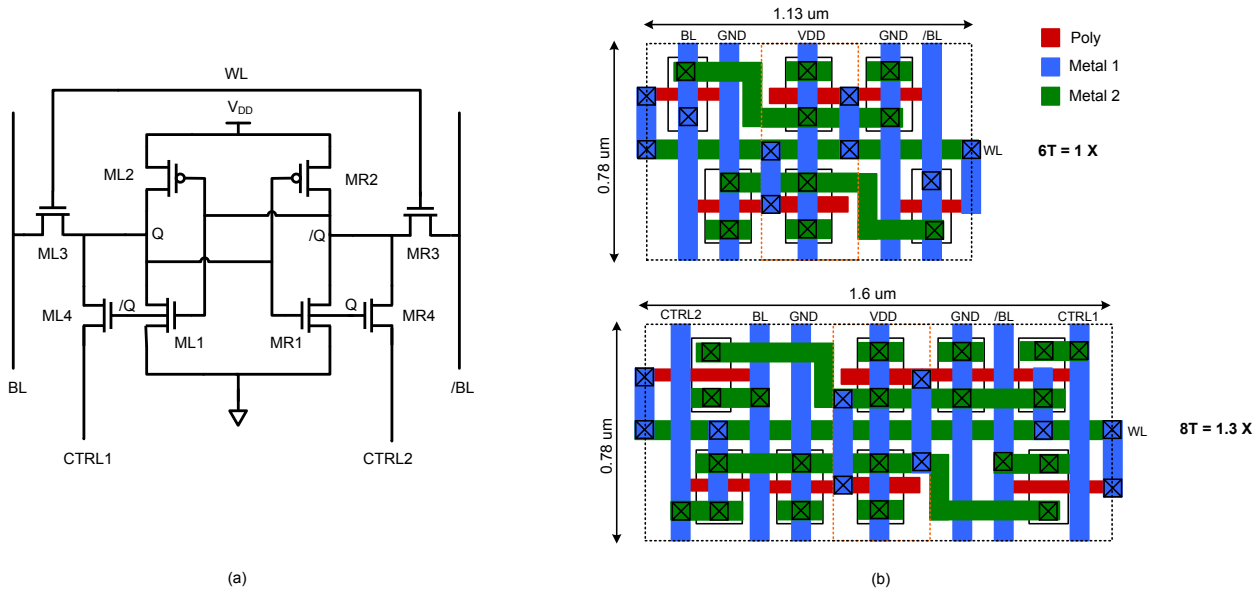


Figure 4.2: (a) Proposed 8T bitcell and (b) one possible 1.25 μm^2 layout in a 32 nm technology.

in the layout can degrade the SNM and WNM. On the other hand, the proposed 8T bitcell gives a better noise immunity, because of its differential operation. It also has symmetric device placement in its layout while providing high area efficiency. Most importantly, it requires no changes compared to a 6T SRAM architecture. Therefore, the proposed 8T bitcell can be used in the present 6T based memory compilers like OpenRAM [64].

The operating principle of this bitcell is illustrated in the timing diagram in Figure 4.3(a). During a read operation, the decoder selects the WL and $CTRL1$ and $CTRL2$ are set to GND. Based on the bitcell stored value, one of the bitlines discharges. If the node Q stores a 0, the MR1 and MR4 will remain off and /BL will stay at V_{DD} during read mode. In this case, node /Q stores a 1 so ML1 and MR4 are turned on and BL will be discharged. Voltage difference between the BL and /BL can be sensed by the sense amplifier to generate the output data. In the proposed 8T bitcell, two current paths exist during a read operation, as shown in Figure 4.3(b): the ML1 current and the dotted line is the ML4 current to GND. Consequently, high speed operation can be achieved by adding the ML4 and MR4 during a read operation. Additionally, the current flow for the two parallel transistors results in increased read current. On the other hand, the bitline discharging speed has a direct impact on the SRAM access time. The proposed 8T bitcell has a faster access time compare to a 6T bitcell, because the extra pull-down transistors are asserted during its read operation.

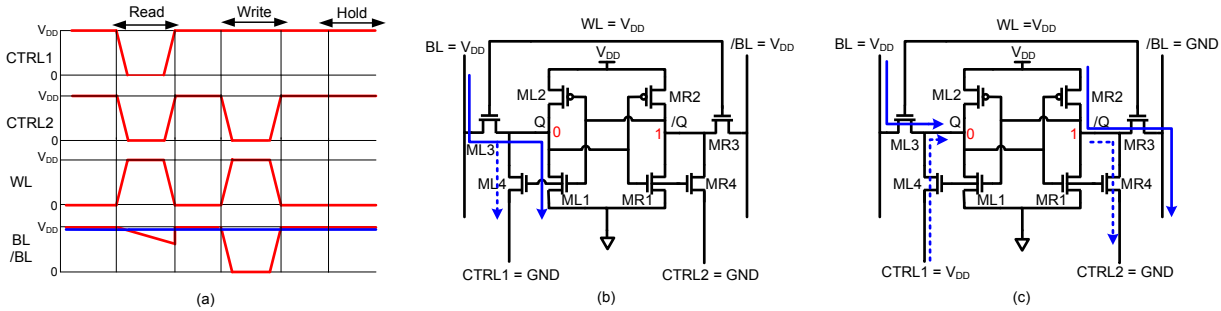


Figure 4.3: (a) Timing diagram, (b) read current paths and (c) write current paths for 8T bitcell.

Figure 4.3(c) shows the current paths during a write mode. As shown in this figure, nodes, Q and /Q, have the values of 0 and 1, respectively and they are written over by charging and discharging of Q and /Q through the ML3 and MR3. In addition, another current from ML4 and MR4 (dotted lines) help to enlarge the write margin in the proposed 8T bitcell by keeping the $CTRL1$ and $CTRL2$ at the same voltages of BL and /BL. In standby mode, BL and /BL stays high, as well as $CTRL1$ and $CTRL2$. The signals $CTRL1$ and $CTRL2$ are changed in accordance with its operation; that is, they are connected to GND in read mode to improve the readout current and during a write operation, $CTRL1$ and $CTRL2$ has the equal value of BL and /BL, which helps the write operation. During its hold mode, $CTRL1$ and $CTRL2$ stay high to limit unnecessary leakage current. In general, this new SRAM bitcell design offers several merits over single-ended and dual-port 8T bitcells; (1) proposed bitcell uses only one single wordline for both read and write operations and unlike the single-ended 8T bitcell, it has a differential-ended structure for better noise immunity. (2) This new 8T bitcell is resilient, since it has a higher SNM and WNM during its read and write operations; plus, the read stability and the writability does not have contrary requirements in the proposed 8T bitcell.

4.4 Comparison of Proposed 8T, Single-Ended 8T, Dual-Port 8T and 6T SRAM Bitcells

The SNM, WNM and readout bitcell current of the proposed 8T, single-ended 8T [2], dual-port 8T [1] and conventional 6T bitcells are compared in the following subsections. Here, all of the comparisons are simulated by re-creating the circuits from scratch and the results stem from the simulations using 32 nm technology. The following transistor sizing guideline is used to find the optimum size for each bitcell; (1)

Table 4.1: Transistor sizing for optimized bitcell area in 32 nm (W/L) [nm].

Bitcell	ML1, MR1	ML2, MR2	ML3, MR3	ML4, MR4
Standard 6T	$208/40$	$104/40$	$156/40$	$-$
Single-ended 8T	$208/40$	$104/40$	$156/40$	$104/40$
Dual-port 8T	$208/40$	$104/40$	$104/40$	$156/40$
Proposed 8T	$208/40$	$104/40$	$104/40$	$104/40$

The pull-up transistors are made weaker to ease the write operation. (2) The pull down transistors choose stronger to facilitate larger SNM and also faster read operation. (3) The access transistors are chosen strong enough to ensure proper write operation. (4) ML4 and MR4 in the proposed 8T bitcell need to be strong enough, since they are involved in the bitcell discharging paths. However, their strength is made comparable to the access transistors to achieve both large SNM and WNM. In summary, all device sizes for the bitcells used in this comparison are as shown in Table 4.1.

4.4.1 Static Noise Margin

The SNM is defined as the maximum possible noise available at the gates of the cross-coupled inverters or storage element that does not flip the bitcell value [11]. Interestingly, adding transistors makes reading the bitcell current become larger inside the 8T bitcell compared to the 6T bitcell, which improves the bitline discharging speed and SRAM access time. Figure 4.4(a) shows the read VTC curves for the proposed 8T, dual-port 8T, single-ended 8T and conventional-6T bitcells at 900 mV and room temperature. Figure 4.5(a) shows the distribution of the SNMs in presence of local and global variations for a $1,000$ point Monte Carlo (MC) simulation at a 900 mV supply voltage. As shown in Figure 4.4(a) and 4.5(a), the single-ended 8T and conventional-6T bitcells have the highest and lowest SNM, respectively. The mean SNM for the proposed 8T bitcell is 170 mV, which is 48% higher than the conventional-6T bitcell. The deviation of the proposed 8T bitcell read SNM is smaller than other bitcells indicate the proposed 8T bitcell has an improved variation tolerance.

4.4.2 Write Noise Margin

The WNM measures how easy or difficult it is to write into the bitcell; it is the highest BL potential that can flip the bitcell data [11]. Figure 4.4(b) shows how the WNM of proposed 8T bitcell is remarkably improved compared to other bitcells, because the strength of access transistor is improved through

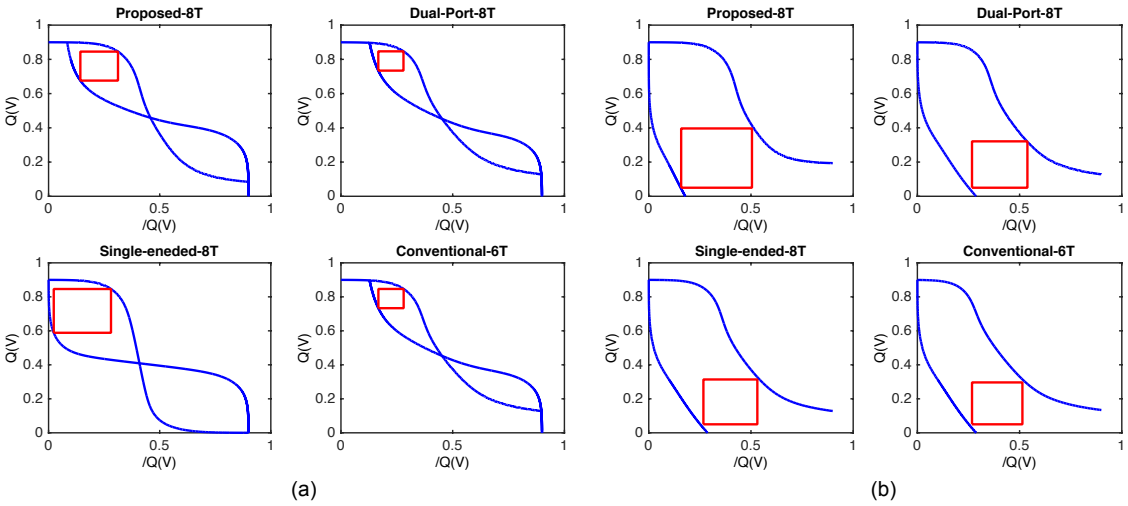


Figure 4.4: VTC curves for (a) SNM and (b) WNM comparison at V_{DD} = 0.9 V.

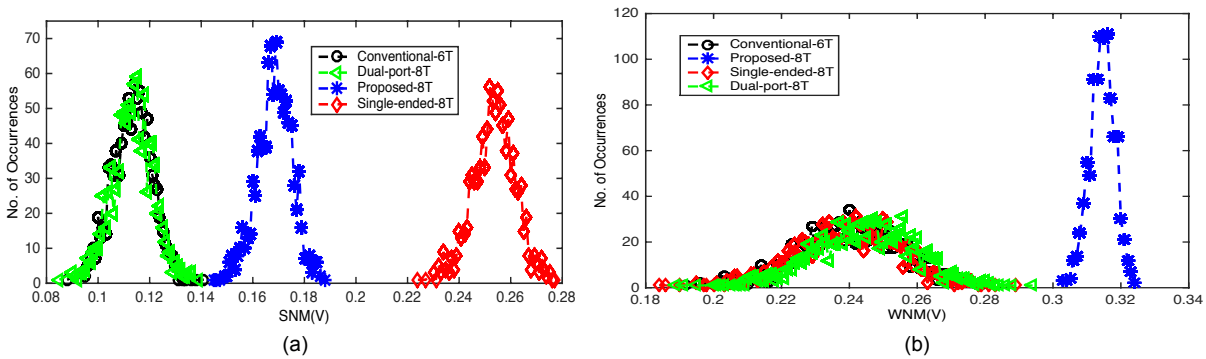


Figure 4.5: (a) SNM and (b) WNM MC simulation results (V_{DD} = 0.9 V, T = 25^{o}C).

adding one more transistors during charging and discharging path operation, as shown in Figure 4.3(c). By adding more transistors to help the write operation, the proposed 8T bitcell gains the largest WNM (= 320 mV) in the same supply voltage (= 900 mV) and environmental conditions. Figure 4.5(b) shows the WNM of the proposed 8T bitcell. A 1,000 point MC simulation is run for mentioned bitcells to derive the WNM distribution and show the impact of local and global V_{TH} variation on the write mode. As shown in this figure, the proposed 8T bitcell increases the WNM by 33% compare to other bitcells while it has the smallest deviation which gives the proposed 8T bitcell better variation tolerance.

4.4.3 Readout Bitcell Current

Another important property of a SRAM bitcell is its readout current. In order to have a fast and reliable read operation, it is desirable to have a large readout bitcell current. The comparison of the readout

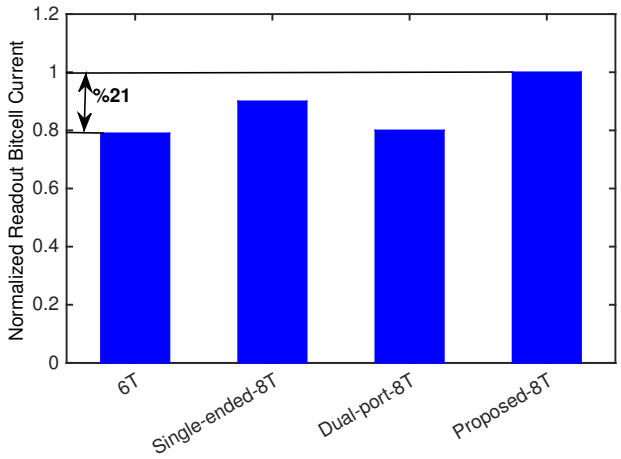


Figure 4.6: Readout bitcell current comparison.

Table 4.2: Specifications of proposed 8T bitcell.

Technology	IBM/Global Foundries cmos32soi 32 nm
Supply Voltage	900 mV
Bitcell layout Size	1.25 um^2
WNM	320 mV
SNM	170 mV

bitcell currents are shown in Figure 4.6. The proposed 8T bitcell increases the bitcell current by 21% compare to 6T bitcell at 900 mV, since the two pulldown transistors are asserted in the read operation. Thus, the proposed 8T bitcell achieves a faster access time than the standard 6T and other 8T SRAM bitcells.

4.4.4 Evaluation Under Process, Temperature and Voltage Variation

To evaluate the efficiency of the proposed 8T bitcell under process, temperature and voltage (PVT) variations, MC simulations are done for both read and write modes at different process corners, different supply voltages and different temperatures. As shown in Figure 4.7, the proposed 8T bitcell has a robust performance in all process corners as well as a wide temperature operability. Besides, the proposed 8T has an acceptable SNM and WNM for smaller supply voltages, below the V_{TH}, where 6T and dual-port 8T cannot be used and single-ended 8T have limitations in WNM. As shown in Figure 4.5, the WNM of dual-port 8T, single-ended 8T and 6T bitcells are almost on top of each other, since all these bitcells have same WNM values and distributions. Table 4.2 summarizes the specifications of proposed 8T bitcell.

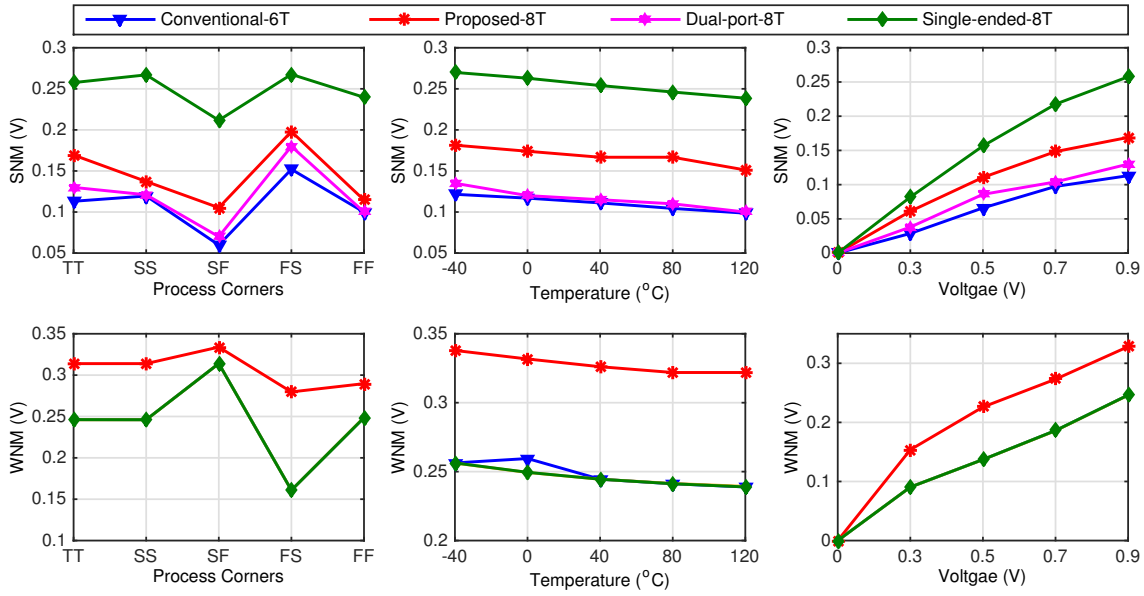


Figure 4.7: Effect of PVT variations on SNM and WNM.

4.5 Multi-Threshold SRAM Design

As feature sizes scale down, leakage and dynamic power consumptions become comparable in high performance SRAM circuits [65]. For years, performance has been considered as the most important factor for SRAM design that leads to non power-efficient memories. Besides, since SRAMs are dense structures and leakage-power is proportional to the number of transistors, large SRAM arrays power consumption tend to increase exponentially [65]. Therefore, there is a need for new SRAM designs to control the leakage-power in sub-100 nm technologies.

There is a wide research from bitcell design to various circuit techniques and SRAM architectures on leakage-power reduction in order to reduce the total power consumption of SRAM circuits. Among the bitcell techniques, body-biasing, although effective for leakage control [66] predicted to become less efficient for technology nodes below 32 nm [65]. Among the circuit techniques which try to reduce the bitline leakage-current, [67] proposes a floating bitline scheme to reduce the leakage-current. However, this technique suffers from performance reduction due to adding an extra precharge phase before read operation. In [68] wordlines are inactivated by negative voltage to cutoff bitline leakage-current. Unfortunately, this technique degrades reliability of the bitcell by over stressing the gate oxide for the access-transistors. On the other hand, [69] uses a power-line control to realize specific low-power operations, however, this increases SRAM area and leads to circuit complexity. Among all leakage con-

trol techniques, multi-threshold bitcell not only shows good reduction in leakage-current, the accurate selection of V_{TH} for each device in a bitcell helps to maintain the SRAM performance [70, 71].

In this section low-leakage and high performance SRAM design using dual-V_{TH} assignment for the proposed 8T bitcell considering process-induced V_{TH} variations are investigated. Multi-V_{TH} technologies bring more flexibility in SRAM design to achieve better data stability. Multi-V_{TH} technique offers no area overhead for saving power compared to other low-leakage techniques [69] which apply extra circuitry to save power.

It is possible to have symmetric and asymmetric configurations of dual-V_{TH} transistors in a 8T bitcell. However, only different symmetric configurations of dual-V_{TH} 8T bitcell are compared under process-induced variations based on read/write margin (stability and write-ability), read delay (access time) and leakage-current (power). Using Monte-Carlo analysis, the impact of process variations on the characteristics of dual-V_{TH} configurations of 8T bitcell is evaluated and a trade-off plot between read/write margin, access time and leakage-current of different configurations is built to show that the choice of best dual-V_{TH} configuration is based on SRAM application.

4.6 Dual-V_{TH} 8T Bitcell

Supply voltage and device selection are two major factors to determine the SRAM power consumption. Using subthreshold and also dual-V_{TH} devices, it is possible to minimize power consumption in SRAM circuits. The total power of SRAM is summation of dynamic (switching) and static (leakage) power during hold, read and write mode. Using dual-V_{TH} devices, it is possible to reduce the leakage-power of 8T bitcell by fabricating some of the transistors with higher V_{TH}. Higher V_{TH} in transistors reduce the leakage-power and, hence, the power consumption. Subthreshold leakage-current occurs in two different paths in 8T bitcell, as shown in Figure 4.8. One path is inside the bitcell and denotes the cell leakage-current (solid line) and other path is the bitline leakage-current through the access-transistors (dotted line). The stored data and type of operation (voltage values on wordline and bitlines) contribute to the amount of leakage-current in a SRAM. Since differential-ended single-port 8T bitcell has a symmetric structure, there is always some bitcell or bitline leakage-current, regardless of stored data and operation mode.

Although high-V_{TH} can reduces the subthreshold leakage-current, because of the trade-off between

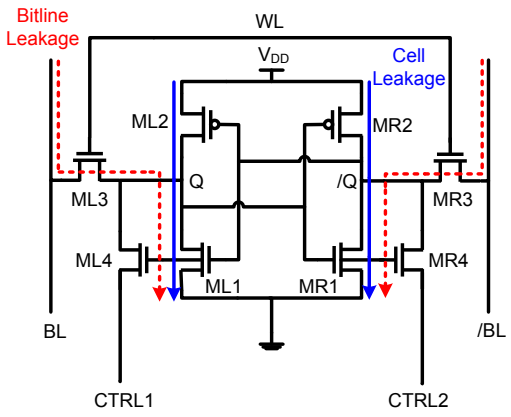


Figure 4.8: Leakage current paths in 8T bitcell.

leakage and delay this will lead to higher access time for 8T bitcell. Therefore it is preferred to use low-V_{TH} transistors in critical paths and high-V_{TH} in non-critical paths to reduce subthreshold leakage-current and maintain performance. Using dual-V_{TH} technology it is possible to balance speed, power and data stability in design of SRAM.

Table in Figure 4.9(a) shows different dual-V_{TH} configurations. Each row in this table corresponds to one configuration by identifying the transistors with high-V_{TH}. In this table, all configurations (C1 to C16) have symmetrically positioned high-V_{TH} transistors and there is no asymmetric configurations. Asymmetric bitcells with low and high-V_{TH} transistors need particular sense amplifier that matches the delay between slow and fast sides of a bitcell due to difference between discharge times for BL and /BL so they are not considered in this dissertation. Figure 4.9(b) shows one possible symmetric dual-V_{TH} 8T bitcell. In this bitcell access and extra transistors have high-V_{TH} and other transistors are implemented with regular-V_{TH}.

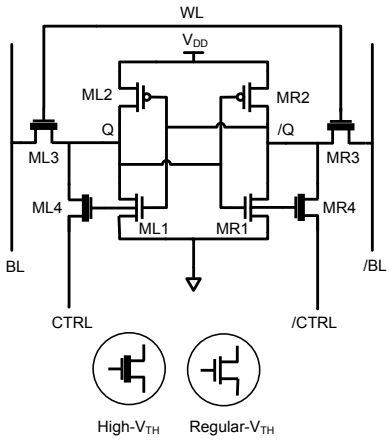
High-V_{TH} and regular-V_{TH} transistors in the 32 nm Silicon-On-Insulator (SOI) models, have V_{TH} values of 0.507 V and 0.293 V for the NMOS and -0.454 V and -0.266 V for PMOS transistors. It is mandatory to decide which transistors can be made low-leakage, since replacement of all transistors with high-V_{TH} can degrade performance.

4.7 Evaluation and Comparison of Different 8T Bitcell Configurations

Process-induced variations cause variability in characteristics of transistors and, hence, 8T SRAM bit-cells. Data stability, read access time and power consumption are among important characteristics of

Configuration	High – VTH transistors
C1	None
C2	ML1, MR1
C3	ML2, MR2
C4	ML3, MR3
C5	ML4, MR4
C6	ML1, MR1, ML2, MR2
C7	ML1, MR1, ML3, MR3
C8	ML1, MR1, ML4, MR4
C9	ML2, MR2, ML3, MR3
C10	ML2, MR2, ML4, MR4
C11	ML3, MR3, ML4, MR4
C12	ML1, MR1, ML2, MR2, ML3, MR3
C13	ML1, MR1, ML2, MR2, ML4, MR4
C14	ML1, MR1, ML3, MR3, ML4, MR4
C15	ML2, MR2, ML3, MR3, ML4, MR4
C16	All

(a)



(b)

Figure 4.9: (a) Table of all configuration and (b) C11 configuration.

SRAM bicells that are subjected to process variations in sub 100 nm technology nodes. In this section the effect of process fluctuations on stability, access time, and leakage-currents of all dual-V_{TH} 8T configurations are evaluated through MC simulations in 32 nm technology.

Assuming a Gaussian distribution for V_{TH}, SNM of all 16 configurations of 8T bitcell are evaluated through MC simulations. Figure 4.10 compares the normalized SNM of all 16 configurations. It is found that among the configurations C4, C11 and C14 have higher SNM.

The subthreshold leakage-current exponentially reduces with a higher threshold voltage [65]. It has been shown that the leakage-power consumed by a SRAM bitcells is data-dependent [65]. In a data storage scenario, all the memory bitcells in a column store the value 1 and in the second scenario all the memory bitcells in a column store the value 0. Measurements and analysis show that read bitline leakage-current is lower when all the memory bitcells store a value 1 as compared to the scenario when all the memory bitcells store a value 0 [65]. For analyzing the impact of using dual-V_{TH} transistors in 8T bitcells and its leakage-power, an all zero-stored data scenario is chosen here to show the worst case leakage-power, as shown in Figure 4.11. Figure 4.10 compares the normalized leakage-power for all 16 configurations in a 256×256 bit SRAM array and worst condition. As expected, C1 configuration has the highest leakage-power because of using regular-V_{TH} transistors. On the other hand, C16 configuration with 8 high-V_{TH} transistors has the lowest leakage-power. As shown in Figure 4.10, the leakage-power of an array with the all high-V_{TH} configuration (C16) has a reduction by 87% compared to the array with all regular-V_{TH} transistors (C1). Also, the configuration with high-V_{TH} pull-up and

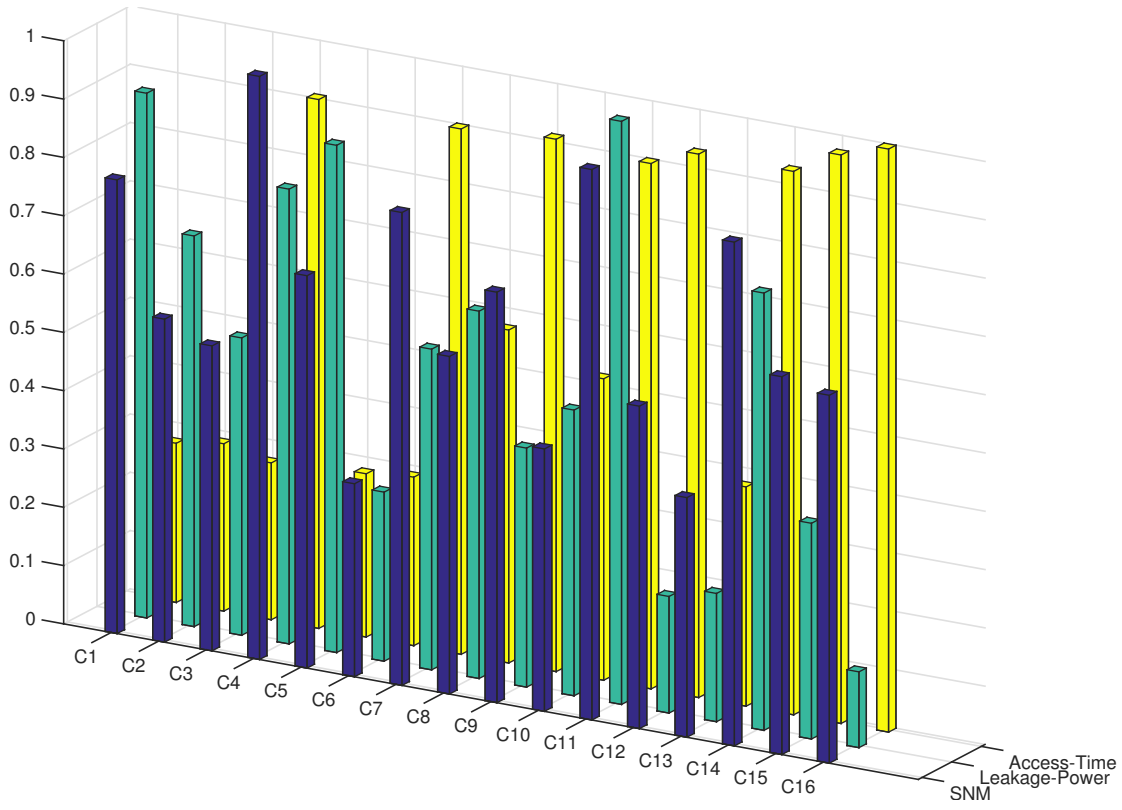


Figure 4.10: Trade-off plot for different V_{TH} configurations of 8T bitcell.

pull-down transistors (C6, C12, C13) result in smaller leakage power, because having high-V_{TH} pull-up and pull-down transistors helps to reduce the cell-leakage, hence, the total leakage power.

In SRAM, a read operation is the most time consuming operation. Read delay (access-time) is measured as the time interval from 50% of a low-to-high transition of a wordline signal until there is a 200 mV differential swing on the bitlines. A 200 mV is the bitline differential voltage when the sense amplifier activates which is greater than the offset voltage of the input transistors in the sense amplifier to guarantee a correct read operation. The normalized read access time of sixteen different configurations are compared in Figure 4.10. As shown in this Figure, C1 which is 72% faster than C16, provides the fastest read operation and C16 provides the slowest one. Configurations with high-V_{TH} access-transistors show a larger read delay because these transistors are in the read path. Although, the SRAM array with the specific C16 configuration is 3.6 times slower than the array with C1 configuration, its leakage power is 7.8 times smaller. This results clearly show the trade-off between leakage-power and access-time in SRAM designs.

Figure 4.10 shows the use of dual-V_{TH} 8T bitcell is application-dependent; for high speed appli-

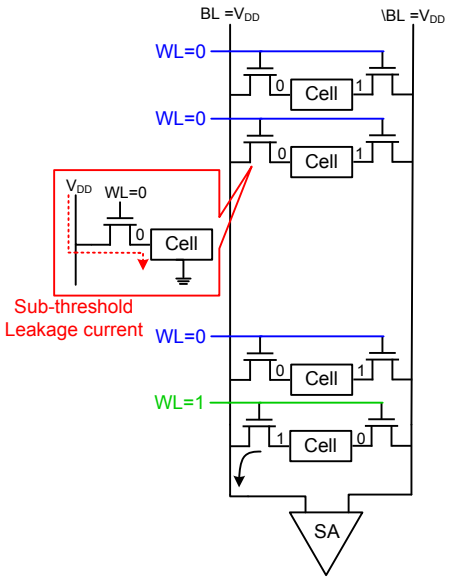


Figure 4.11: Worst case data storage scenario which leads to maximum leakage-current.

cations where power is not an issue C1, C2, C3, C5 and C6 are good options while for higher energy efficiency and battery operated appliance, such as biomedical devices, where speed is not critical, C12, C13 and C16 can be good choices. It is important to note that the Figure-Of-Merit (FOM) in Equation 4.1 can be utilized to evaluate the overall performance of a SRAM circuit:

$$FOM \approx \frac{SNM}{P_{leakage} \times T_{access}} \tag{4.1}$$

where the SNM is the static noise margin, $P_{leakage}$ is the subthreshold leakage-power and T_{access} is the read access time. As all 16 configuration have same area, layout is not a concern to define the FOM. Comparing the FOM of 16 SRAM memory circuits shows that C13 shows the highest FOM for having a relatively fast, power efficient and stable SRAM. This configuration is used in 64 kb SRAM array shown in Figure 4.12. In order to get accurate signal timing for read operation, SRAM structure uses multi replica bitline delay [3] to control the timing of latch-type sense amplifiers. Details on design and operation of multi replica bitline delay technique is presented in chapter 5. The word width is 64 bit, therefore, a 4:1 column-multiplexer is used to choose the selected word based on the input address. The $CTRL1$ and $CTR2L$ signals are controlled with pass gate switches in read mode and transmission gates in write mode. Table 4.3 summarizes the specifications of 64 kb SRAM array with C13 configuration

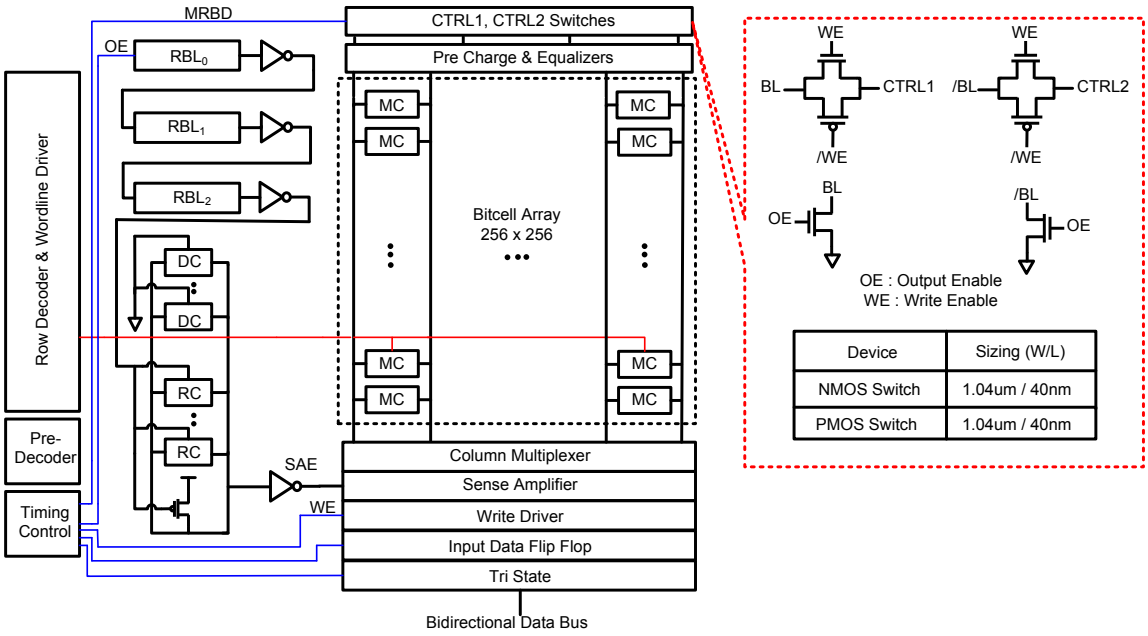


Figure 4.12: SRAM array structure with multi replica bitline delay technique [3].

Table 4.3: Comparison of 64 kb SRAM array with C13 and C1 configurations in 32 nm.

bitcell Config.	**C1**	**C13**
Technology node	32 nm	32 nm
SRAM Organization	256×256 bits	256×256 bits
Supply Voltage	0.9 V	0.9 V
Read-Power cons.	19.32 mW	4.80 mW
Write-Power cons.	18.9 mW	5.19 mW
Leakage power	23.7 mW	5.11 mW
Read access time	44.3 ps	91.6 ps
Write Delay	112 ps	212 ps
No. of Transistors	533, 758	533, 758

and compares it with the same size array with all regular-V_{TH} devices in 8T bitcell (C1). As shown in Table 4.3, using a dual-V_{TH} 8T bitcell demonstrates an excellent trade-off between power, speed and stability and helps to reduce the read and write power by almost 75% with an 50% increase in read and write access time, adding no area overhead or design complexity.

4.8 Summary and Conclusions

In this Chapter, a differential-ended single-port 8T bitcell is proposed which is tolerant to process variation and also achieves a faster access time by increasing the bitcell current by 21% compare to 6T bitcell.

A 30% area penalty is incurred with the addition of two extra NMOS transistors, but the 8T SRAM bitcell can allow for continued scaling beyond what is possible with the 6T bitcell. Besides, its circuit structure does not require any changes compared to a 6T SRAM memory architecture. Also various dual-V_{TH} configurations of differential-ended single-port 8T SRAM bitcell are examined considering process variations using MC simulations. Under process variations each configuration is evaluated based on its stability, read delay and leakage-power in 32 nm SOI technology. Using high and regular threshold voltage devices available in IBM/Global Foundries cmos32soi 32 nm technology, the optimal device combination for power and access-time minimization and data stability maximization without any extra circuit techniques and area penalty is defined.

Chapter 5

MRBD and RRBD Techniques for Optimum Sense Amplifier Set Time

5.1 Introduction

As SRAM continues to occupy most of the area in VLSI systems, the speed and power consumption significantly impacts the system performance [72]. In recent years, power dissipation has become an important consideration due to increasing integration and operating speed of devices, as well as the growth of battery operated appliances. Consequently, the demand for fast memories with lower power consumption continues to be an important consideration for future architectures. However, as process technology scales below 100-nm feature sizes for functional and high yields in silicon the traditional design approach needs to be modified to survive increasing amounts of variation [73, 74].

To decrease the energy consumption for portable applications, circuit designers have been continually decreasing supply voltages and SRAMs are no exception to this trend. Unfortunately, however, the V_{TH} has not scaled down as fast as the supply voltages. Moreover, fluctuations in the V_{TH} cause delay variability of low power circuits across process corners [75, 76]. In case of the SRAMs, the large delay across process corners will demand larger time margins to discharge the bitline path and also will result in larger power dissipation and loss of speed.

The read operation in SRAM is the most time consuming access procedure. Generally, the Sense Amplifier (SA) amplifies the small voltage difference on the bitlines at the proper sense timing to realize

high-speed operations. Therefore, the Sense Amplifier Enable (SAE) signal is extremely important for high speed and low power SRAMs. Unfortunately, with the increased variation effects, such as random dopant fluctuation, accurate generation of timing signals in SRAM are not easy, because the optimum timing for the SAE is sensitive to PVT variations. Fortunately, the timing generation circuit for SAE in SRAM also undergoes similar variation as read path in SRAM array which can be modeled by a normal distribution [77, 78, 79].

If the SAE arrives early before the bitline difference reaches the SA input offset, a read failure may occur and a late-arrived SAE would consume more unnecessary time, thereby wasting the power. Figure 5.1 depicts the distribution of the T_{BL} (time that bitline voltage is sufficient for sensing) and the T_{SAE} (time that sense amp activates) considering process variation. Figure 5.1(a) shows the correct sensing when $T_{BL} < T_{SAE}$ and Figure 5.1(b) shows the wrong sensing when $T_{BL} > T_{SAE}$. As the technology scales down, these distributions become wider and the probability of wrong sensing could potentially change. Therefore, it is necessary to consider a timing margin in SRAM design as shown in Figure 5.1(c) and by increasing the timing margin it is possible to guarantee safe sensing operation. The problem is how to determine the adequate timing margin in order to keep the performance high [80, 81].

The conventional way of generating SAE is to use a RBL that consists of an additional column of dummy cells (DC) and replica cells (RC) that track the random process variation in SRAM array [43]. However, the increased local variations in scaled technologies, causes the replica column characteristics to vary significantly. Consequently, the random V_{TH} variation cannot be tracked well by the RBL technique, which causes read failures and the cycle time deterioration. To suppress variation of the RBL delay, several technique has been proposed [82, 83, 84, 85, 86]. However, there are limitations for all these techniques and they cannot track variations well.

In design of SAE tracking circuitry, following effects should be considered:

1. Threshold voltage variations due to dopant fluctuation in transistor channel results in different bitcell read current and different discharge rates for bitlines.

2. Different input offset voltage due to process variation or aging degradation makes the SA and memory cell asymmetric and results in incorrect sensing. Accurate and symmetric layout helps to minimized the SA input offset voltage but still local threshold voltage variation results in mismatch. This mismatch can make the SA metastable and cause a slow sensing operation when the

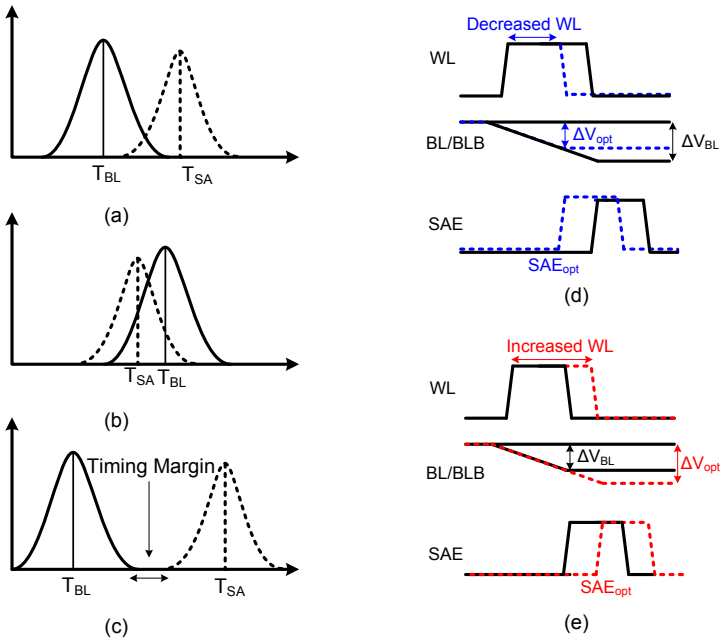


Figure 5.1: (a) A correct read operation when $T_{SA} > T_{BL}$, (b) an incorrect read operation when $T_{SA} < T_{BL}$, (C) timing margin between T_{BL} and T_{SA} due to random process variation effect, (d) a decreased SAE and wordline pulse width to improve the access time and (e) an increased SAE and wordline pulse width to reduce the failure rate.

developed swing on the bitlines is equal to threshold voltage of SA input transistor.

3. Leakage current of unselected SRAM bitcells on the active bitline is dependent on stored data pattern. The worst-case leakage occurs when all unselected cells have the opposite data value from selected cell. Leakage current and data pattern change the set time of SA.

4. When stress is applied on the device continuously, it results in aging degradation. The main reason of device aging degradation is Bias Temperature Instability (BTI) which leads to a shift in threshold voltage for devices that are at strong inversion for a long time [87]. Negative BTI happens for PMOS when $V_D = V_S = V_{DD}$ and $V_G = 0$ and positive BTI occurs for NMOS when $V_D = V_S = 0$ and $V_G = V_{DD}$ ($V_{DS} \approx 0$, $V_{GS} \neq 0$). BTI can be applied asymmetrically to a SRAM cell that stores a fixed value for a long time and results in weakening one side of a cell. This effect makes the SRAM cell asymmetric and weak in reading a specific value which results in higher read failure probability.

5. Temperature, voltage, resistance and capacitance variations increase the failure probability as

well.

By considering the above-mentioned effects, in case of using the conventional RBL, the time margin between T_{SA} and T_{BL} should be increased for proper functionality in scaled technologies. Unfortunately, increasing the time margin by considering the worst case scenario to reduce the SRAM failure rate, deteriorates the performance and increases the SRAM power consumption. Therefore, it is necessary to design new timing control circuitries that allow SAE timing calibration after silicon fabrication to reduce the access time of SRAM based on the characteristics of specific chip. Post silicon calibration allows the ability to dynamically improve SRAM yield and power consumption and recalibrate the SRAM when parameters changes from their nominal values by aging degradation.

Figure 5.1(d) shows how a variable timing controller can decrease the SAE delay to not only resolve the data correctly but also reduce the wordline pulse width in order to reduce the swing on bitlines and save energy. Figure 5.1(e) shows how an incorrect read operation due to insufficient swing can be avoided by delaying the SAE, increasing the wordline pulse width and developing more swing on bitlines that can be sensed by the SA to generate the correct digital output.

This Chapter presents two SAE tracking architectures to suppress the effect of variation on SAE signal. The first proposed architecture utilizes a Multi Replica Bitline Delay (MRBD) technique for the SA read control timing. This technique is more efficient in area compare to other schemes that use replica bitline, such as [83] and [86], because this technique uses less number of replica bitlines to suppress the variation. The simulation results using IBM/Global Foundries cmos32soi 32 nm technology show that MRBD reduces the timing variation approximately 50% using a 0.9 V supply voltage compared to a conventional RBL.

The second technique is a Reconfigurable Replica Bitline Delay (RRBD) to find the optimum timing with minimum deviation for the set time of SA. This technique is suitable for SRAMs that support a wide range of different applications in different operating voltages. The RRBD technique can be used more safely compared to conventional RBL technique and achieves the best time tracking for SAE under PVT variations. This technique not only finds the optimum SAE signal to improve the power, performance and yield, but it also allows calibration after fabrication and recalibration due to device aging degradation. Due to constant stress on silicon devices, SRAM cells become weak over time and cannot perform at their designed performance. However, the RRBD technique generates faster and

slower SAE based on specific characteristics of chip and environmental parameters through a digital control code. The RRBD technique allows calibration for PVT and time-dependent variations and brings the SAE signal to its optimum value, when needed. Therefore with the proposed scheme there is no need to design the SRAM for its worst-case scenario.

5.2 Background

The conventional RBL delay [43] used to be the most common technique to generate accurate SA timing signal in SRAM. The RBL matches the delay of voltage swing at the bitlines and the delay of the SA signal activation. In this design shown in Figure 5.2, both data path (i.e., solid red line) and the SAE control path (i.e., dashed blue line) are driven by SRAM bitcells. Consequently, the effect of global PVT variations is same for both paths. Therefore, the RBL technique can track the optimal SAE timing in presents of variations.

The SAE signal is generated as follows. At first, bitlines in a bitcell array and the RBL column are charged to V_{DD}. Thus, during read mode, selected memory cells discharge the bitlines based on their stored value and develop a differential swing that can be sensed and amplified by the SA. At the same time, The RBL is discharged by a replica cell (a memory cell that is hard-wired to always store zero) to generate the SAE signal. Other memory cells on the RBL column are used as dummy cells to mimic the same parasitic loads of the main bitline. The height of RBL is a fraction of main bitline (normally 10% at nominal voltage) therefore, the RBL is discharged faster than main bitline and generates its SAE signal when bitlines are discharged by a small amount. The differential swing that is developed on the main bitlines must be bigger than SA input offset voltage to be sensed and amplified. And, the size of the RBL column and number of replica cells (RC) determine the activation time of the SA and is determined through circuit simulations at design time. While the RBL is able to track the global variation, it cannot handle the effect of local variations between cells on the bitcell array and replica bitline and may result in a non-optimum setup time for the SA.

The Configurable Replica Bitline (CRBL) technique is presented to decrease the RBL delay variation [82]. A CRBL chooses the RCs with smaller variation to discharge the RBL while these bitcells are chosen based on a post-silicon test. Although, the CRBL technique can decrease the delay variation, it increases the costs especially in large size SRAMs, by adding excessive tests. In [85] multiple replica

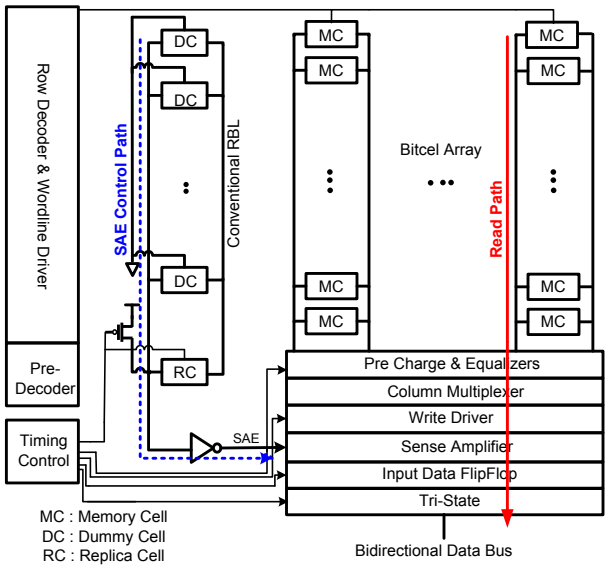


Figure 5.2: SRAM array with conventional replica bitline delay technique (solid red line : read path, dashed blue line: sense amplifier enable path).

bitcells (K) are used in parallel to reduce the effect of V_{TH} variation on the SAE timing. In this technique, the SAE standard deviation is divided by $K \cdot \sqrt{K}$ [85]. However, there is an upper limit for the number of RCs for low supply voltages and this technique cannot suppress the variations for low V_{DD} voltages.

The goal of using a multi-stage replica bitline technique is to suppress the SAE timing by increasing the number of stages (RBLs) [83]. However, inserted inverters between stages increases the delay as number of stages grows which causes a large mismatch between the normal bitline and RBL timing. To keep the total delay constant, the RBL delay variation cannot be dramatically decreased using this technique. To reduce the variation of the RBL delay, a technique called digitized replica bitline delay [84] increases the number of replica cells (K) and then uses a timing multiplier circuit to obtain the final timing for SAE. While increasing K decreases timing variation, quantization noise of the timing multiplier circuit increases variation, because the number of gates in this circuit is proportional to K as well.

The above mentioned schemes and also MRBD technique which is proposed in section 5.3 reduce the effect of process variations in SAE by averaging the signal. Although these techniques are more effective to generate less sensitive SAE signals and improve SRAM access times compared to conventional RBL, they are still all fixed at design time and cannot be adjusted later (e.g., post-production

silicon). And, most importantly these designs cannot be easily ported from one operating voltage to another. On the other hand, statistical methods to generate SA timing, such as methodologies in [77], [88] and [80], are based on the worst-case delay generation that results in extra power consumption and performance degradation.

In [89] a built-in self-test timing-tracking scheme is presented to automatically define the control code for optimum variable delay element. In this technique the output of RBL is connected to an inverter chain and several transmission gates. Transmission gates are controlled by predefined control codes and add delay to the SAE signal, when needed. This technique only generates SAE timings that have equal or bigger delay compare to conventional RBL circuitry and does not produce faster SAE signal with smaller delays. Besides, by using transmission and inverter gates, variation on the SAE signal increase which eventually results in larger access times. In another technique, a programable delay element [81] provides a variable delay for the SAE signal. Again, this technique does not generate faster SAE signal and increases the timing deviation (σ) of the generated SAE (considering a normal distribution for SAE timing with μ_{SAE} (mean) and σ_{SAE} (deviation)).

In section 5.4, a reconfigurable replica bitline is proposed to find the optimum timing with minimum deviation for the set time of SA. This technique is suitable for SRAMs that support a wide range of different applications in different operating voltages. This technique not only finds the optimum SAE signal to improve the power, performance and yield, but it also allows calibration after fabrication. The proposed RRBD technique generates faster and slower SAE based on specific characteristics of chip and environmental parameters through a digital control code.

5.3 Multi Replica Bitline Delay (MRBD) Technique with 8T bitcell

To suppress the effect of random V_{TH} variation on SAE timing, MRBD technique is proposed here, shown in Figure 5.3 [3]. As shown in this figure, compared to conventional RBL, two discharge paths are realized by connecting BL and /BL lines in replica column and also connecting the ML1, ML4, MR1, MR4 gates to V_{DD} in the proposed 8T replica cell. Also, replica column is divided to M segment and in each segment K replica cells are used.

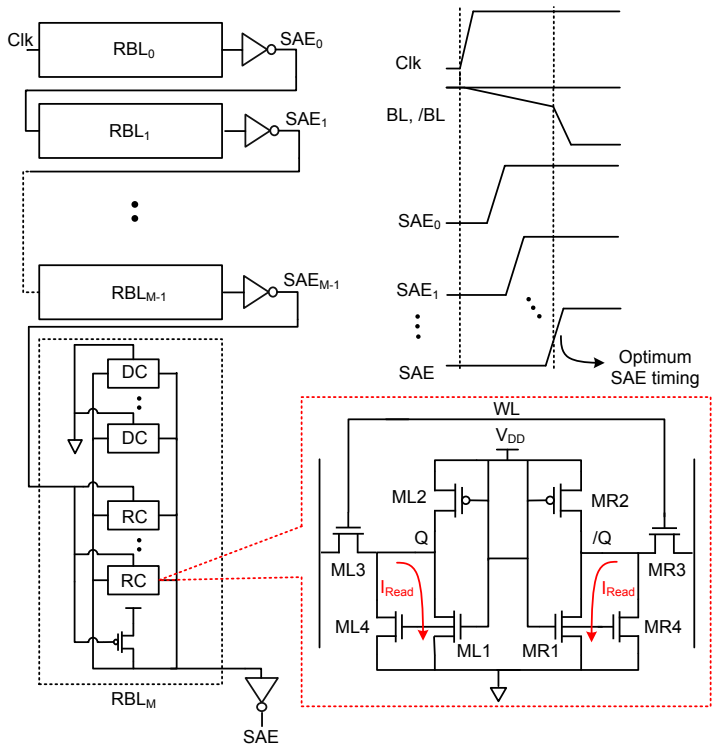


Figure 5.3: MRBD technique with proposed 8T bitcell as replica cell.

The mean and deviation of SAE timing can be expressed by Equation 5.1:

$$\mu_{SAE} + \sigma_{SAE} = \frac{C_{BL} \cdot V_{DD}}{I_{Read} + \Delta I} \tag{5.1}$$

In this equation μ_{SAE} and σ_{SAE} are the mean and deviation of SAE signal, C_{BL} is the bitline capacitance, V_{DD} is the supply voltage, I_{Read} is the bitline discharge current and ΔI shows the effect of variation (random dopand fluctuations in transistor cause threshold voltage variation and eventually current variation). Using the Euler's transformation [90] and knowing $\Delta I \ll I_{Read}$, Equation 5.1 can be written as follows:

$$\begin{aligned} \mu_{SAE} + \sigma_{SAE} &= \frac{C_{BL} \cdot V_{DD}}{I_{Read} \cdot (1 + \frac{\Delta I}{I_{Read}})} \\ &\approx \frac{C_{BL} \cdot V_{DD}}{I_{Read}} \cdot \left(1 - \frac{\Delta I}{I_{Read}}\right) \approx \frac{C_{BL} \cdot V_{DD}}{I_{Read}} - \frac{C_{BL} \cdot V_{DD} \cdot \Delta I}{I_{Read}^2} \end{aligned} \tag{5.2}$$

The deviation of SAE signal in MRBD is suppressed as follows. MRBD uses both bitlines in its

discharge path while conventional RBL uses only one bitline. The structure of replica cell is changed in order to provide two discharge path as shown in the Figure 5.3. This replica cell has the same size transistors as the memory cell, so it adds the same parasitic capacitance and resistance on the RBL as the memory cell adds on the main bitline. And, having two discharge path provides a smaller deviation for the SAE timing with the proposed technique through averaging. Using both bitlines doubles both the capacitance load ($2 \times C_{BL}$) and discharge current ($2 \times I_{Read}$) of each replica column, therefore, based on Equation 5.1 it would not change the mean value of the SAE but divides the deviation by $\sqrt{2}$.

Increasing the number of activated replica cells (K replica cell instead of one) in each replica column leads to more discharge current ($K \cdot I_{Read}$, $\sigma^2 = K \cdot \Delta I^2$ or $\sigma = \sqrt{K} \cdot \Delta I$). According to Equation 5.1 using K replica cells divides the mean value by K and deviation by $K \cdot \sqrt{K}$. Therefore, the MRBD technique generates the SAE signal with less delay and smaller variation by increasing the value by K.

Besides, each replica is column divided to M segments, thus, with the same number of RCs, the capacitance load of each segment will be C_{BL}/M, in compared with a conventional RBL design. This will divide the deviation of the SAE signal by $\sqrt{M}$ and the delay of each segment would be $1/M$ of conventional RBL delay. By placing an inverter between output of one segment and input of another one, all segments are connected together to form the MRBD, as shown in Figure 5.3. The delay of M segments then are added together, so the final delay of SAE would be the same as RBL while the σ value is divided by $\sqrt{M}$.

Having inverters between each segment increases the variation of SAE signal and also shifts the mean value of SAE to the right. However, using k RC in each segment, helps to bring the SAE mean value to its proper value. Therefore, the optimum mean value for SAE can be reached while its σ value is divided by $\sqrt{M}$ and $K \cdot \sqrt{K}$ due to having M segments and K replica cell in each one. Also, σ is divided by $\sqrt{2}$ because in discharging of each segment, both BL and /BL are used. As a result, σ value of SAE signal is divided by $K \cdot \sqrt{(2 \cdot M)}$ which means compared to conventional RBL scheme, variation of SAE is considerably reduced. The probability distribution of the SAE timing for the conventional RBL compared to the MRBD techniques is shown in Figure 5.4. By adjusting the values of M and K, the MRBD technique is able to achieve lower variation for SAE timing.

Figure 5.5 shows the block diagram of the 256×256 bits designed SRAM. Each word is 64 bit and word selection is performed using a decoder/ multiplexer combination. To control the voltages of $CTRL1$ and $CTRL2$ in read and hold modes a single pass transistor is used while a transmission gate

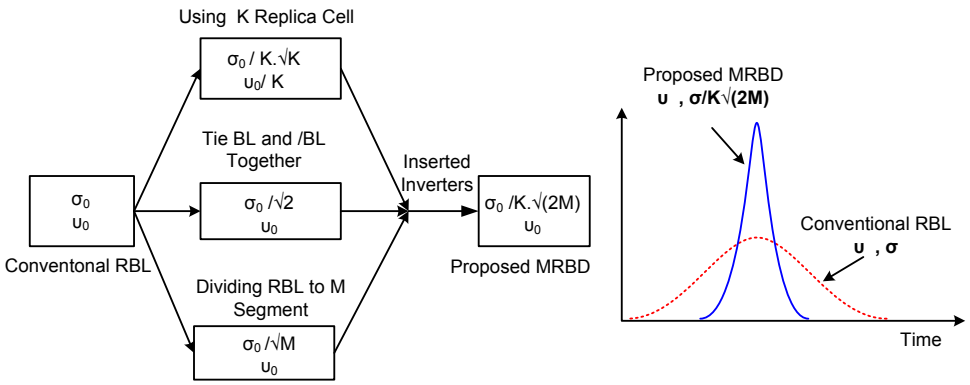


Figure 5.4: Comparison of conventional RBL and MRBD timing variations.

with complementary control signals is used in write operation to transmit high and low signals. A write driver is skewed to increase the speed of the write operation and reduce the delay of driving the bitline with the input data. In read operation, a small difference voltage swing on the bitline is sufficient to detect the stored value. To apply the MRBD technique, a latch-based SA is chosen and SAE signal will be generated when minimum readable differential swing is available on bitlines. This minimum readable value is limited by the offset of SA input transistors. The minimal target value of the required swing on the BL depends on the SA design, its sizing and the technology node. An increase in random V_{TH} variations increases the offset voltage mismatch for the input transistors of the SA and requires the use of upsized SAs for a correct read operation. By increasing the size of the input transistors in the SA, an offset voltage can be reduced, but this up-sizing directly increases the energy consumption.

In Figure 5.6, a $1,000$ point transient MC simulation of the SAE and BL discharge signals during a read operation at 0.9 V supply voltage in IBM/Global Foundries cmos32soi 32nm technology are shown. Figure 5.6(a) shows the results for a conventional RBL using a 6T bitcell and Figure 5.6(b) shows the waveforms for the MRBD with the proposed 8T bitcell. MC transistor models for cmos32soi technology are supported by manufacturing vender. In transient MC simulations, 3σ random dopant variation and device geometric mismatch are included to model the most complete representation of statistical variation during chip manufacturing. The best choice determined by simulation for the number of stages and RCs using the MRBD technique for this technology are 4 and 4, respectively. The standard deviation of the SAE signal is 30 ps with the MRBD scheme, which is 50% less compared to conventional RBL scheme, as shown in Figure 5.6(a) and (b). Figure 5.6(c) and (d) show the SAE and

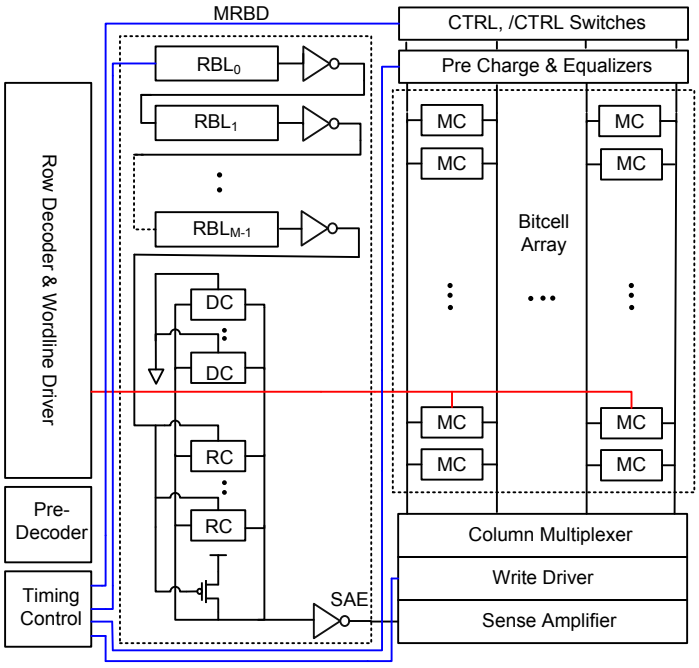


Figure 5.5: Block diagram of SRAM circuit using the MRBD technique.

the BL discharge signals results in FF (Fast-NMOS, Fast-PMOS) and SS (Slow-NMOS, Slow-PMOS) process corners using the MRBD technique and the proposed 8T bitcell. As shown in this figure, the MRBD technique provides the SAE with 11 ps deviation for FF corner and 52 ps deviation for SS process corner, which are the best and worst process corners, respectively. These results show SRAM with MRBD and proposed 8T bitcells leads to less access time even at worst process corner compared to a 6T SRAM array with the RBL technique.

Table 5.1 summarize the performance of the MRBD design using the proposed 8T and compares it with conventional RBL which uses a 6T bitcell. The MRBD technique has more transistors and a larger power consumption, however, this power and area cost are acceptable due to significant reduction of SAE timing variation and improvement in the SNM, WNM and readout bitcell current and more importantly great improvement in operation frequency. To achieve the same operation frequency with 6T and conventional RBL a higher supply voltage is needed which can lead to more power consumption, since power is proportional to square of the voltage. Thus, the MRBD saves overall power consumption in SRAM. The MRBD technique with the proposed 8T bitcell can be applied to the different row and column configurations, so it can be used in traditional memory compilers such as OpenRAM [64].

Figure 5.7(a) shows the read delay of proposed 8T and 6T SRAMs at different supply voltages.

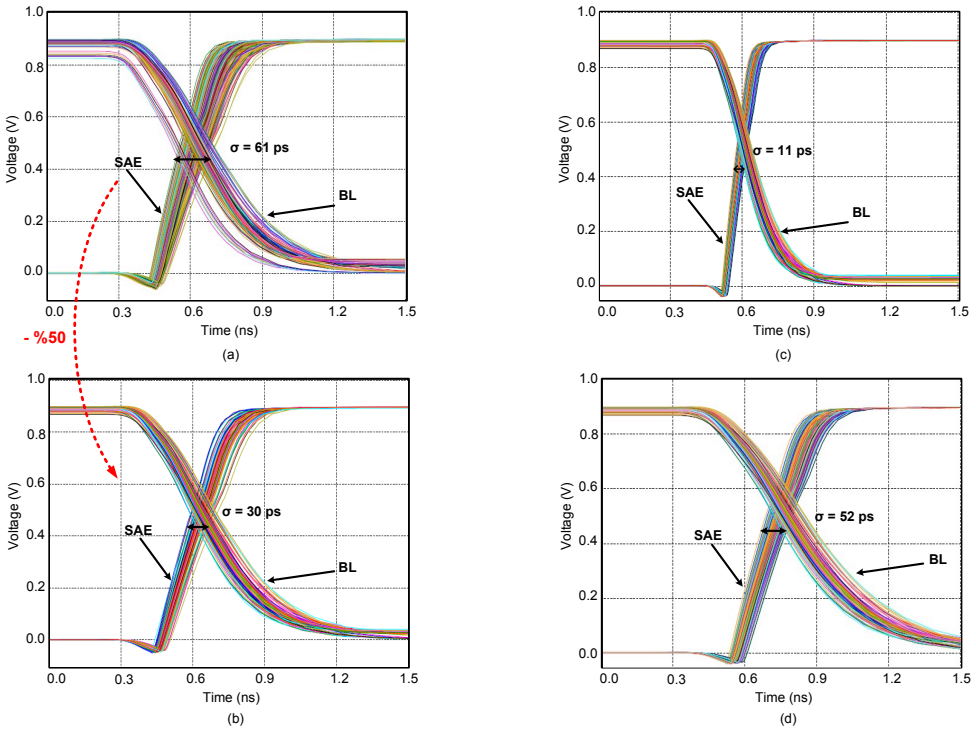


Figure 5.6: MC simulation results of SAE and BL timing variation of (a) conventional RBL with 6T bitcell @ TT corner, (b) MRBD with proposed 8T bitcell @ TT corner, (c) MRBD with proposed 8T bitcell @ FF corner and (d) MRBD with proposed 8T bitcell @ SS corner (256×256 SRAM array, 0.9 V).

Table 5.1: Summary of MRBD/8T and conventional RBL/6T design comparison in IBM/Global Foundries cmos32soi 32nm technology.

Design	**Conventional RBL with 6T**	**MRBD with proposed 8T**
SRAM Organization	256 rows 256 cols	256 rows 256 cols
Operating Frequency	250 MHz	500 MHz
Power Dissipation	3.100 mW	5.851 mW
Supply Voltage	0.5 V	0.5 V
Bitcell SNM	68 mV	115 mV
Bitcell WNM	118 mV	231 mV
No. of Transistors (Area)	402, 939 (1 x)	531, 020 ($\approx$ 1.3 x)

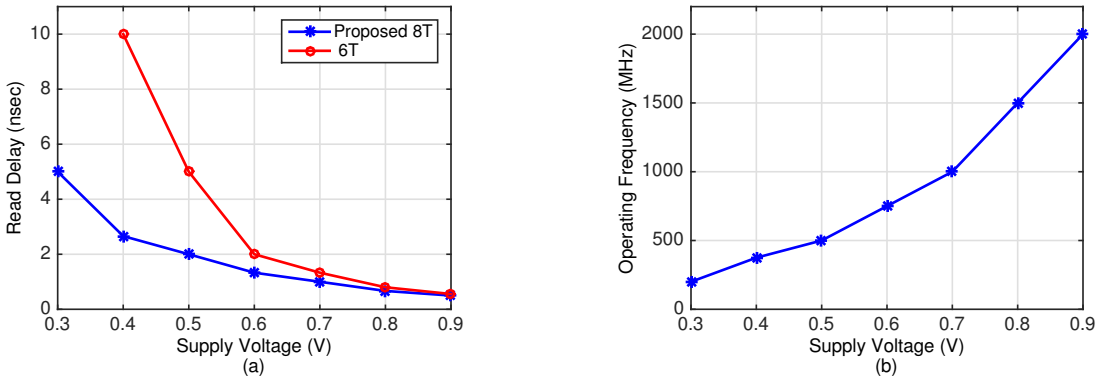


Figure 5.7: (a) Read delay of proposed 8T and 6T bitcells in 256×256 SRAM array and (b) operation frequency of 8T SRAM array at different supply voltages.

Here, the read delay (access-time) is measured at the time interval from 50% of a low-to-high transition of a word-line signal until there is a $V_{DD}/2$ differential swing on the bitlines. As it is shown, the read delay of proposed 8T is less than the 6T, since the proposed 8T has one more transistor in its discharge path and it has 21% more bitcell current. Also, Figure 5.7(a) shows that using proposed 8T bitcell, it is possible to have a faster SRAM at low voltages. As shown in this figure conventional 6T cannot operate at low voltages (0.3 V) because it does not have enough stability while proposed 8T works with small access time. Figure 5.7(b) shows the maximum operating frequency of the proposed 8T design versus different supply voltages. This figure shows that while the array can perform at 2 GHz with 0.9 V, it is also able to function at low voltages without the need for secondary or dynamic power supplies. A high frequency operation of 200 MHz at 0.3V is enabled by the incorporation of proposed 8T bitcell. The minimum V_{DD} of the SRAM macros is limited by noise margin as shown in Figure 4.7.

5.4 Reconfigurable Replica Bitline Delay (RRBD) Technique

Figure 5.8 shows the proposed reconfigurable replica bitline as the control circuitry of SAE. This design generates variable delay based on input control code Aji and Bj (i is the number of replica cell in each column and j is the number of RBL column).

If local variation in memory cells on the main bitline or RBL causes an early-generated SAE signal when sufficient swing is not developed on the bitlines, it is possible to increase the SAE delay using RRBD technique. This extra delay allows SRAM to develop bigger differential swings to overcome the

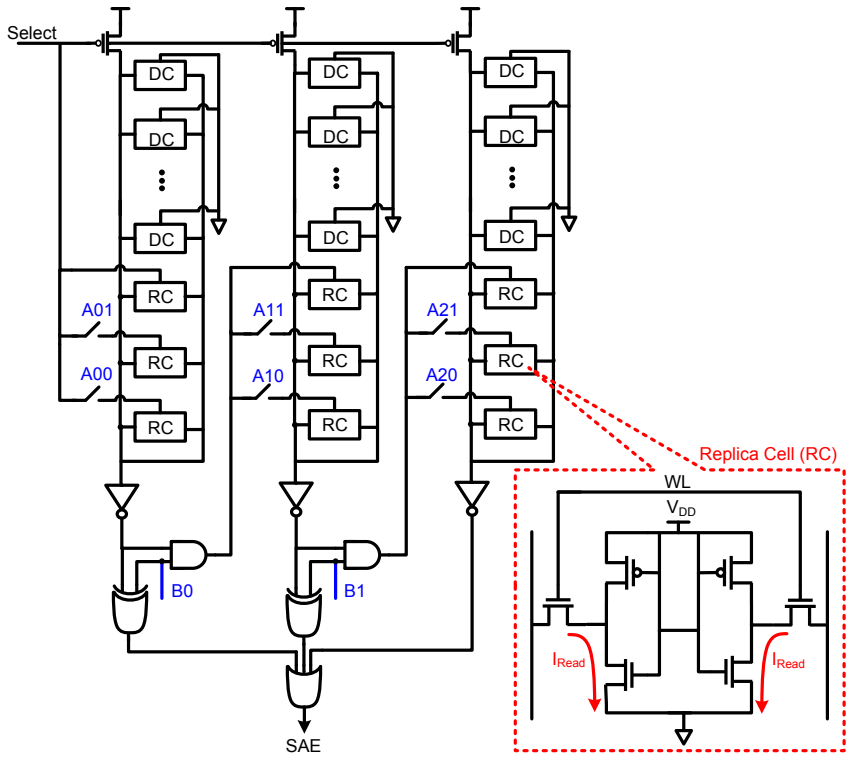


Figure 5.8: Proposed reconfigurable replica bitline delay (RRBD) scheme (A_{ji} and B_j are control code bits).

input offset voltage of SA. Conversely, if the SAE signal enables late, the RRBD technique can decrease the SAE delay to avoid extra discharge on bitline and eliminate the excessive power dissipation and access time.

Figure 5.8 shows how the control code changes the SAE delay. When all bit of digital code are zero (Aji =0 and Bj = 0), RRBD acts like MRBD technique. In cases that less SAE delay is needed, control bits in the first replica column (A0i) are set to high one by one to increase the number of activated replica cells (increase the discharge current) and reduce the SAE delay.

In case a slower SAE compared to what a conventional RBL generates is needed, control bit Bj, goes high one by one and adds extra delay to the SAE signal. It is possible to tune the SAE delay by both coarse tuning bits (Bj) and fine tuning bits (Aji) to get the optimum value for SAE. The mean value of the SAE signal changes when RRBD adds extra delay (e.g., when B0B1= 10, μ_{SAE} equals $2 \cdot C_{BL} \cdot V_{DD}/I$). Therefore, having two RBL column multiplies the mean and deviation by 2 and $\sqrt{2}$, respectively, based on Equation 5.1. However, because the deviation on each RBL is divided by $1/\sqrt{2}$ by using both bitlines, there would be no increase in deviation and the proposed design results in the

same amount of deviation as conventional RBL. It is worth mentioning that techniques in [89] and [81] increase the deviation by using inverter gate delays and eventually result in more unnecessarily access time. In addition, using fine tuning bits in the second RBL column allows the ability to determine the optimum SAE **and** also reduce the deviation. RRBD technique can be extended to more RBL columns and replica cells in each column for better tuning in smaller supply voltage, near threshold and subthreshold regions, where conventional RBL and techniques in [89] and [81] are not effective.

The flowchart in Figure 5.9 shows how the control code for RRBD technique can be defined based on specific parameters, process and environment variations. The height of each RBL column, number of replica cells in each column and number of RBL columns can be defined by simulation at typical, fast and slow process corners, respectively. And the digital control code can be defined based on desired yield for SRAM through few iterations as shows in Figure 5.9.

RRBD technique allows the optimum SAE to be found for each chip based on its specific characters. The proposed technique not only allows the ability to dynamically manage the power-yield trade-off based on an application, but also provides a wide supply voltage for a SRAM array. By using RRBD technique the setup time for the SA can be increased or reduced for lower or higher supply voltages compared to a nominal one. This feature allows the SRAM array to be used in different applications and a wide-voltage ranges.

The RRBD technique is simulated with a 64 kb (256×256 bits) SRAM array and is compared with a conventional RBL at different supply voltages, temperatures and input offset voltages of SA. The number of RBL columns and replica cells are 3 and 3 based on simulations results in slow and fast process corners and the height of RBL is 26 cells that is defined by simulation in typical corner. To define the number of RCs, RBLs and RBL height (SAE delay), nominal supply voltage for 32 nm (0.9 V) and a 99.9% yield (0.1% read failure) are considered in all simulations. To evaluate the effectiveness of proposed design thousands of MC simulation in IBM/Global Foundries cmos32soi 32 nm technology are done. MC transistor models for this technology are supported by the manufacturing vendor. In transient MC simulations, 3σ random dopant variation and device geometric mismatch are included to model the most complete representation of statistical variation during chip manufacturing. Before a read/write operation, the bitlines are precharged to V_{DD} in the first half of the cycle and read/write is done in the second half.

Table 5.2 shows how the mean and deviation of SAE signal changes based on the proposed digital

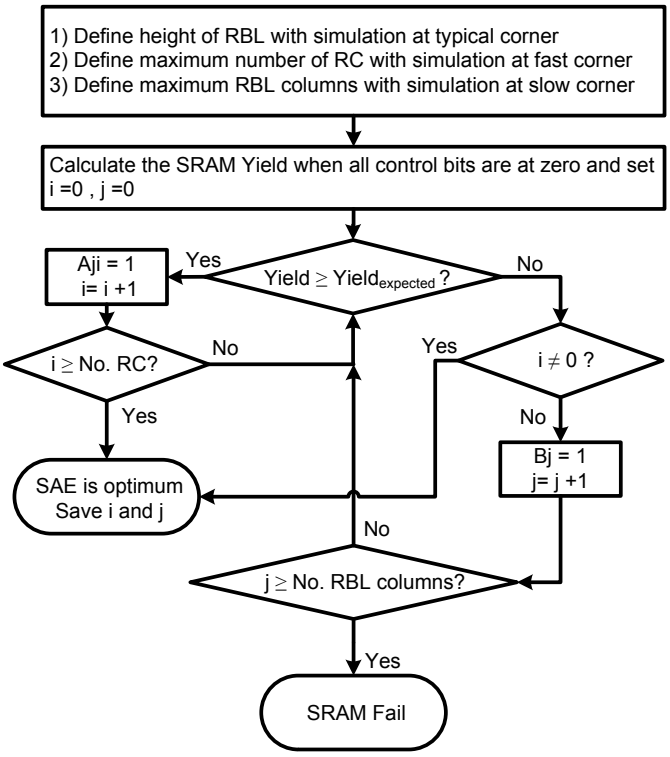


Figure 5.9: Optimum SAE timing generation Flow.

Table 5.2: $1,000$ point MC Simulation for $\Delta\mu_{SAE}$ & $\Delta\sigma_{SAE}$ based on digital control code (Δ : RRBD - RBL).

Control Code (A_{ji} B_j)	$\Delta\mu_{SAE}$ **[pSec]**	$\Delta\sigma_{SAE}$ **[pSec]**	**Variation** (6σ) **improvement**
000000 00	0	0.00	0.000%
100000 00	-41	-7.00	35.554%
110000 00	-60	-10.00	48.500%
000000 10	81	-0.90	4.785%
001000 10	55	-5.20	23.745%
001100 10	39	-7.10	35.605%
000000 11	140	0.75	-3.525%
000010 11	107	-0.50	2.425%
000011 11	92	-5.20	23.745%

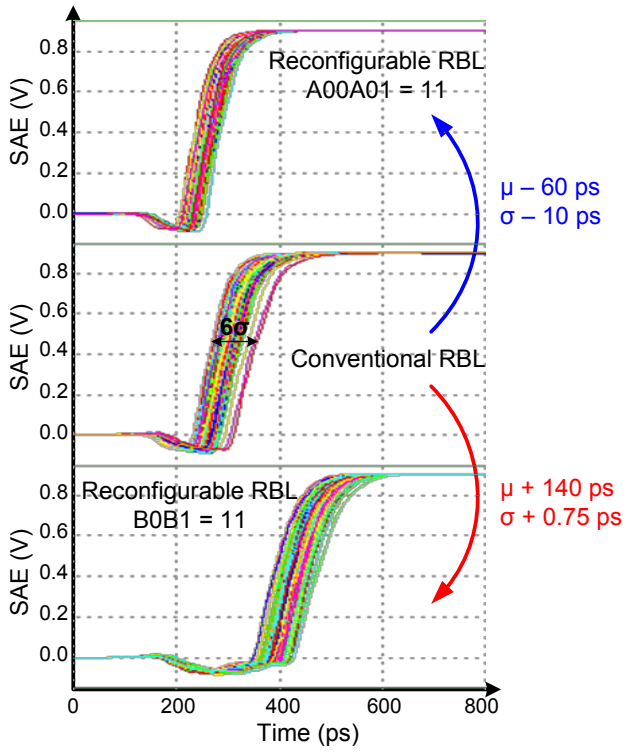


Figure 5.10: Decrease in μ_{SAE} and σ_{SAE} by setting A_{00} and A_{01} bits and increase in μ_{SAE} with no degradation in deviation by setting B_0 and B_1 bits of control code in RRBD.

control code. When all control bits are at zero, the RRBD technique acts like a conventional RBL and no extra delay is added or subtracted from the SAE signal. When B_{0-1} are at zero and A_{00-01} goes high, the RRBD generates SAE signals with less delay and deviation compared to conventional RBL. As shown in Table 5.2 by setting A_{00} and A_{01} to 1, the SAE delay decreases by 41 and 60 ps, and deviation decreases by 7 and 10 ps, respectively. By activating B_0 and B_1 switches, the mean value of the SAE signal increases while its deviation decreases or stays almost the same as a conventional design. The generated SAE signal for different control bits of the RRBD technique and comparison with conventional RBL are shown in Figure 5.10.

Figure 5.11 shows how operation frequency of SRAM changes based on the fastest and slowest SAE timing in proposed design, compared to conventional RBL. As shown in this figure operation frequency of an SRAM can increase by 20% using reconfigurable RBL compared to an conventional RBL that is designed for worst case scenario.

A maximum of 0.15 V is considered as input offset voltage of the SA, because of its small layout size (i.e., SA layout must be pitch match with 6T layout). Figure 5.12(a) shows how the TSAE increase with

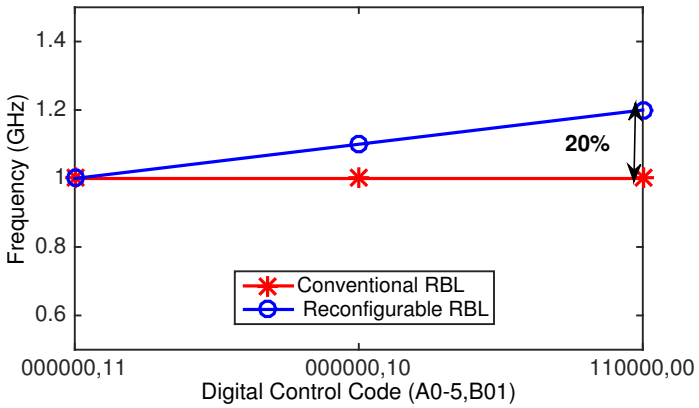


Figure 5.11: Operation frequency of a 64 kb SRAM array at 0.9 V voltage based on digital control code (operating frequency of conventional RBL is fixed at design time for worst case).

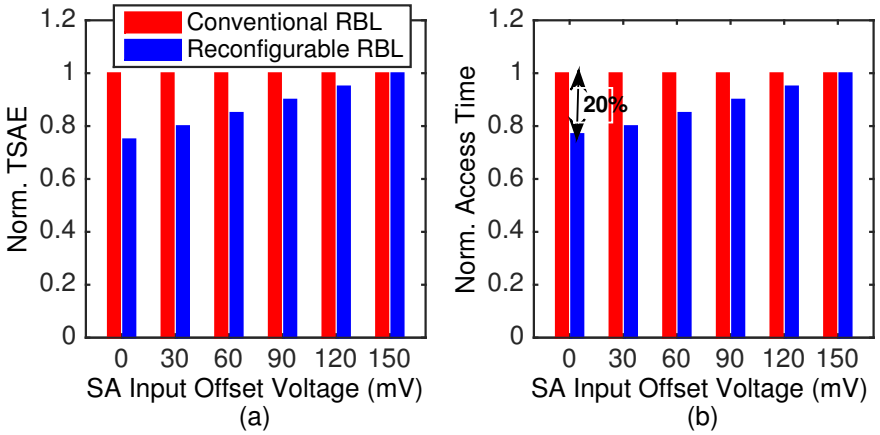


Figure 5.12: (a) SAE enable time and (b) access-time comparison when conventional RBL is designed with 150 mV input offset voltage for SA.

an increase in input offset voltage of the SA using reconfigurable RBL while it is fixed in conventional RBL design because the conventional RBL is designed for the worst case input offset voltage (0.15 V). Figure 5.12(b) shows how a design for the worst input offset voltage results in 20% bigger access time for conventional RBL. On the other hand, Figure 5.13(a) shows the case when the conventional RBL is designed considering zero input offset voltage for SA. Figure 5.13(b) shows how the yield decreases with an increase in the SA offset voltage for a conventional RBL and the RRBD is able to tune the delay of SAE in order to keep the SRAM yield.

Figure 5.14 shows how reconfigurable RBL generate sufficient SAE delay in different supply voltages. In lower voltages the effect of variation is more and therefore a bigger time margin between T_{BL} and T_{SA} is needed (more access time). This larger time margin allows more swing on the bitlines at the

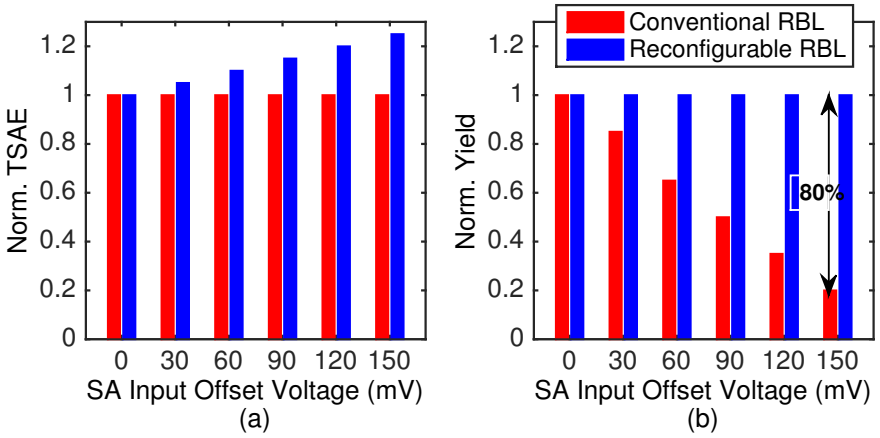


Figure 5.13: (a) SAE enable time and (b) yield comparison when conventional RBL is designed with zero input offset voltage for SA.

time of the SA activation and results in less failure. Based on the transient MC simulation for 64 kb array sufficient swing on bitline for a 99.9% yield at 1.1 and 0.7 V are 100 mV ($\approx$ 10% V_{DD}) and 320 mv ($\approx$ 50% V_{DD}), respectively. As shown in Figure 5.14 RRBD technique can generate the delay to meet this yield, while conventional RBL with fixed number of replica and dummy cells fails at lower voltages (when extra SAE delay is needed).

Figure 5.15 shows how the SAE timing of the proposed design changes with temperature variation. Higher temperature leads to more subthreshold leakage current and it takes more time for SRAM to develop the sufficient differential swing on the bitlines. Therefore, as temperature rises more SAE delay is needed to maintain the yield. Proposed design provides more delay by adjusting the control code while conventional RBL results in extra delay at lower temperature if it is designed for worst case temperature. Again, proposed design results in less access time and less power consumption compared to traditional RBL scheme for temperature variation. Proposed design does not have any dependency on SRAM configuration and can easily be applied to SRAM compilers such as OpenRAM [64]. Table 5.3 summarizes and compares the specifications of SRAM arrays with conventional and proposed RRBD technique. As shown in this table, SRAM array with RRBD can set the performance (clock frequency) based on supply voltage using the digital control code in order to have the minimum access-time/power consumption. On contrary, the performance of the SRAM array with conventional RBL is fixed for a specific supply voltage at design time and any PVT or time dependent variation may result in performance degradation or yield loss for this array.

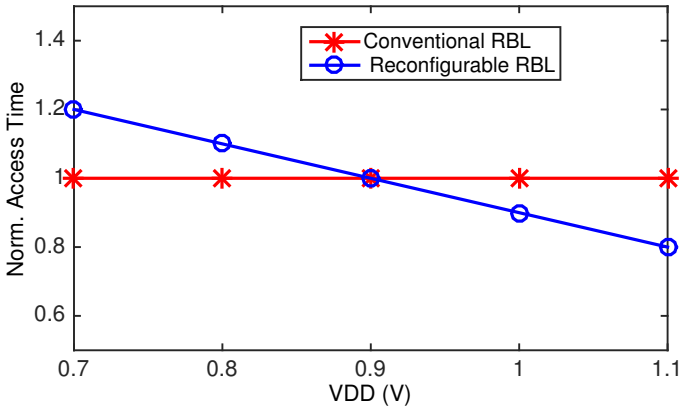


Figure 5.14: Access time in different voltages (variation in supply voltage). Conventional RBL results in extra access time in higher voltages and more read failure in lower voltages while proposed RRBD sets the optimum SAE time.

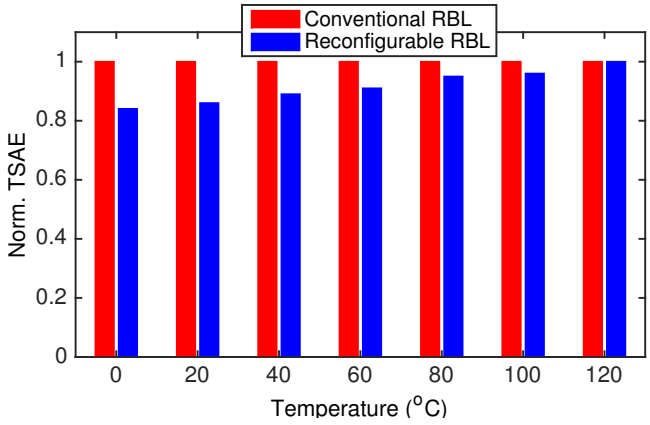


Figure 5.15: RRBD can generate optimum TSAE with temperature variation.

Table 5.3: Characteristics of 64 kb SRAM array with conventional RBL and RRBD at 32nm technology.

	Conventional RBL	**Reconfigurable RBL**
SRAM Macro Organization	256×256 bit	256×256 bit
Power Supply Voltage	0.9 V	0.9 V $\pm 25\%$
Average Power Consumption	4.06 mW	$3.10 - 4.45$ mW
Number of Transistors	$402,939$	$403,327$
Clock Frequency	1 GHz	1 GHz $\pm 20\%$

5.5 Summary and Conclusions

In this Chapter a multi replica bitline delay technique is proposed to improve the random variation tolerance of sense amplifier timing signal in SRAM. MRBD uses differential-ended single-port 8T bitcell as dummy cell and replica cell. Compared with conventional RBL scheme with 6T bitcell, this design shows a 50% variation reduction in SAE timing at 0.9 V supply voltage in an IBM/Global Foundries cmos32soi 32 nm technology while there is a negligible increase in power consumption. Also, final SRAM architecture for proposed 8T bitcell and MC simulation results are shown. The proposed 8T along with the MRBD technique are demonstrated in a 64 kb SRAM array designed in a 32 nm technology that operates at 2 GHz with a 0.9 V and 250 MHz with a 0.3 V supply voltage.

Also, a reconfigurable replica bitline technique to determine the optimum set time of SRAM sense amplifier under PVT and time-dependent variations is presented. RRBD technique sets the sense amplifier activation time to the minimum required value for reliable operation based on desired yield. Proposed design allows SRAM to be used for different application in different voltages. A 64 kb SRAM array with RRBD technique is simulated in IBM/Global Foundries cmos32soi 32 nm technology at 0.9 V and shows 20 % less access time compared to the same size array with conventional RBL technique.

Chapter 6

A Half-Select Disturb-Free Subthreshold 12T SRAM Bitcell

6.1 Introduction

Subthreshold operation enables suppressing dynamic power consumption and extends the battery life time of low power devices. SRAM scaling is one of the major bottlenecks for the reduction of supply voltage in current and future CMOS technology nodes. Although SRAM can achieve low power dissipation in subthreshold region, it must face the ever increasing process variation challenges in this region. With an increase in process variations for lower supply voltages, it is becoming difficult to balance the read and write stability for a 6T SRAM bitcell due to its conflicting design requirement in read stability and writability [54]. Besides, in sub 100 nm technology nodes, subthreshold leakage power is a substantial portion of total power consumption and 6T bitcell doesn't have any mechanism to control this leakage. During read operation subthreshold leakage leads to a read failure for 6T bitcell in small supply voltages. Although, techniques such as multi-threshold bitcells [71] and local-global bitlines [91] help to decrease leakage current of bitcells and long bitlines, these techniques degrades the performance or increase the area overhead.

Traditional 6T bitcell has a simple structure but suffers from write half-select issue. The half-select disturbance occurs when there is a half-selected column in write mode. During this occurrence, the bitcell in the unselected column is disturbed because the wordline is raised to turn on the access

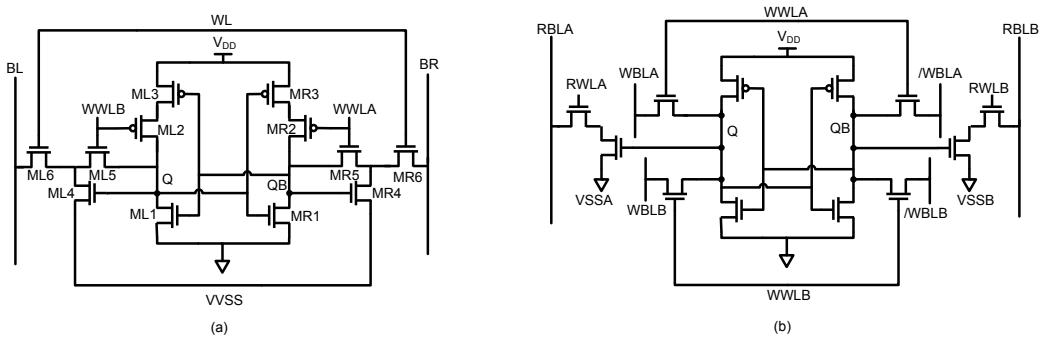


Figure 6.1: (a) Single-port 12T bitcell (12T-S) [4] and (b) quadruple-port 12T bitcell (12T-Q) [5].

transistors for selected bitcells that need to be written [92]. Due to half-select issue, most of the SRAM designs cannot be bit-interleaved and lose data through multi-bit soft error.

Different configurations for SRAM bitcells have been proposed to improve the read stability, writability and subthreshold leakage control in low-voltage operation. 8T [2] eliminates charge sharing between the bitlines and internal storage nodes and improves the SRAM stability in low voltages. However, this bitcell suffers from a reduced swing on bitlines due to leakage as well as poor noise immunity due to its single-ended structure. In addition, an improvement of the access time is not expected since read operation is single-ended and a full rail sensing is necessary. [60] and [93] solve the 8T bitline-leakage problem during read by stacking three MOS transistors in read path. Again although these bitcells improve the bitline leakage in low voltages, they have poor noise immunity due to their single-ended structure and read operation is slow due to full rail sensing. Besides both bitcells still have the half-select issue in write mode. In [14] a differential-ended 10T bitcell makes use of the voltage difference between BL pair during a read operation to make this bitcell a suitable candidate for high-speed applications. This cell uses decoupled read port to improve the read stability and has two wordlines which helps to control the half-select issue and use bit-interleaving structure. However, this bitcell cannot be utilized on long bitline SRAMs because of its poor mechanism to control the leakage in read mode.

12T bitcell in [4] (12T-S) as shown in Figure 6.1(a) employs a cross point write structure with a data aware column based write wordline to eliminate the half-select disturb, therefore, can be used in bit-interleaving structure. However this bitcell cannot be used in long bitlines due to bitline leakage current in read mode which leads to a read failure in small supply voltages. Another 12T bitcell (12T-Q) [5] as shown in Figure 6.1(b) uses two differential ports for write and two single-ended ports for read

operation to be read and written simultaneously. Again single-ended decoupled read port improves read stability while increase access time and reduces noise immunity. This cell doesn't control half-select problem and it is not suitable for bit-interleaving structures. Besides, this bitcell does not have leakage control mechanism and cannot be used in high density SRAMs with long bitlines. In addition, this cell has four wordline and six bitlines which leads to a significant increase in bitcell area compared to 6T bitcell and due to its multiple port structure, it requires extra circuitry to be controlled that potentially leads to an increase in power and as well as a larger area penalty.

To overcome the limitations on SRAM bitcell in low-voltage and low-power operations, a novel 12T bitcell is proposed here with the following features:

1. This bitcell provides greater improvements in the static read and write noise margins by decoupling the bitline from storage node during read and boosting the gate voltage of access transistors during write mode. Hence, it can withstand the ever increasing process variations in scaled technology nodes.

2. The proposed bitcell has a fully differential structure and layout therefore shows better noise and mismatch immunity compared to single-ended schemes.

3. The proposed 12T bitcell utilizes differential sensing for read operation which leads to faster operations and less access time.

4. This cell has a row-based wordline and a column-based control signal, therefore it can eliminate the half-select issue during write by isolating the stored data from bitline.

5. The proposed 12T bitcell can be implemented in a bit-interleaving structure and allows to solve the multi-bit soft errors by conventional error correction code (ECC) techniques.

6. This bitcell has a leakage control ability which helps to reduce the bitline leakages in read and hold modes and provides a fast robust read operation in low voltages. Therefore, it is a suitable candidate for arrays with long bitlines for high density SRAMs.

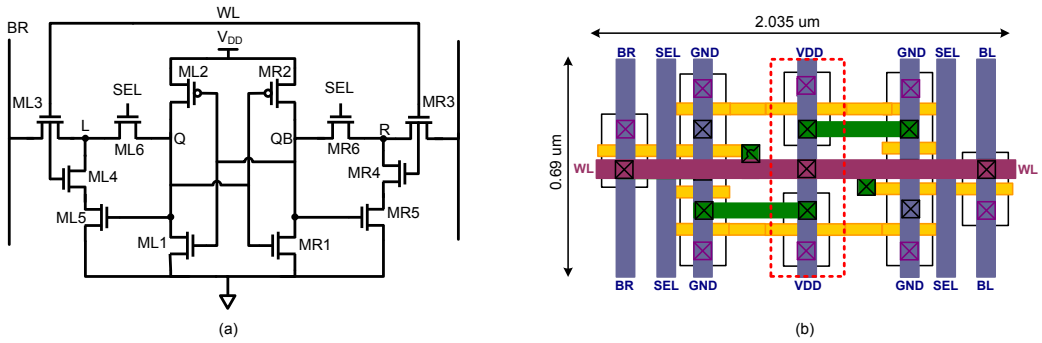


Figure 6.2: (a) Proposed single-port differential-ended 12T bitcell, (b) one possible 1.4 μm^2 layout in 32 nm CMOS SOI technology.

Table 6.1: Transistor sizing for proposed 12T bitcell.

	ML1, MR1	**ML2, MR2**	**ML3, MR3**	**ML4, MR4**	**ML5, MR5**	**ML6, MR6**
W (nm)	104	104	104	104	104	104
L (nm)	40	40	40	40	40	40

6.2 Proposed 12T SRAM Bitcell

Figure 6.2(a) shows the schematic of the proposed 12T bitcell. This bitcell is fully differential hence has a good noise and mismatch immunity. The proposed 12T bitcell consists of a cross-coupled inverter pair (ML1, ML2, MR1, MR2) that keeps the stored data, write access transistors (ML3, ML6, MR3, MR6) and decoupled differential read ports (ML3, ML4, ML5, MR3, MR4, MR5). Wordline signal (WL) is row-based while SEL signal is column-based. As shown in Table 6.1 all the transistors are minimum sized because read and write path does not have conflicting design requirement in this bitcell. Besides, in subthreshold operation, since the ratio of PMOS to NMOS current depends exponentially on threshold voltage, sizing is not a strong knob for improving noise margin in read or write mode. Figure 6.2(b) shows one possible thin cell layout of the proposed 12T bitcell. Although this 12T bitcell adds more area overhead relative to 6T SRAM bitcell, the overall area penalty is less because more bitcells can be included in the bitlines.

6.2.1 Read Operation

Figure 6.3(a) shows the 12T bitcell timing diagram in read, write and hold modes. Figure 6.3(b) shows the 12T bitcell current path during the read mode. The bitlines (BL, BR) are precharged to V_{DD} before

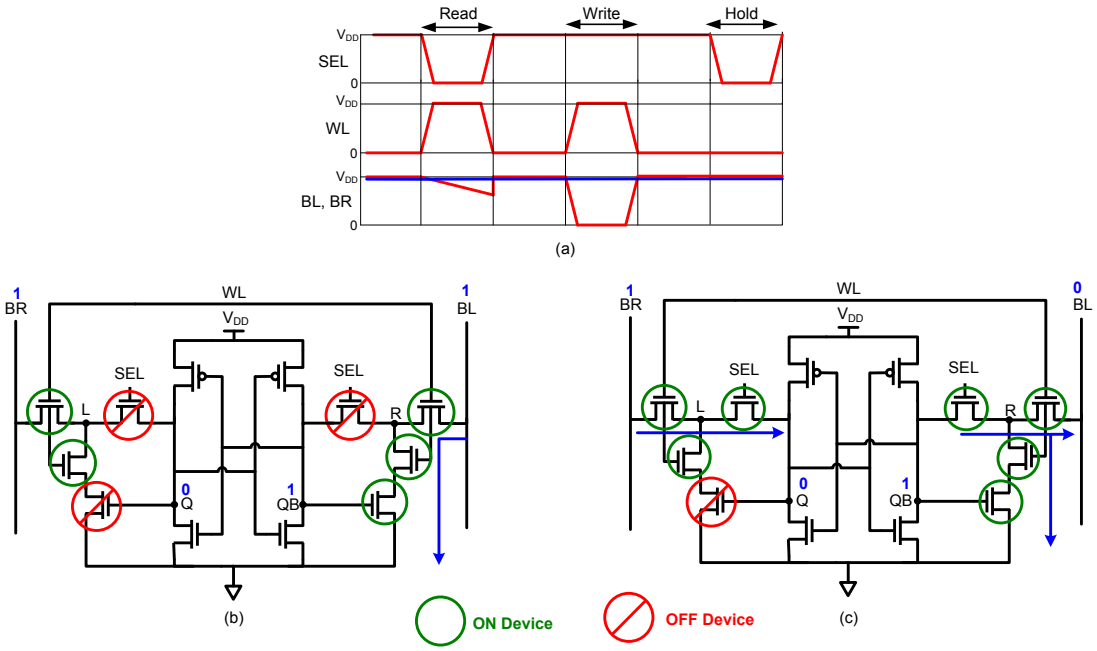


Figure 6.3: (a) Timing diagram, (b) read current path and (c) write current paths of proposed 12T bitcell.

the cell is accessed. When wordline is enabled and SEL remains disabled (WL =1 and SEL =0), BL is discharged through pull-down transistors MR3, MR4 and MR5. In this case Q has the value 0 which leads to a discharge in BL while BR stays high. A latch-type sense amplifier is used to sense the differential swings on BL and BR in order to speed up the read operation. In proposed 12T, the read value is the inverted signal of stored value, hence, position of BL and BR are exchanged in this bitcell. The cell storage node is decoupled from the read bitline, therefore SNM during read is almost equal to Hold Noise Margin (HNM) of conventional 6T bitcell.

The SNM is defined as the maximum possible noise available at the gates of the cross-coupled inverters or storage element that does not flip the bitcell value [11]. The read Voltage Transfer Characteristic (VTC) of 12T bitcell can be measured by sweeping the voltage at storage node Q with both BL and BR and WL biased at V_{DD} while monitoring the node voltage at QB. The SNM can be quantified by the side of the largest square embedded between the read VTC curves. Figure 6.4(a) shows the read VTC curves for the proposed 12T bitcell and compares it with the SNM of traditional 6T bitcell. The 12T bitcell has a SNM of 86 mV at 0.3 V while that of a 6T bitcell is 30 mV; the 12T bitcell gains 65% improvement compared with 6T bitcell.

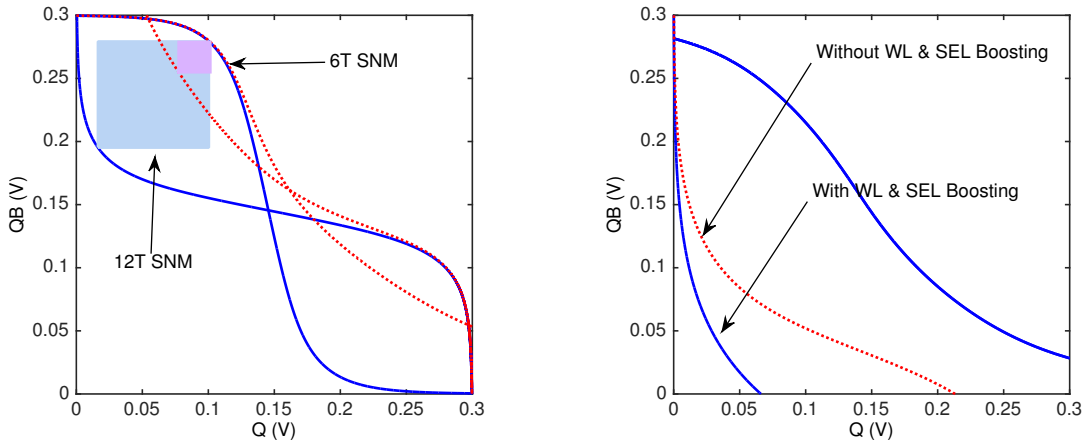


Figure 6.4: Read VTC curves comparison, 6T vs. proposed 12T bitcell and write VTC curves comparison, with an without WL and SEL voltage boosting

6.2.2 Write Operation

Figure 6.3(c) shows the 12T bitcell during the write mode; Here Q and QB have values of 0 and 1, respectively and are tried to be written over by opposite values. The bitlines BR and BL charge and discharge to V_{DD} and GND, respectively. When both wordline and select-line are enabled (WL =1 and SEL =1), BR is discharged through ML3, ML6. As position of BL and BR are exchanged in this bitcell, write data is also inverted for correct writing. Series access transistors in 12T bitcell can degrade the writability, therefore, in this work WL and SEL are boosted by 100 mV (at 300 mV supply voltage) to increase the current of series access-transistors for writability improvement.

The WNM measures how easy or difficult it is to write into the bitcell; it is the highest BL potential that can flip the bitcell data [11]. The write VTC of 12T bitcell is measured by sweeping the voltage at the storage node Q with BR, SEL and WL biased at V_{DD} and BL biased at GND while monitoring the node voltage at QB. This VTC should be used in combination with the VTC measured by sweeping the voltage at the storage node QB while monitoring the node voltage at Q. WNM can be quantified by the side of the smallest square embedded between the VTC curves. Figure 6.4(b) shows the write VTC curves with WL and SEL voltage boosting for the proposed 12T bitcell and compares with VTC curves of bitcell without voltage boosting. This figure shows how 100 mV voltage boosting for both WL and SEL improves WNM of proposed 12T bitcell.

Figure 6.5 shows the distribution of the SNM, WNM and HNM of proposed 12T bitcell and also

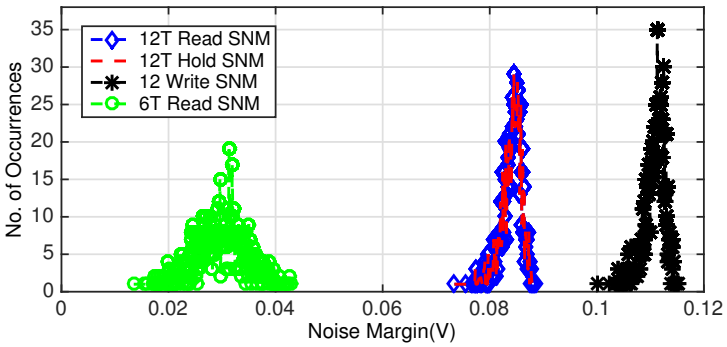


Figure 6.5: Monte Carlo simulation results for noise margin distribution in read, write and hold mode for proposed 12T bitcell (V_{DD} = 300 mV, 25^{o}C, TT).

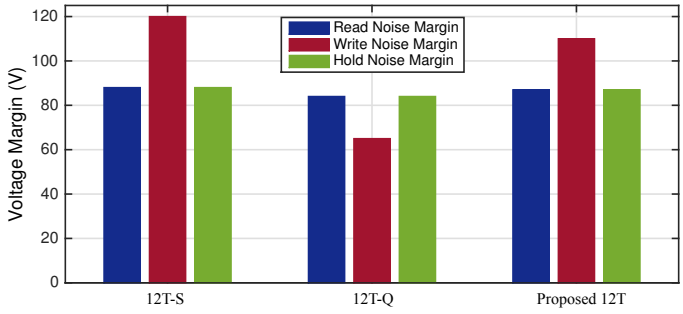


Figure 6.6: Comparison for noise margin in read, write and hold mode (V_{DD} = 300 mV, 25^{o}C, TT).

SNM of traditional 6T bitcell, in presence of local and global variations for a $1,000$ point Monte Carlo simulation at a 300 mV supply voltage. The mean value for SNM, WNM and HNM for the proposed 12T bitcell are 86 mV, 78 mV and 86 mV, respectively while SNM for 6T bitcell is 30 mV which means this cell cannot be used in small supply voltages.

Figure 6.6 compares the values of SNM, WNM and HNM of proposed 12T bitcell with 12T-S and 12T-Q bitcells in presence of variations for a $1,000$ point Monte Carlo simulation at 300 mV supply voltage. As shown in this figure all bitcell have almost equal SNM and HNM. Although 12T-S bitcell shows bigger mean value for WNM, proposed 12T bitcell has sufficient WNM to be used in 300 mV.

To evaluate the effectiveness of the proposed 12T bitcell under Process, Voltage and Temperature (PVT) variations, Monte Carlo simulations are done for both read and write modes at different process corners, supply voltages and temperatures. As shown in Figure 6.7, proposed 12T bitcell has adequate values for both SNM and WNM in all process corners as well as a wide range temperature. As shown in this figure, even at worst case conditions, 12T bitcell has bigger values for SNM and WNM compared

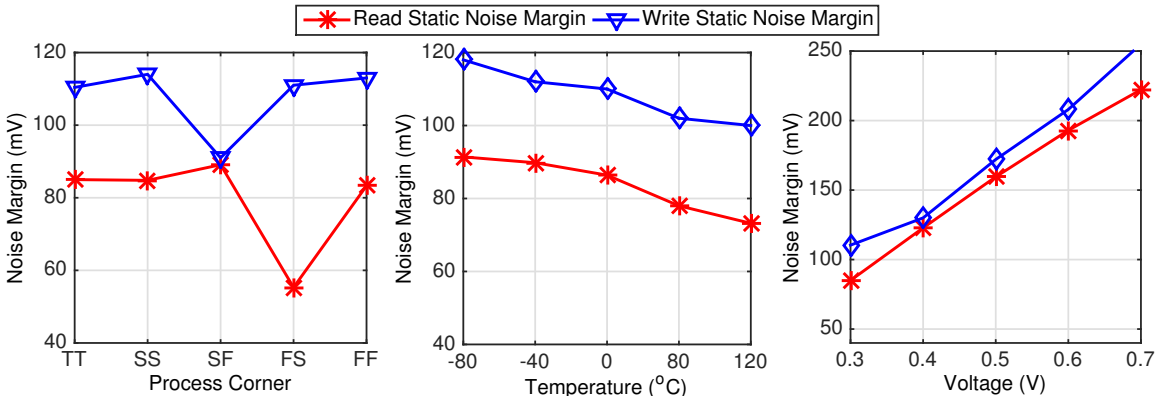


Figure 6.7: Read and write noise margin of proposed 12T bitcell in different process corners, temperatures and supply voltages.

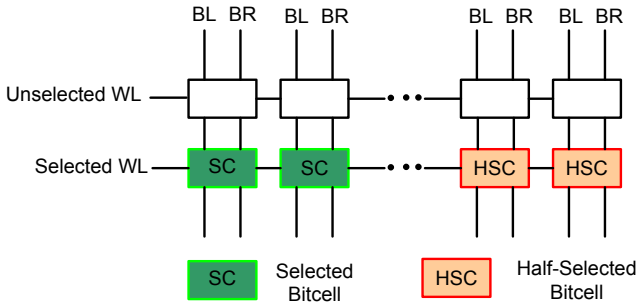


Figure 6.8: Half-select disturbance in SRAM bitcell array.

to 6T bitcell. Besides, the proposed 12T has acceptable SNM and WNM values for supply voltages in sub- and supper-threshold regions.

6.3 Half-Select Free Bitcell and Efficient Bit-Interleaving

Similar to 6T bitcells, 8T bitcell in [2], 10T bitcells in [93, 60] and 12T bitcell in [5] still suffer from the problem associated with the half-select disturbance effect. The half-select disturbance occurs when there is an half-selected column during a write operation, as shown in Figure 6.8. During this occurrence, the bitcell in the unselected column is disturbed because the wordline is raised to turn on the access transistors for selected bitcells that need to be written. The bitcell current flowing in the access transistors should be large for the written bitcells to flip the data while it should not be too large for the disturbed bitcells to avoid the data corruption [92].

Another problem in designing of robust SRAM bitcell is solving multi-bit soft errors that occur when

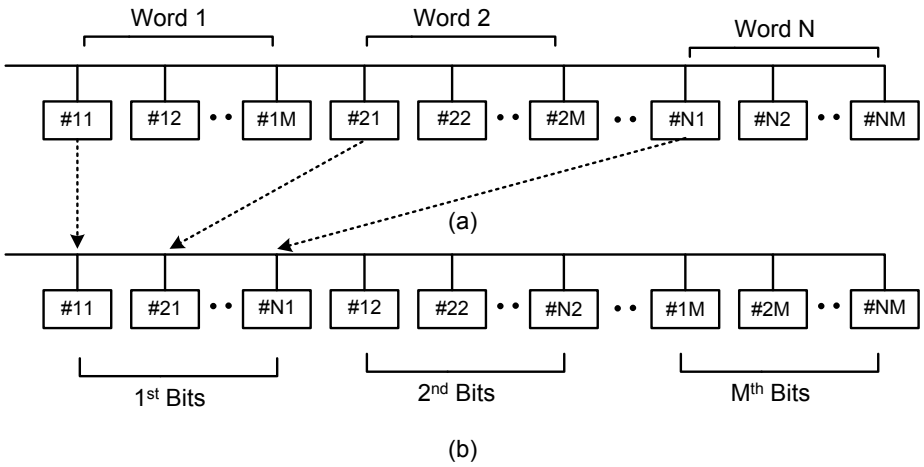


Figure 6.9: (a) Shared wordline and (b) bit-interleaving schemes.

an alpha particle or comic ray hits the memory and causes it to lose data [94]. When SRAM operates near to threshold region, cosmic rays can induce soft errors more easily because the critical charge is reduced. Multi-bit errors from a single strike usually occur in two to three adjacent bitcells [94]. Thus, to prevent multi-bit error from occurring in a single word, bits from different words should be interleaved. Shared-wordline and bit-interleaving (column-multiplexing) are common ways of arranging the words in SRAM as shown in Figure 6.9 [95]. In the shared-wordline scheme, which is widely used because of its simplicity, the probability of multi-bit errors is high because all the bits of a word are next to each other. Multiple bit errors are regarded as one single bit error in the bit-interleaving structure that is detectable and easy to correct with conventional ECC techniques. However, because of half-select issue, most of the SRAM designs cannot be bit-interleaved, and can only be implemented in shared-wordline architecture [14]. In [2] and [60], to avoid the half-select, the entire cells on a row are written at the same time which makes these SRAMs exposed to multi-bit soft errors.

In the proposed 12T bitcell, only the accessed bitcells in a row are activated for a write operation through their respective WL and SEL signals. As shown in Figure 6.10(a), although, other bitcells on the same row are selected with same wordline signals, their respective SEL are at a low level to avoid any disturbance for stored value. For the half-selected bitcells in an active column as shown in Figure 6.10(b), WL is at low level, so stored value cannot be disturbed by bitline voltage. Thus, using the proposed 12T bitcell, only one word is turn on while others are not disturbed, therefore, it is possible to implement a bit-interleaving structure with the proposed bitcell. Besides, bit-interleaving

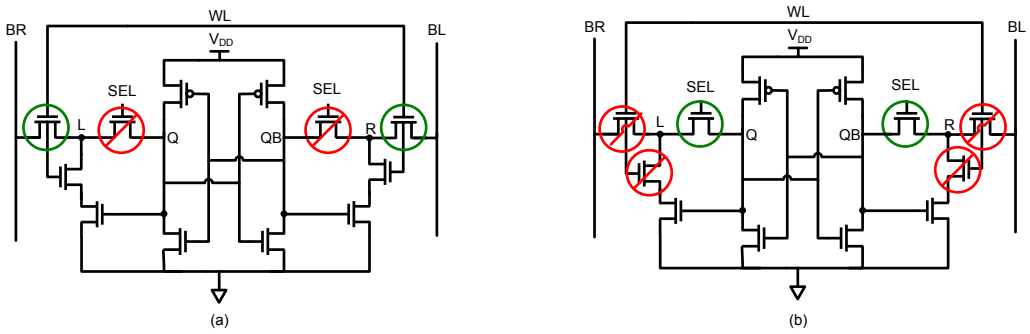


Figure 6.10: (a) Write half-selected bitcells in active row, (b) write half-selected bitcells in active column.

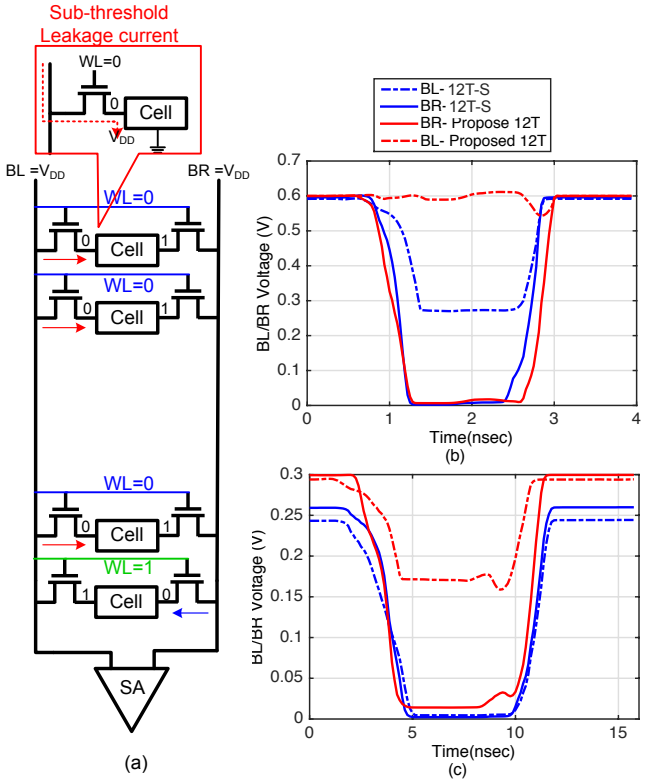


Figure 6.11: (a) Worst case bitline leakage scenario in read mode, BL/BR voltage of a column with 256 bitcells at (b) V_{DD} = 0.6 V and (c) V_{DD} = 0.3 V (12T-S [4]).

in SRAM designs allows better pitch matching between the bitcell array layout and read/write circuitry that ultimately helps to increase the bitcell density.

6.4 Leakage Control Mechanism in Read and Hold Modes

In the read operation, one bitline discharges and the other one stays high. As soon as the differential swing on bitlines exceeds the input voltage offset of sense amplifiers, data is ready on the data bus. Compare to the single ended read bitline in 12T-Q bitcell which needs a full swing to provide the correct output, in proposed 12T, differential structures speeds up the read operation with less access time. Figure 6.11(a) shows the worst case scenario for bitline leakage during read; when accessed cell holds value '1' and all other cells store '0'. In this case, leakage current by unselected cells is comparable to read current which may cause a read failure. Transient simulation results in Figure 6.11(b) and 10(c) verify the effectiveness of ML4/MR4 transistor in leakage control during read mode. Proposed bitcell have three series stacking transistors in read path which helps to reduce the leakage and allows to put more bitcells on bitline which enables smaller bitline partitioning and less area overhead. As shown in figure 6.11(b), 12T-S bitcell operates in 0.6 V supply voltages, but due to large leakage in read mode, this cell fails at 0.3 V (figure 6.11(c)) while proposed 12T control the leakages in small supply voltages and can be used in long bitlines. During hold mode (WL = 0 and SEL = 0), ML4/MR4 adds an off device in leakage path through BL and BR to GND and decreases the leakage through ML3/MR3 transistor. Besides, node L and R (Figure 6.2(a)) are float above 0 and make the V_{GS} of ML3/MR4 negative, therefore reducing the leakage current exponentially.

6.5 Final SRAM Architecture and Simulation Results

Figure 6.12 shows the block diagram of the 64 kb SRAM array with peripheral circuitries. Word selection is performed using a decoder/ multiplexer combination. To suppress the effect of random V_{TH} variation on timing variation of sense amplifier enable signal, multi replica bitline delay [3] technique is used to generate the timing of sense-amplifiers. Here, all of the comparisons are simulated by re-creating the circuits from scratch and the results stem from the simulations using 32 nm technology. To have a fair comparison with proposed 12T bitcell, an iso-configuration SRAM array is created (256×256

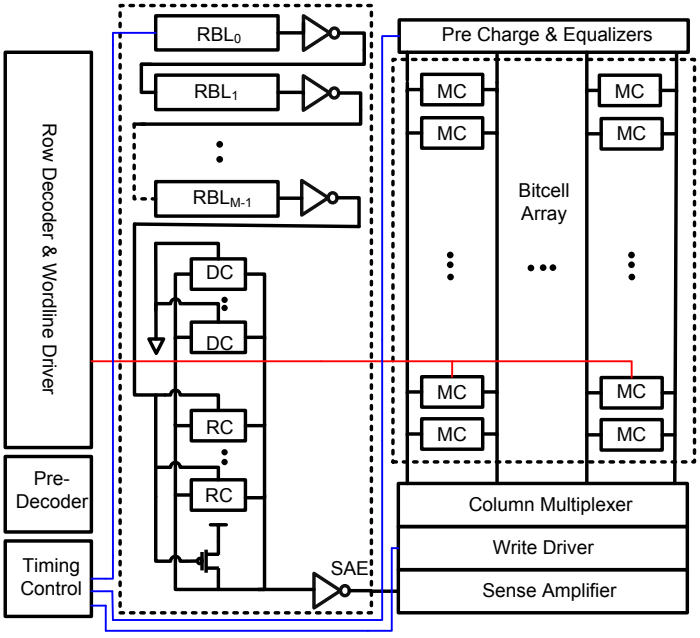


Figure 6.12: Final SRAM architecture with multi replica bitline delay [3].

bit array) for 12T-S and 12T-Q bitcells and proposed bitcell. The Monte Carlo simulation includes both the local and global variations that provide a complete representation of the variations during chip manufacturing.

Figure 6.13 shows the maximum operating frequency of the 64 kb array versus different supply voltages. This figure shows that 12T SRAM array can perform at 50 MHz with 0.3 V and also functions at 2 GHz with 0.9 V supply voltages. Therefore, This bitcell is a good option for both subthreshold and high performance operations. The minimum V_{DD} of the SRAM array is limited by read operation as the read current becomes weak with smaller V_{DD} and, therefore, a full bitline sensing is necessary for correct read operation. Figure 6.14 compares the read delay of 12T SRAM arrays at different supply voltages. Read delay is defined as the time interval from 50% of a low-to-high transition of a wordline signal until there is a 100 mV differential swing on the bitlines for proposed and 12T-S bitcell. Since 12T-Q bitcell has a single-ended read bitline, a full swing is needed which leads to bigger read delay. As shown in figure 6.14, the 64 kb array has a read delay of 0.87 ns for proposed 12T bitcell at 300 mV supply voltage and room temperature while 12T-S bitcell does not work at this voltage due to large bitline leakage and 12T-Q bitcell is very slow with a 3.9 ns read delay which is almost 4.5x slower than proposed bitcell. The leakage, dynamic and total power consumption of the SRAM arrays at different

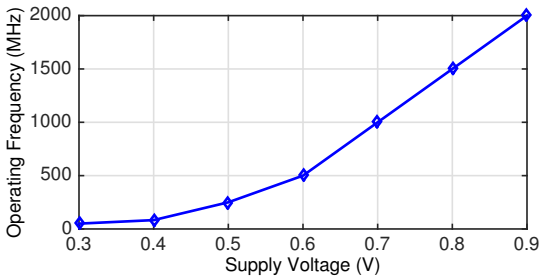


Figure 6.13: Maximum operating frequency for 64 kb array of proposed 12T bitcell versus supply voltage.

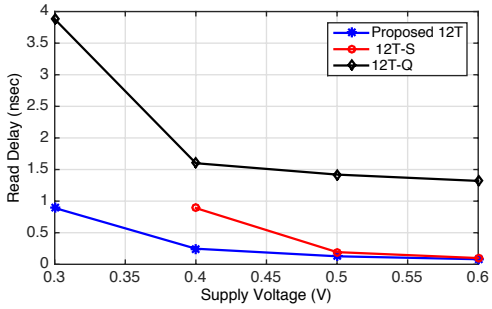


Figure 6.14: Read delay for a 256×256 array in different supply voltages (12T-S [4] fails to read at 0.3 V due to large leakage).

operating voltages is also explored, as shown in Figure 6.15. At 50 MHz and 0.3 V the leakage power for proposed 12T bitcell array is 0.95 mW which is 53% and 63% less than leakage power in arrays of 12T-S and 12T-Q bitcells, respectively. Also total power consumption in smaller using proposed 12T-S bitcell different supply voltages, as shown in Figure 6.15.

Table 6.2 summarizes and compares the performance of the proposed 12T bitcell. All SRAM arrays have same configuration and are simulated in the minimum supply voltage and maximum operation frequency. Proposed 12T bitcells shows improvement in leakage and dynamic power reduction. It has faster read operation and works in higher frequency compared to 12T-S and 12T-Q bitcells. Proposed 12T bitcell can be used on long bitlines to get less area overhead.

6.6 Summary and Conclusions

A novel subthreshold, single-port, differential-ended 12T SRAM bitcell with high performance is proposed which improves read stability and writability and allows continued scaling beyond what is possible with the 6T SRAM bitcell. This bitcell uses read buffer to improve the read stability and achieves

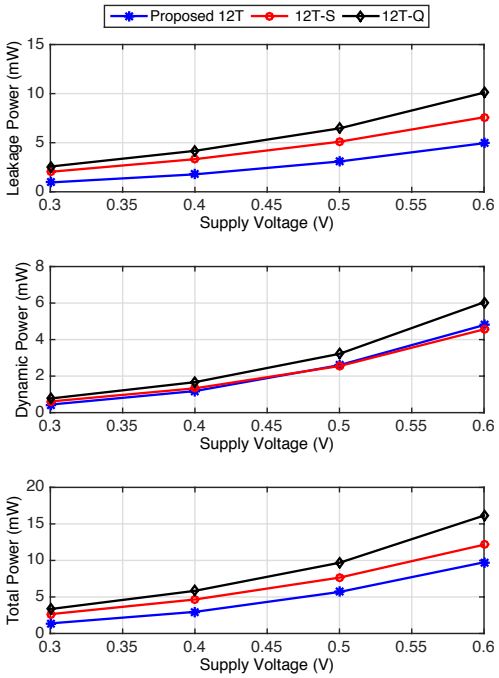


Figure 6.15: Leakage, dynamic and total power consumption comparison in different supply voltages for a 64 kb array (12T-S [4] and 12T-Q [5]).

Table 6.2: Feature list of works in 32 nm technology.

Design	**proposed 12T**	**12T-S [4]**	**12T-Q [5]**
Technology	32 nm	32 nm	32 nm
Capacity	64 kb	64 kb	64 kb
Organization	256 × 256 bit	256 × 256 bit	256 × 256 bit
Max. Frequency @ Min. V_{DD}	50 MHz@ 0.3 V	50 MHz@ 0.4 V*	25 MHz@ 0.3 V
Total Power Consumption @ 0.3 V	1.38 mW	2.64 mW	3.34 mW
Leakage Power Consumption @ 0.3 V	0.95 mW	2.04 mW	2.57 mW
Bit-Interleaving & Half-Select Free	Yes	Yes	No

* Fail to read at 0.3 V due to large leakage.

read SNM equal to hold static noise margin. Using a column-based select signal this bitcell provides half-select and read-disturb free features, facilitating bit-interleaving structure to reduce multi-bit soft errors by conventional error correcting code techniques. By boosting wordline and select signal voltage, this bitcell can read and write with no error at 300 mV. Bitline leakage suppression in 12T bitcell allows more bitcells per bitline for high density SRAMs and provides faster read operation. A 64 kb 12T SRAM macro is designed in 32 nm CMOS SOI technology that operates down to 300 mV with 50 MHz operating frequency while it functions at 0.9 V with 2 GHz operating frequency as well.

Chapter 7

Approximate SRAM Architecture for Low-Power Video Applications

7.1 Introduction

Multimedia applications require both intensive computations and large amounts of embedded memory that aggravate the total amount of power consumption. Consequently, it is of great importance for portable devices with multimedia applications to extend their battery life by lowering their power dissipation [96]. The key sources of power dissipation in multimedia applications is large embedded memory access power. Large embedded data memories are desired because of lower cost, higher performance and more reliable operation due to single packaging. However, by integrating larger blocks of embedded memory into a chip, memory failure increases and the manufacturing yield of the system drops sharply. Therefore, embedded memory is becoming the critical focus on SoCs to reach higher yield and lower power.

Supply voltage scaling is one of the most effective techniques for power reduction in VLSI systems, as switching power dissipation is quadratically dependent on V_{DD} [97]. However, during low-voltage operations, the failure probabilities of SRAM bitcells significantly increase, especially with pronounced PVT variations in scaled technologies. And, SRAM bitcell's noise margin degradation is the main reason for SRAM failure at low voltages [12]. To improve the stability and error resiliency in memory bitcells, SRAM structures with 8 [98, 2], 10 [60] and even 12 transistors [15] have been proposed.

Although using extra transistors in the bitcell helps to improve the noise margin in smaller supply voltages, it leads to more silicon area and cost. Also, to recover data from defective memory bitcells after fabrication and increase the embedded memory yield, redundant rows/columns/blocks and the use of ECC have been used in most designs. However, these redundant circuitry techniques lead to a large area overhead in current nano-scale devices with undesirable high error rates. Besides, most ECC systems can correct only a single bit of error without significant overhead while suffering from a delay penalty for the encoding/decoding of data. Therefore, to handle the high rate of errors in sub-100 nm designs, area efficient ECC techniques tend not to be as effective as they used to be.

Supply voltage reduction has been a popular technique for error tolerant designs [99, 100]. In error tolerant applications, error during storage or computation are acceptable if the design maintains an adequate quality of the output signal [100, 101]. One example of error tolerant applications is with video/image processing. For the human visual system, it is typically highly sensitive to the High Order Bits (HOBs) of the luma and chroma pixels in video data as opposed to the Low Order Bits (LOBs) [96]. Figure 7.1 demonstrates the quality of image when error is injected in a single bit position and shows how quality degradation is stronger when error happens in HOBs. The image quality degradation is acceptable until the 5th bit (assuming a lower bound of 30 dB for PSNR), but errors in 6th, 7th and 8th bits can result in significant quality degradation. Therefore, a high Bit Error Rate (BER) can be tolerable if errors happen within the LOBs, whereas, the BER in HOBs has to be considerably smaller to reach a sensible level of quality. Moreover, video data memory does not need to be 100% error free and it can have partial data loss without serious quality degradation. Besides, nano-scale systems that use 100% error free on-chip memory are unrealistic and impractical [100].

This chapter discusses an architecture to reduce the impacts of error in low voltage video memories and improve the energy efficiency. In this chapter, image processing applications are considered as case study while the idea can be easily extended to other error tolerant applications. The proposed design scales the supply voltage to improve the overall energy efficiency while HOB bitlines use a higher cell-supply during read mode and LOBs use a smaller one to maintain the quality as well. Likewise, using a supply voltage switching network within an idle (standby) mode decreases the leakage power by utilizing a lower cell supply voltage. More importantly, the proposed approach allows a dynamic reconfiguration for the number of bits using higher or lower supply voltages at run time based on a given accuracy requirement.

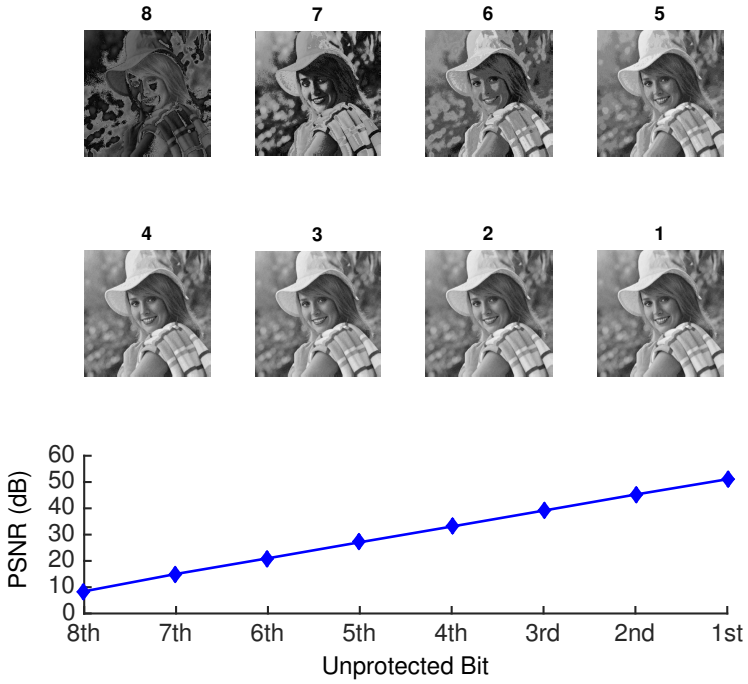


Figure 7.1: Quality degradation due to injected errors in single bit positions (8th bit is the high order bit and 1st bit is the low order bit).

Another objective of this chapter is to document the process in determining the Peak Signal to Noise Ratio (PSNR) value versus error position in approximate SRAM architectures. Previous methods have been able to extract much of this information from chip measurement [102] however, proposed approximate SRAM in this chapter extracts its information through SPICE/MATLAB simulations.

In this chapter an extensive background and comparison on different memory designs for error tolerant applications are discussed. Also, the impact of voltage scaling, terminal voltage dependencies and SRAM bitcell failure due to inadequate noise margin are explained. Detailed description of proposed SRAM array architecture is presented with simulation results and comparison to show the effectiveness of this design.

7.2 Background

In order to improve embedded memory yield in error tolerant applications while keeping the area overhead, power consumption and design cost low, several approximate SRAM designs have been proposed. Approximate SRAM designs try to decrease the failure probability by increasing the bitcell size, utilize

higher supply voltages and using ECC codes for high order bits based on specific applications. In this section, existing approximate SRAM designs are analyzed and proposed design is presented.

To reach a higher yield in lower supply voltage, two hybrid SRAM structures which are a mixture of 6T/8T SRAM bitcells [96] and 8T/10T bitcells [103] are proposed for MPEG4 video processors. In these SRAMs, higher order luma/chroma bits are stored in 8T (10T) bitcells while the lower order bits are stored in 6T (8T) bitcells. Although 8T and 10T bitcells can work in lower voltages, the 6T SRAM fails in those low supply voltages and it is more efficient to drop these bits. In [96] peripheral circuits are needed to combine a differential single wordline 6T cell with a single-ended double wordline 8T cell in the same row. Therefore, the architecture in [96] results in design challenges to combine the differential cells in the same row. Also, the [103] structure leads to an area penalty when using both 8T and 10T SRAM bitcells. Although this design tries to reduce the area overhead with bit-truncation, skipping bits of the pixel and replacing them with zero, it results in more quality degradation. Moreover, a dynamic energy-quality trade-off at run time is not possible in both designs because the BER and number of HOBs are fixed at design time.

Random dopant fluctuation are the dominant source of permanent defects in digital circuit designs that can be alleviated by increasing transistor sizing [97]. This fact is the design choice of a heterogeneous SRAM sizing algorithm for the embedded memory of a H.264 video processor [104]. In [104] the HOBs are stored in the relatively larger 6T SRAM bitcells and the LOBs are stored in the smaller ones. Compared to the [96] architecture, this heterogeneous approach offers a simpler SRAM design with straightforward layout because it only uses 6T memory cells. However, the 6T bitcell in [104] has limited improvements in stability for small supply voltages below 500 mV, and since no read/write assist circuitry is considered, this design is not applicable in near or subthreshold designs. More importantly, it does not lead to significant amount of power reduction and the number of HOBs is fixed at design time, hence, it is unable to dynamically track the time-varying quality requirement.

In another work, a partial memory protection scheme is presented that protects the SRAM data blocks with HOBs using a higher supply voltage while allowing errors in the blocks with LOBs using lower supply voltages [105]. This partial protection memory architecture has dual power rails with a static choice of the number of bits in low-voltage mode that can lead to an unreasonable quality degradation. In [106], this problem is addressed by presenting a dynamically reconfigurable SRAM array by using a lower voltage for the LOBs and a nominal voltage for the HOBs. This architecture

allows reconfiguring the number of bits within a low-voltage mode to change the error characteristics of the array. However, even with the HOBs at a higher supply voltage (faster cells), the performance is limited by the access time of the slower cells (LOBs). Moreover, each column needs a separate supply connection (bitlines, wordline and cell-supply are all at the same voltage level), therefore a big modifications is required in the array layout. Besides, to achieve large amounts of energy reduction, supply voltages need to be scaled significantly so that most LOBs fail.

The design in [107] scales the supply voltage to improve the overall energy efficiency and also adopts write-assist circuitry and error correction codes to a small number of HOBs. This design proposes a dynamic error-quality trade-off by consuming a small amount of extra energy for a few HOBs to improve the output signal quality. In [108] and [102], instead of powering the least important bits at a lower supply voltages, the LOBs are simply dropped to save more power. It is shown in [102] that the quality of the dual-V_{DD} scheme in [106] is approximately the same as the bit dropping technique in [107]; that is, the error increases in the LOBs at lower voltages which makes the LOBs mostly incorrect (i.e., it is equivalent to dropping them). In [102], a single error correction technique is utilized that reuses the dropped LOB as a check bit to protect the HOB.

In [109], a priority based error correction code for SRAM memories in H.264 processors is proposed. In this design, the single error correction Hamming ECC is used for the LOBs and the BCH (Bose-Chaudhuri, Hocquenghen) code that presents a better error correction performance is used for the HOBs. However, The ECC adds delay for data encoding/decoding as well as incurring an area overhead for the ECC logic. In addition, although the ECC provides an energy benefit over voltage scaling, at low voltages the failure rate becomes so high that double and higher order errors are more likely to occur and the single ECC scheme in [102] is not able to correct them.

In digital circuits, the switching power dissipation is often expressed as $\alpha \cdot C \cdot V_{DD}^2 \cdot f$, where α is the activity factor, C is the capacitance, V_{DD} is the supply voltage, and f is the operating frequency [11]. The switching power can be improved by reducing α, f and V_{DD} for the memory accesses. Decreasing the value of C is not easy since it is defined based on the fabrication technology. The approach in [110] reduces the bitline switching activity using a prediction-based approach and reduces the energy per access through reducing α. This design uses a 10T SRAM cell with 5 control signals per cell and leads to a serious area overhead and design complexity. Similarly, the work in [111] exploits statistical similarity in images by using an inversion bit for each word to reduce read-bitline transitions (α).

All the previously mentioned techniques, require either complex peripheral circuits that result in large penalties in performance and/or layout area or are not effective at low supply voltages. In this chapter, an area efficient 6T SRAM with three supply voltages to combat the stability degradation and reduce the leakage power at low supply voltages is proposed. Besides, the proposed architecture allows a dynamic energy-quality trade-off at run time for different inputs and various applications.

For image/video memories, error can happen when writing (encoding) due to write failure and in reading (decoding) due to read disturbance failures. The proposed dynamically reconfigurable SRAM array for low-power multimedia applications allows the architecture to reconfigure the number of bits in the low/high voltage mode to reduce the error characteristics of the SRAM array during both read and write modes. Additionally, the proposed design is different than the structure in [106] that uses a lower voltage for cells storing low-order bits and a nominal voltage for cells storing higher order bits. In this design, both HOB and LOB cells can have low or high voltage based on the operation. This spatial voltage scaling can be extremely useful for improving effective yield of multimedia memories. The proposed design provides a real-time modification, depending on the error tolerance of the application, which leads to a better energy/quality trade off. It provides a bit-level-robustness enhancement and makes it adaptable for different applications. The proposed design also allows the use of conventional 6T bitcells that leads to significant area savings compared to designs that use an 8T or 10T SRAM bitcells [96, 103]. Compared to existing error-tolerant SRAM designs, the simplicity of the proposed scheme, leads to minimizing the design effort and considerable power saving which is desired for battery-supported multimedia applications. Finally, the proposed design allows aggressive voltage scaling beyond what is possible with a single-supply design and leads to additional energy reduction.

7.3 SRAM Bitcell Failure and Read/Write Assist Techniques

Figure 7.2(a) shows the schematic of a 6T SRAM bitcell. The single-supply conventional SRAM bitcell displays limitations of competing requirements for read and write noise margin. A stronger cross-coupled pair compared to the access transistors is needed for better read stability while weaker cross-coupled pair leads to easier write operation and improvement in writability [97]. Having a dynamic cell-supply rail (Figure 7.2(b)) that switches between high and low V_{DD} offers the advantage of improving both read and write noise margins, reducing the gate leakage [112] and also decreasing subthreshold

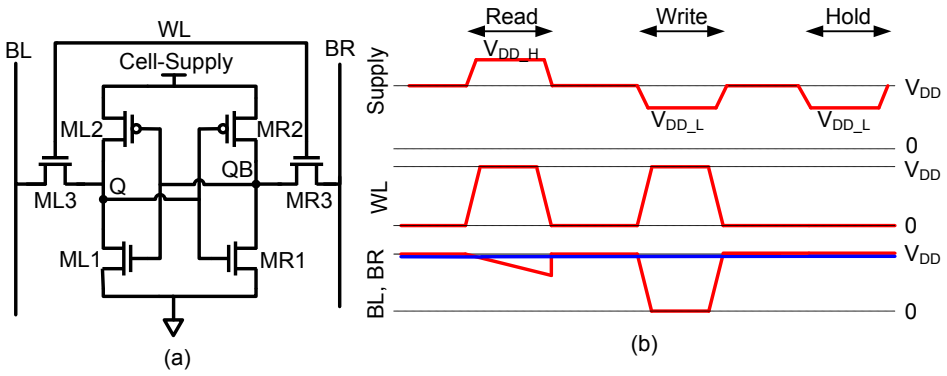


Figure 7.2: (a) 6T SRAM bitcell, (b) dynamic power supply for the bitcell.

leakage through Drain Induced Barrier Lowering (DIBL) effect [113] [114] in standby mode.

Supply voltage scaling is an effective techniques for power reduction, however, there is a limit on the minimum cell-supply voltage by the Hold Noise Margin (HNM) for SRAM bitcells. The HNM is the maximum amount of noise that SRAM bitcell can tolerate in idle mode before losing its stored value [11]. The HNM can be calculated using the VTC or butterfly curves, as shown in Figure 7.3(a). The HNM is the side of the largest square embedded between the retention VTC curves [11]. As shown in Figure 7.3(a), decreasing the supply voltage from 500 mV down to 100 mV degrades the HNM by a great factor and leads to a zero static noise margin with a 100 mV supply voltage and, therefore, produces an unstable bitcell. Random process variations make the HNM value even worse in small supply voltages. To calculate the effect of random process variations on the HNM, a $10,000$ points DC Monte Carlo simulation is performed on a 6T bitcell at different supply voltages and the results are shown in Figure 7.3(c). As shown in Figure 7.3(b), the noise margin follows a Gaussian distribution. The mean (μ) value for this distribution is the value of the HNM at that specific voltage and sigma (σ) is variation across the chip. Figure 7.3(b) also shows the fraction of distribution which has failed due noise margin constraints. Decreasing the supply voltage reduces the mean value and, therefore, degrades the stability of cell. Based on the MC simulation results in Figure 7.3(c) and approximately a 35 mV threshold voltage variation (σ value in 32 nm SOI CMOS for a minimum-sized transistor) and by considering a minimum of 50 mV for the HNM, the 6T bitcell failure probability is 99.4% in 200 mV and 2% at 300 mV. Therefore, 300 mV can be the lowest cell-supply voltage for the 6T in an error tolerant (not error free) SRAM design.

Now that minimum cell-supply voltage for 6T in 32 nm is defined, the read SNM, WNM and bitcell

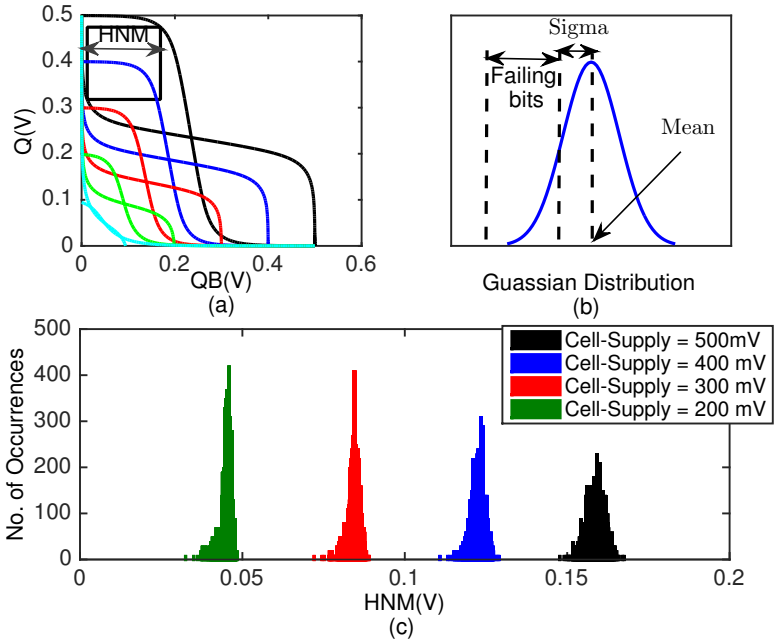


Figure 7.3: (a) 6T SRAM bitcell VTC curves in hold mode at different supply voltages, (b) failing bitcell measurement using noise distribution (c) Monte Carlo simulation results for HNM at different cell-supplies.

failure rate at different supply voltages can be calculate using the distribution of the MC results in read and write modes. Considering the dynamic supply voltage for the bitcell (Figure 7.2(b)) and to obtain a better SNM, the wordline and bitlines should be connected to voltages lower than cell-supply. Conversely, the wordline and bitlines should be connected to voltages higher than cell-supply for WNM improvement. Regular-V_{TH} transistors in the 32 nm SOI CMOS model, have V_{TH} values of 0.293 V and -0.266 V for the NMOS and PMOS transistors, respectively, and the nominal operating voltage is 0.9 V. To reduce the power consumption a 0.45 V supply voltage is used in this design.

The SNM in read mode is defined as the maximum possible noise available at the gates of the cross-coupled inverters or storage element that does not flip the bitcell value [11]. Again, read values from the VTC of the 6T bitcell can be used to measure the SNM value at different voltages. Three roots of intersection in the VTC curves are desired to indicate bistability in read operation and the SNM can be quantified by the side of the largest square embedded between the read VTC curves. Increasing the cell-supply makes a larger gate overdrive on the pull-down NMOS (M1/MR1 in Figure 7.2(a)), reduces the overlap between VTC curves and improves the SNM. Figure 7.4(a) shows the VTC curves for the 6T bitcell at different cell-supplies. The 6T bitcell has a SNM of 45 mV and 85 mV at cell-supply of

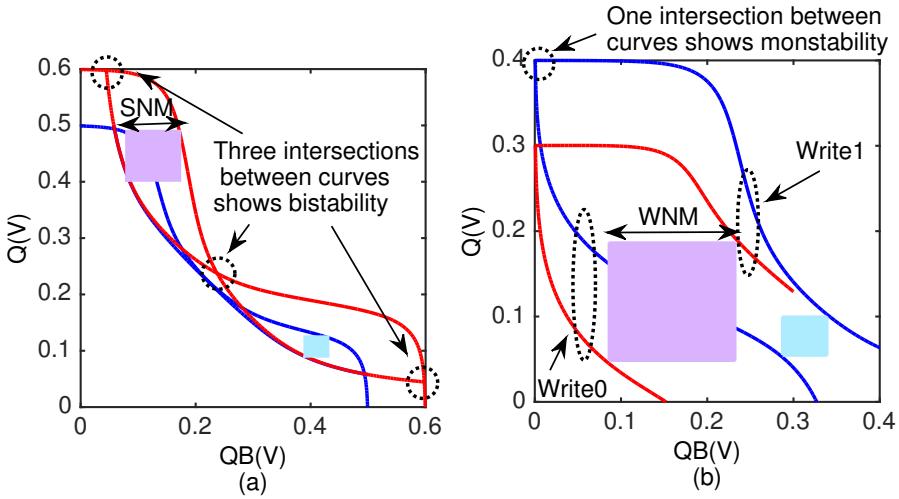


Figure 7.4: (a) VTC curves and SNM in read mode (higher cell-supply increases the SNM), (b) VTC curves and WNM during write (lower cell-supply increases the WNM). Both cases are at typical corner, wordline and bitlines are at 0.45 V and only the cell-supply changes.

0.5 V and 0.6 V, respectively, at a typical process corner (while the WL, BL and BR in both cases are at 0.45 V). Figure 7.4(a) shows how the SRAM bitcell gains 61% improvement in the SNM with only a 100 mV increase in cell-supply. Figure 7.4(b) shows the 6T VTC curves during write mode; here, Q and QB are writing over by opposite values. A monostable bitcell must have only one intersection between VTC curves during write mode. As shown in Figure 7.4(b), the WNM improves with a smaller cell-supply. Lowing the cell-supply in write mode, lowers the Write0 curve or raises the Write1 curve, therefore, improving the WNM. With only a 100 mV increase in cell-supply, the bitcell gains 110% improvement in WNM.

The bitcell static write and read noise margins are also affected by random process variations. The slow NMOS-fast PMOS (SF) corner has a fast pull-up PMOS, skews the bitcell beta ratio (M2/ M3) and decreases the ability of the NMOS access transistor to overwrite the bitcell, thereby reducing the WNM. On the other hand, the fast NMOS-slow PMOS (FS) corner skews the bitcell alpha ratio (M1/M3) by having a fast NMOS access transistor and reduces the SNM. In this design, to calculate the failure probabilities, worst process corners; *FS* corner for read and *SF* corner for write mode are considered.

Figure 7.5(a) and (b) show the distribution of the SNM and the WNM of the 6T bitcell at different cell-supplies, in presence of local and global variations and worst process corners using a $10,000$ point MC simulation (wordline and bitlines are at 0.45 V in both read and write modes). As shown in this figure, increasing the cell-supply in read mode and decreasing it during write operation, improves the

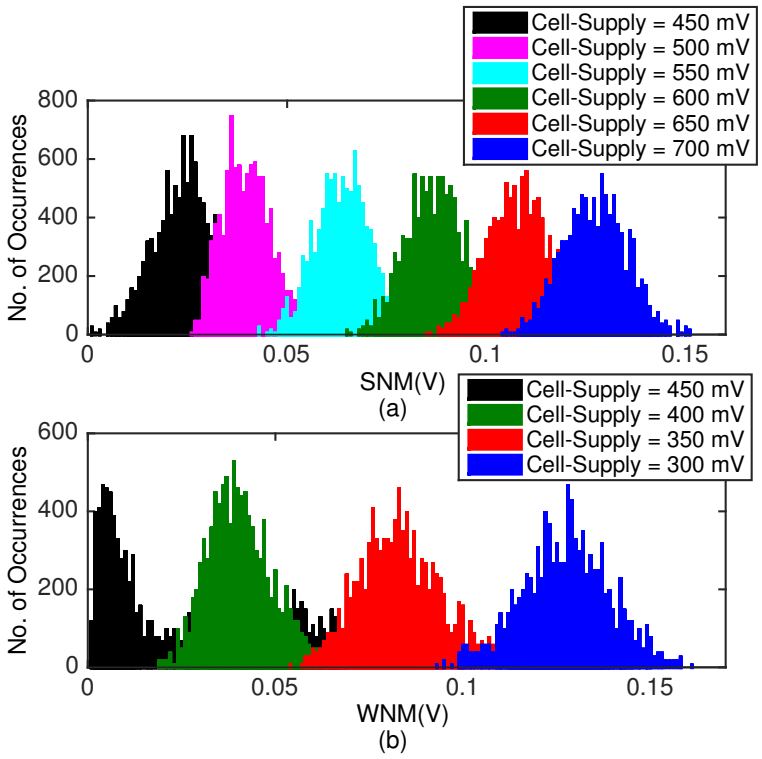


Figure 7.5: Monte Carlo simulation results for 6T SRAM cell with the dynamic cell-supply in 32 nm SOI CMOS technology (a) SNM and (b) WNM at different cell-supplies and worst-process corners (wordline and bitlines are at 0.45 V).

Table 7.1: Noise margin comparison of triple-supply and single-supply SRAM cell in 32 nm at 0.70 V.

	Triple-supply cell	**Single-supply cell**
SNM* [mV]	130	58
WNM [mV]**	135	63

* Noise margin at *FS* corner.**, Noise margin at *SF* corner.

mean value of noise margin distribution in both modes. This figure clearly shows the impact of dynamic cell-supply biasing on static noise margins. Table 7.1 summarizes the noise margin values for both read and write mode at their worst process corners and compares them with noise margin values of a single-supply SRAM cell. For the results, the sizing of SRAM transistors is the same in both cells and is provided in Figure 7.8.

The μ and σ values of SNM and WNM distributions at different supply voltages are used to predict the read and write fail probabilities, as shown in Figure 7.6(a). In read mode, the cell-supply smaller than V_{DD} (0.45 V) results in an abrupt increase in failure probability, whereas, during write mode a cell-supply greater than V_{DD} leads to a sudden increase in failure. As shown in Figure 7.6, a 100 mV

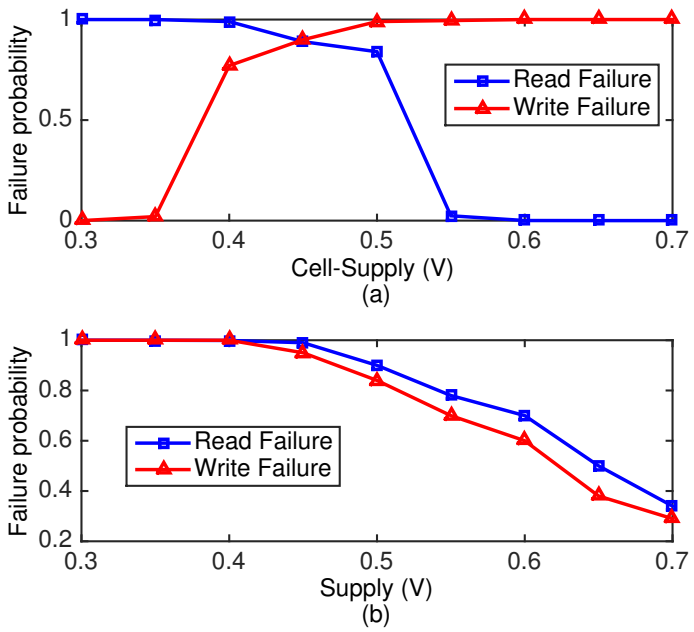


Figure 7.6: Read and write failure probability at different voltages and worst process corners (a) 6T with dynamic cell-supply (wordline and bitlines are at 0.45 V) and (b) conventional single-supply 6T.

increase in cell-supply (550 mV) in read mode and a 100 mV decrease in cell-supply (350 mV) during write operation (compared to wordline and bitlines at 0.45 V), can reduce the failure rate by 80% and 75%, respectively. Figure 7.6(b) shows the failure probability of the conventional 6T cell when the wordline, bitline and cell-supply are at the same level. As shown in Figure 7.6(b), only increasing the supply voltage results in failure probability reduction in both read and write modes. Compared to Figure 7.6(a) which demonstrates that write failures decrease with a smaller cell-supply, it is possible to reduce the failure probability while decreasing the power consumption by using a dynamic cell-supply, as shown in Figure 7.2(b).

7.4 Leakage Power Reduction in Low Voltage SRAMs

In sub 100-nm channel-lengths, the supply voltage must be decreased to assure reliability for scaled transistors; also, the threshold voltage must be decreased to maintain its performance [114]. However, reducing the threshold voltage in scaled technologies leads to increases in leakage power. In battery-supported applications, such as cell phones, leakage power can dominate power consumption and reduces the battery life time. Therefore, it is necessary to manage the leakage power in low threshold

voltage devices of new technologies.

Leakage current in SRAM not only exists in idle mode, almost all bitcells in active mode are in retention mode and consume leakage power, as well. To suppress the leakage power in SRAM different techniques have been proposed. One technique utilizes power switches to reduce the leakage and although this is effective in significantly decreasing the leakage power, it leads to great amounts of data loss and cannot be applied to SRAM design due to a noticeable performance penalty [115].

Body biasing has been proposed as an approach in [116] to reduce the leakage current without speed degradation. However, this technique is applicable in process technologies that allow independent biases for P and N wells (this requires a twin-tub technology so that the substrates of individual devices can be adjusted) and also increases the supply voltage of the SRAM that results in more dynamic power and, therefore, more total power consumption. Besides, this technique requires modifications of the SRAM bitcell structure that can result in an observable area penalty. Moreover, as the substrate bias increases, pn junction breakdown can occur and the reversed substrate bias can aggravate the effect of threshold voltage variations in scaled technologies. Optimal values of the reverse bias continue to get smaller for many sub-100 nm technologies (since the breakdown voltage of pn junctions decreases), therefore, body biasing may not be as useful in future bulk CMOS technologies.

Other techniques use a gate-grounded NMOS in the pulldown path of the SRAM bitcell [117]. This NMOS is ON in active mode and is OFF during the standby mode to reduce the leakage current through stacking transistors in the leakage path. When the NMOS is OFF, there is a virtual ground and not an actual ground which causes noise margin degradation. Also, transistor stacking makes the read operation slower and increases the dynamic power consumption. Drowsy caches use dynamic voltage scaling in idle SRAM bitlines by using a lower V_{DD} to reduce the subthreshold leakage power and preserve stored values [118]. This technique leads to over 70% of leakage power reduction while it has additional delay and power requirements for waking up a drowsy bitline. It is also shown in [112] that lowering the cell-supply voltage not only reduces the subthreshold leakage current due to a reverse body bias effect, but also reduces the gate leakage by 80% in standby mode.

In this design, a lower V_{DD} in standby mode is used to reduce the leakage power. When V_{DS} of a transistor reduces, it increases the height of potential barrier near the source terminal which, increases the threshold voltage and, therefore, decreases the leakage current. This phenomena is known as the DIBL effect [114]. In scaled technologies, such as 32 nm, the DIBL effect is more pronounced be-

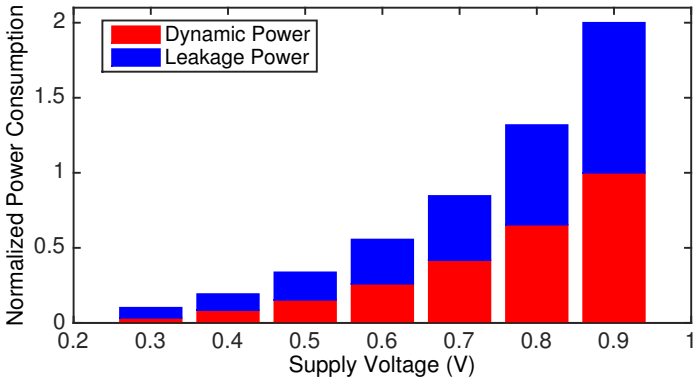


Figure 7.7: Leakage and dynamic power consumption at different cell-supplies.

cause of smaller dimensions in transistors. By supplying the unselected SRAM bitcells with a smaller cell-supply voltage, the leakage current of the SRAM can be decreased. The combined effect of less leakage current and smaller cell-supply voltages during standby mode leads to a recognizable reduction in leakage power.

Figure 7.7 shows the dependencies of the SRAM bitcell leakage power versus supply voltage in a 64 kb SRAM array architecture. Although smaller supply voltages lead to smaller amounts of leakage power, the stability of the bitcell must be considered when scaling the supply voltage. As discussed in section 7.3, there is a limit for cell-supply voltage in standby mode due to the data retention voltage. Using the minimum voltage (300 mV as data retention voltage calculated in section 7.3) may disrupt the stored value in memory bitcell through the noise on the supply rail and radiation particles [119]. Therefore, it is mandatory to consider a guard band of 50 mV over minimum voltage to combat the effect of voltage ripples on supply rail, soft errors, process variation effects and also temperature fluctuations [119]. Hence, a 300 + 50 mV cell-supply during standby mode to reduce the leakage current is considered.

7.5 Video Encoding/Decoding and Video Quality

The bit-rate of video data is quite large and without compression it is almost impossible to transmit raw data directly. H.264 and MPEG-4 are among the most powerful video compression algorithms for mobile video applications, Internet video streaming and digital signal broadcasting. Encoding and decoding are important parts of coding and the quality of video is directly related to the quality of the

embedded SRAMs for motion estimation and buffering. In a H.264 video stream, failures in the embedded SRAM array increases the video quality degradation and, subsequently, the main factor affecting the quality degradation is the locations of the SRAM failures [120]. A H.264 system allows the operation frequency to operate as low as 20 MHz, hence, delay failure can be easily satisfied using a 450 mV supply voltage in 32 nm SOI CMOS technology. Therefore, only functional failures due to stability degradation should be considered in failure calculations.

The proposed SRAM architecture allows protecting the video memory partially rather than making it error-free. A partial memory protection scheme allows combining protected and unprotected bits of a pixel data for both encoding and decoding operations. In error-tolerant applications, the error position is more important than the error total number, therefore, the bit error rate at the HOBs are reduced rather than reducing the total error rate. Figure 7.1 clearly shows the average PSNR versus error position. As shown in Figure 7.1, the error on HOBs contributes more in quality degradation compared to the LOBs.

The PSNR is the ratio of the largest pixel value and the rms of error and can be expressed as:

$$PSNR = 20 \cdot log_{10}\left(\frac{255}{\sqrt{MSE}}\right) \tag{7.1}$$

where the MSE is the mean square error between the original videos and the degraded videos and expressed as:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[Original(i,j) - Degraded(i,j)]^2 \tag{7.2}$$

an acceptable image quality is obtained in the order of 30 dB or higher [121]. Here, PSNR is used to compare the quality of images. The PSNR is typically a validated metric for quantitatively evaluated image quality [121].

7.6 SRAM Array Architecture

Figure 7.8 shows the overall architecture of the reconfigurable 64 kb (256×256 bit array, where each row includes four words and each word is 64 bits; eight 8 bit pixels per word). The proposed triple-supply

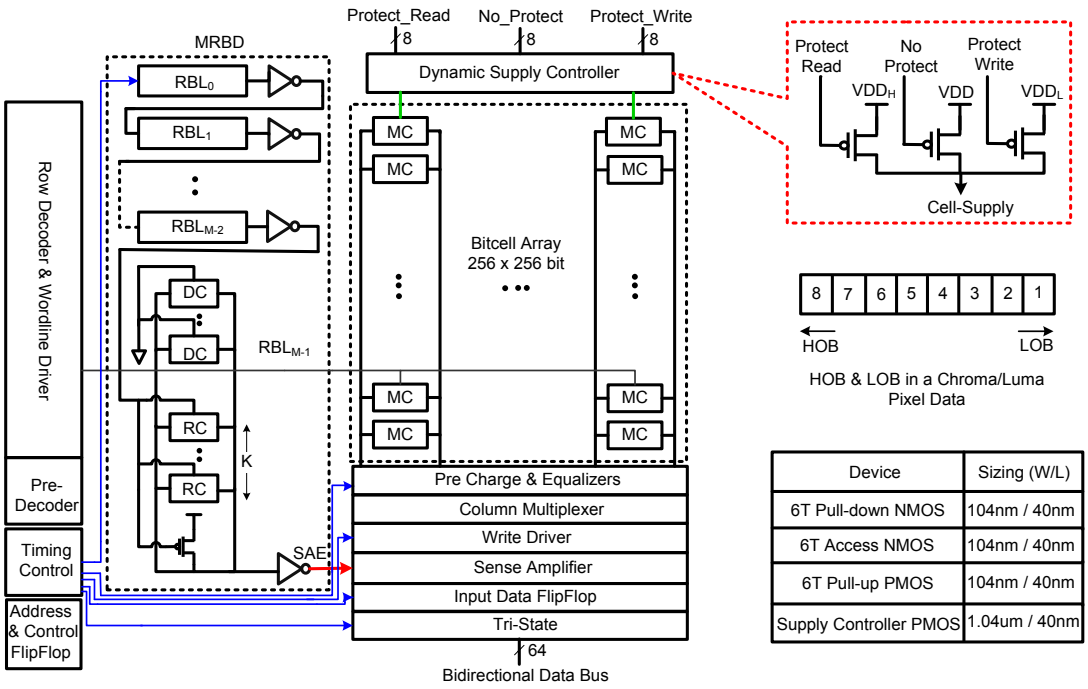


Figure 7.8: SRAM architecture with the MRBD [3] technique to control the timing of sense amplifier and transistor sizing table.

SRAM architecture has the following features:

1) In this memory, the wordline signal is shared by the cells in the same row and the bitline signals are shared by the bitcells in the same column. Supply voltage for the cell is routed vertically and supply voltages of the SRAM bitcells in one column are connected together. There is a dynamic supply controller per column, as shown in Figure 7.8, that allows the cell-supply voltage to be controlled based on its operation and desired image quality. A column based supply allows bit-by-bit reconfiguration and consist of three PMOS transistors (one connected to V_{DDH}, one to V_{DD} and one to V_{DDL}). The gate terminal of PMOSes are controlled by *protect-read*, *no-protect* and *protect-write* to reconfigure the cell-supply voltage. An 8-bit control signal is needed to control a pixel data bit by bit, so the array can work in both high-error and low-error modes.

2) Precharge, write driver, sense amplifier and decoder are connected to the scaled voltage (V_{DD}); so is the voltage level on bitline and wordline. During read mode, supply voltage of protected SRAM cells will be connected to V_{DDH} and in write/retention mode it would be connected to V_{DDL}. The cell-supply of bitcells which are not protected are connected to V_{DD} in read mode. During write operation, power supply voltage lines of protected bitcells are connected to V_{DDL}. This smaller cell-supply makes

the bitcell easier to write. However, the power supply lines for half-selected columns remain at the V_{DD} to avoid disturbances to the half-selected columns during the write mode. In read mode, power supply voltage lines of protected bitcells increases (V_{DDH}) which helps to discharge the bitline quicker and improve the read stability.

3) Although soft-error is not a serious problem for video memories, bit-interleaving is considered in this implementation to provide better protection for possible soft errors. Multi-bit soft errors occur when an alpha particle or comic ray hits a memory device and causes it to lose data [94]. When SRAM operates near to the threshold region, cosmic rays can induce soft errors more easily because the critical charge is reduced. Multi-bit errors from a single strike usually occur in two to three adjacent bitcells [94]. Thus, to prevent multi-bit errors from occurring in a single word, bits from different words should be interleaved. Multiple bit errors are regarded as a single bit error in a bit-interleaving structure and is easy to correct with conventional ECC techniques. Besides, bit-interleaving in SRAM designs allows better pitch matching between the bitcell array and read/write circuitry layout that ultimately helps to increase the bitcell density.

4) There is a negligible area overhead due to the power supply switches in dynamic supply controller. The main concern in a multiple-supply voltage design is the delay overhead of transition between operation modes. However, The operation frequency in video applications is around tens of MegaHertz and the power requirement is much more demanding than performance. Therefore, the delay overhead is a small fraction of cycle time and is not a crucial concern in multimedia applications. However, the voltage drop of the power supply switches should be low to not affect the noise margin. All modules use the minimum sized transistors except the 6T bitcell and dynamic supply controller. The sizing of transistors for these two module are shown in Figure 7.8.

5) The differential structure of the 6T SRAM bitcell allows a high speed read operation by using voltage sense amplifiers. In this design, to suppress the effect of random V_{TH} variations on timing of sense amplifier enable (SAE) signal, MRBD technique [3] is used. The MRBD technique utilizes 6T memory bitcells to control the drive delay. Replica techniques are common elements within high-speed SRAM architectures. The delay driven memory cells in the control path are the same as that of read path as well as the delay shift according to the PVT variation. Therefore, the MRBD technique attains self-timed tracking with optimal SAE timing [3].

7.7 Results

To calculate the PSNR of an image at different voltage levels (different failure probabilities), SRAM bitcell failure probabilities at different cell-supplies and worst process corners are measured through extensive SPICE Monte Carlo simulation in both read and write modes for 6T with dynamic cell-supplies and without dynamic cell-supplies (traditional 6T cell). The failure probability in read mode increases by decreasing the cell-supply while it increases in write mode by increasing the cell-supply, as shown in Figure 7.6. Using MATLAB simulations, the calculated bitcell failure are randomly added to image to mimic the behavior of random variations of die. An error equal to amount of failure is generated and randomly injected to faulty cells.

The protected cells are affected by smaller error than unprotected cells, therefore, in this approach not only is the location of failures examined, but the random distribution of fault within protected and unprotected cells are considered, as well. The bit-error rate and number of protected bits are changed in each MATLAB simulation to calculate and compare the image quality. SPICE simulations for different number of protected bits and different voltages are done to calculate the dynamic and leakage power consumption. The flow of simulation steps to get the desired PSNR value for an image is shown in Figure 7.9.

As shown in Figure 7.10, protecting 8 bits of each pixel results in a higher quality in both read and write modes. The voltage axis in this figure represents the power consumption and shows a significant amount of power can be saved with image quality degradation in read mode while in write mode quality of image increases with a decrease in power consumption. This figure shows, if the number of reconfigurable bits and bit error rate (voltage level) changes simultaneously, better power-quality trade-off can be obtained. Figure 7.10 shows an agreement with Figure 7.6(a); increasing the cell-power above 0.35 V in write mode and decreasing it below 0.55 V in read mode, leads to increases in bitcell failure by a significant factor and would not allow the image quality to improve by increasing the number of protected bits. Based on the results from this figure, protecting 8 bits/pixel in write mode at 0.35 V not only leads to higher PSNR, but also results in smaller amounts of power consumption.

The total power is the average of dynamic and leakage power in read, write and standby modes and is measured using SPICE simulation at 20 MHz. The power consumption is compared with a regular single-supply SRAM array in Figure 7.11. Since the triple-supply SRAM uses a smaller cell-supply

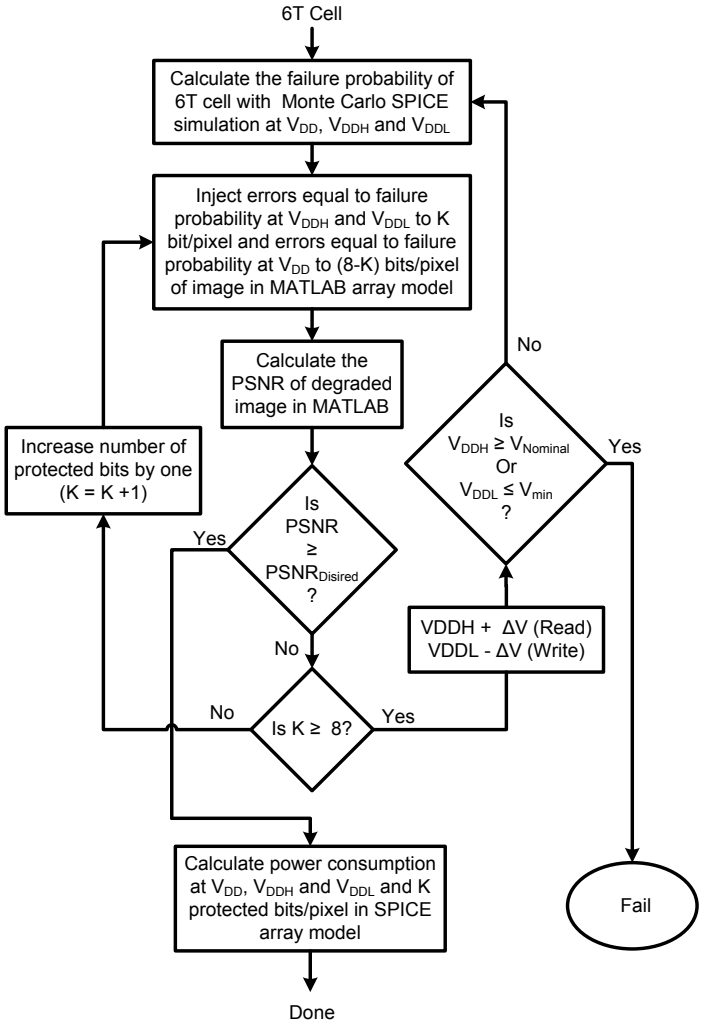


Figure 7.9: Simulation steps to get desired PSNR value using a triple-supply SRAM.

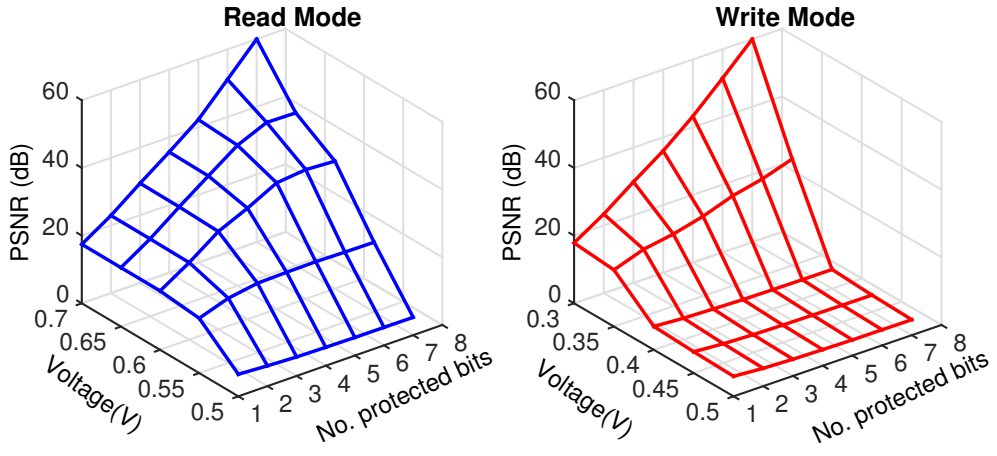


Figure 7.10: Average PSNR versus number of protected HOBs at different cell-supplies in read and write modes.

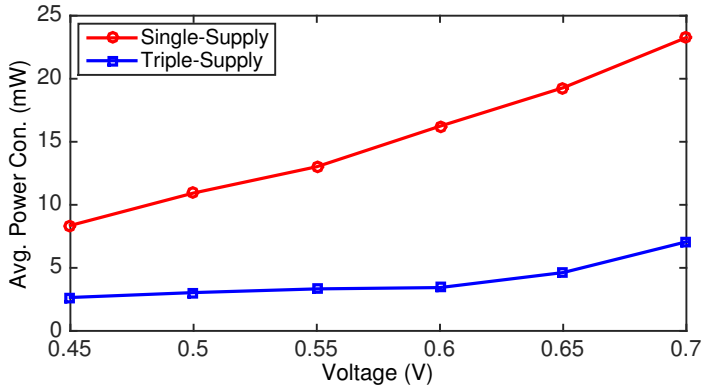


Figure 7.11: Power consumption for the single-supply and proposed triple-supply SRAM arrays (In the triple supply SRAM, the cell-supply at write and standby modes is 0.35 V and only the cell-supply in read operation is changed).

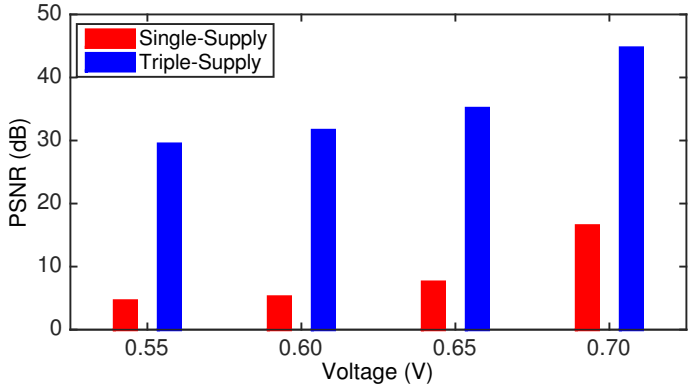


Figure 7.12: PSNR comparison at different supply voltages (In the triple-supply SRAM, the cell-supply during write and standby modes is 0.35 V and only the read cell-supply is changed, all bits are protected in the triple-supply design).

(0.35 V) during a write, the standby mode results in significant amount of power savings compared to a single-supply array architecture. At 0.70 V, the proposed triple-supply array decreases the power consumption by 69%. This design not only results in a significant power reduction, but demonstrates great improvement in image quality, as shown in Figure 7.12. As shown in this figure, the proposed triple-supply array improves the PSNR value, mainly through increasing read and write stability and decreasing the 6T bitcell failure rate. The proposed triple-supply design results in 63% improvement in image quality with no area penalty and design difficulty at 0.70 V. This figure also shows voltage scaling in a single-supply array results in producing low and impractical PSNR images.

Figure 7.13 shows how PSNR improves in higher read cell-supply and more protected number of bits. The X axis in this graph can be interpreted as the power axis (more protected bits means more

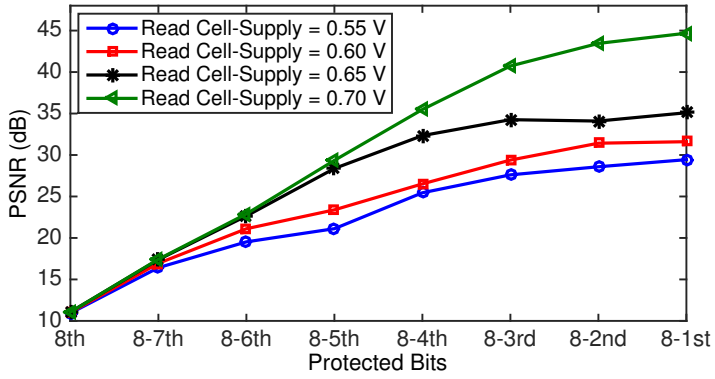


Figure 7.13: PSNR value for different cell-supplies and the number of protected bits (In the triple-supply SRAM, the cell-supply at write and standby mode is 0.35 V and only the read cell-supply is changed).

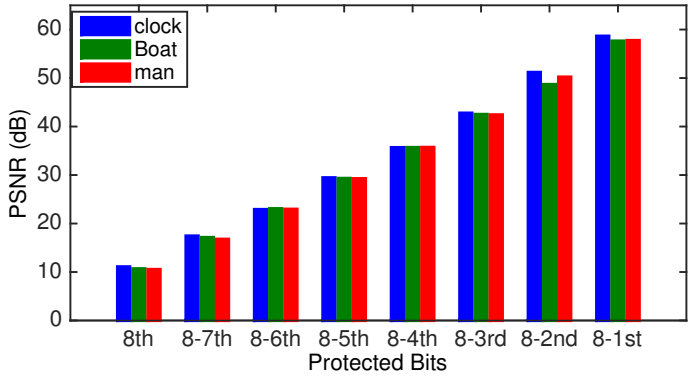


Figure 7.14: PSNR comparison for different test benches (read cell-supply = 0.70 V and write/standby cell-supply = 0.35 V).

columns are connected to V_{DDH} and more power is consumed), and shows better PSNR comes with more power consumption. This graph also shows that to have a practical value of PSNR ($\approx$ 30 dB) at 0.55 V all eight bits in pixel must be protected while for higher cell-supplies like 0.60 V and 0.70 V protecting six and three most significant bits, respectively, is acceptable. As shown in Figure 7.12, it is not possible to get a practical PSNR value for a single-supply array at 0.70 V.

The results on the proposed SRAM design is evaluated with three different benchmarks [122] (**man**: gray-scale, 1024×1024 pixels, **Boat**: gray-scale, 512×512 pixels and **clock**: gray-scale, 256×256 pixels). Figure 7.14 compares the PSNR of different benchmark at different number of protected bits using proposed design and shows negligible variation in PSNR for different images. Simulation results show that PSNR is consistent for different image benchmarks with 2 dB deviation, confirming the PSNR is a suitable metric for image quality measurements. Table 7.2 quantitatively compares the proposed

Table 7.2: Quantitative comparison of approximate SRAM designs.

	[96]	**[103]**	**[104]**	**[105]**
Design technique	Hybrid 6T&8T	Hybrid 8T&10T	Heterogeneous 6T sizing	Dual-V_{DD}
Area penalty	Yes	Yes	Yes	No
Design difficulty	High	High	High	Low
Dynamic power-quality management	No	No	No	No
Same voltage for HOB & LOB	Yes	Yes	Yes	No
	[106]	**[107]**	**[109]**	**This Work**
Design technique	Dual-V_{DD}	Write Assist & ECC	ECC	Triple-V_{DD}
Area penalty	No	No	No	No
Design difficulty	High	High	High	Low
Dynamic power-quality management	Yes	Yes	Yes	Yes
Same voltage for HOB & LOB	No	No	Yes	No

design with similar approximate SRAM designs. As one can conclude from this table, the proposed triple-supply design is the only design that provides dynamic power-quality trade-off at run time with no area overhead and no design complexity. With the proposed design it is possible to improve the power-quality trade-off even when chip is skewed to worst process corners for read or write mode.

Figure 7.15 shows several sample images for the proposed and single-supply SRAM arrays. As shown in this figure, scaling the supply voltage in single-supply 6T SRAM array decreases the quality of image by a significant amount while the proposed SRAM keeps the quality of image through protecting most significant bits of pixel data in read mode and protecting all bits during write mode. As shown in this figure, the practical value of the PSNR can be achieved at different voltages by protecting a different number of bits. Figure 7.16 shows the trade-off between image quality and power consumption in proposed design.

To achieve higher image quality, more bits need to be protected in read mode and this results in more power consumption. This figure shows how quality can be traded with power consumption based on application and the desired output signal quality using a triple-supply SRAM. On the other hand, in the single-supply SRAM, as shown in Figure 7.11 and Figure 7.12, a higher supply voltage leads to more power consumption and adds no improvement in the PSNR. Table 7.3 summarizes the performance of a 64 kb array using single-supply and proposed triple-supply design in 32 nm SOI CMOS technology. As shown in this table at the same voltage and operating frequency triple-supply array results in less power consumption and better image quality with a negligible increase in area.

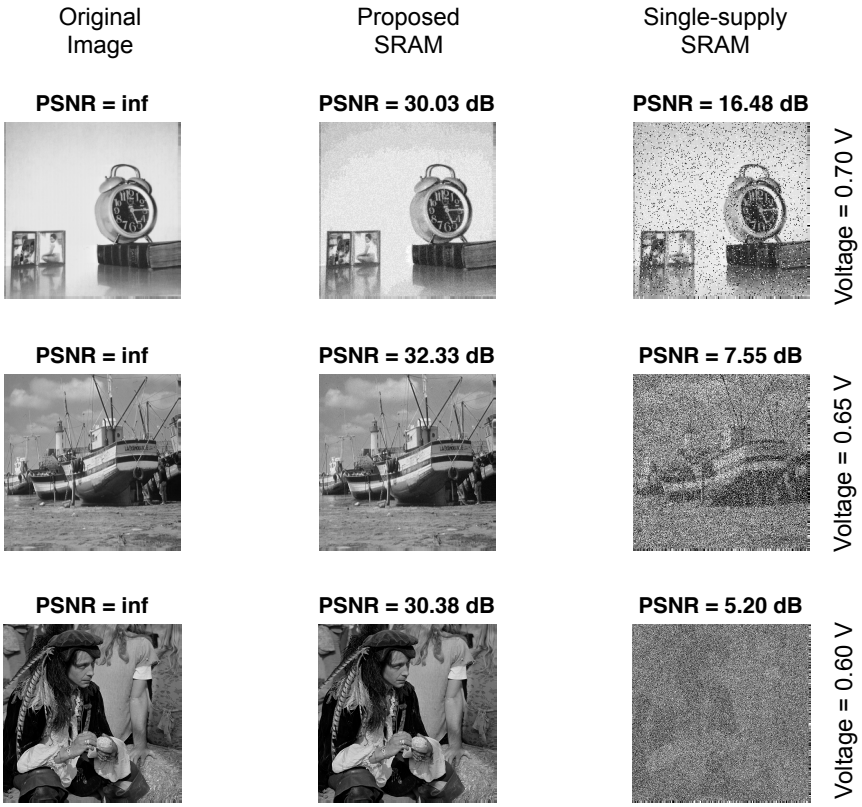


Figure 7.15: Sample images at different voltages for the triple-supply and single-supply SRAM. (@ 0.70 V bits 8th-6th are protected, @ 0.65 V bits 8-5th are protected and @ 0.6 V bits 8rd-3th are protected). In the proposed SRAM, the cell-supply during write mode is 0.35 V for all cases and only the cell-supply in read mode changes.

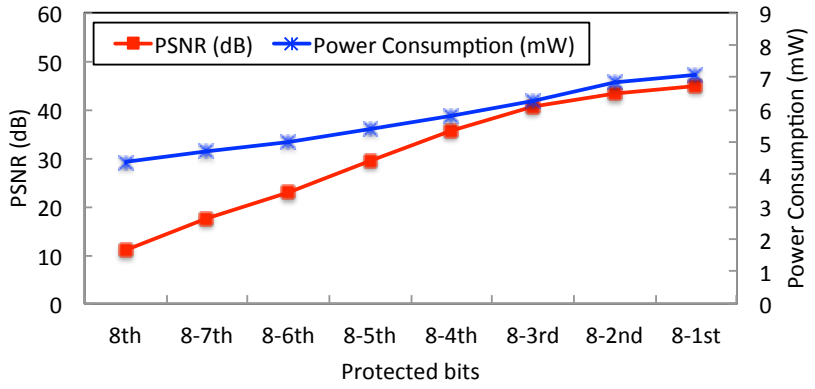


Figure 7.16: PSNR and power trade-off for the proposed SRAM architecture (cell-supply at write mode is 0.35 V and at read mode is 0.70 V).

Table 7.3: Comparison of 64 kb triple-supply and single-supply SRAM designs in 32 nm.

Deign	**Triple-Supply**	**Single-Supply**
Voltage [V]	$V_{DDH} = 0.70$ $V_{DDR} = 0.45$ $V_{DDL} = 0.35$	$V_{DD} = 0.70$
No. of protected bits/pixel	8	0
Max PSNR [dB]	58	16
Power consumption [mW]	7.069	23.268
Frequency [MHz]	20	20
Total No. of transistors	$402,918$	$402,150$

7.8 Summary and Conclusion

Voltage scaling is used to reduce the power consumption in an error tolerant SRAM video application. To improve the SRAM bitcell noise margin in scaled voltages, cell-supply is decreased in write mode and is increased in read mode respect to wordline and bitline voltage level. In video memories, higher order bits of pixel data are stored in protected bitcells (bitcells with larger cell-supply during read and smaller cell-supply in write mode) while lower order bits are stored in unprotected bitcells (same cell-supply voltage as wordline and bitlines voltage level). This approach allows to improve the image quality and at the same time keep the power consumption low. The proposed triple-supply approach achieves 63% improvement in image quality and 69% reduction in power consumption compared to a single-supply 64 kb SRAM array at 0.70 V. The proposed design allows low-power SRAM implementation with minimum design changes and negligible area overhead compared to conventional 6T SRAM arrays and also provides a dynamic power-quality trade-off at run time. In proposed design power can be traded with output signal quality at run time, therefore, this SRAM can be used in different approaches of video memories from mobile video applications to video broadcasting.

Chapter 8

Conclusions and Future Work

8.1 Thesis Contributions

SRAMs have become a standard component embedded in all SoC, ASIC, and micro-processor designs. Their wide application leads to a variety of requirements in circuit design and memory configuration. The regular structure of memories leads well to automation that produces size and configuration variations quickly, but developing this with multiple technologies and tool methodologies is challenging. Because, memory designs play a significant role in overall system performance and costs, memory compiler is a critical tool. In Chapter 3 an open-source memory compiler called OpenRAM, is proposed. OpenRAM is a technology independent memory compiler, written in Python, flexible and portable, therefore can be used by students and researchers to generate different types of memory in different technology nodes. Table 8.1 summarizes the needed input and dependencies and generated output files by OpenRAM compiler.

As variability concerns mount in future CMOS technologies, SRAM cell stability, which depends on delicately balanced transistor characteristics, becomes a significant concern. In conventional 6T, bitcells must be stable during read and writable during write mode. Ignoring redundancy, such functionality

Table 8.1: Inputs, dependencies and outputs of OpenRAM compiler.

Inputs & Dependencies	**Outputs**
Technology Library (layermap, tech rules, transistor models,..)	GDSII layout & .lef file
User specifications (word size, memory size, aspect ratio,..)	Spice & Verilog netlists
Spice simulator (ngspice, hspice,..)	Liberty (.lib) file
Python2.7, Layout viewer/editor & Calibre	DRC & LVS check results

must be preserved for each bitcell under worst-case variation. Transistor strength ratios must be chosen such that cell static noise margin and write margin are both maintained, which presents conflicting constraints on the 6T SRAM bitcell transistor strengths. This delicate balance of transistor strength ratios can be impacted by device variation, which dramatically degrades stability and writability, especially in scaled technologies.

To circumvent variability problems, Chapter 4 introduced a novel 8T SRAM bitcell to improve the read stability and writability of SRAM in ever increasing process variations of scaled technology. Proposed 8T bitcell is fully differential with one port and one wordline, so it doesn't need any architectural change in current memory compilers for 6T bitcell and can be used as a replacement for 6T bitcell. 8T bitcell improves the stability during read mode by increasing the strength of pull-down transistors while it enhances the writability by making access-transistors stronger.

High-speed and low-power SRAM is greatly desired for various applications, especially for mobile applications. As V_{DD} is decreased, the power is reduced in proportion to the square of V_{DD} and cycle time deteriorates. It is essential to suppress cycle time deterioration at low-voltage operation. In SRAM read operation, discharging the bitline is the most time consuming procedure. Therefore, the timing for sense amplifier enable signal is extremely significant for the high-speed and low-power SRAM. The optimum timing for sense amplifier enable signal exists for the high-speed SRAM read operation while it is shifted by PVT variations. Therefore, the sense amplifier timing must be determined in relation to the PVT variations.

In order to improve the access time of SRAM and speed up the read operation, multi replica bitline delay (MRBD) and reconfigurable replica bitline delay (RRBD) techniques are proposed in Chapter 5. MRBD generates a more accurate sense amplifier enable signal compared to conventional replica bitline [43] technique, leads to less access time for SRAM. MRBD uses multiple replica bitline and replica cells to suppress the effect of process variations and shows 50% less variation in SAE signal compared to RBL with a negligible area overhead. RRBD has the same structure as MRBD but the number of replica columns and replica cells change with a digital control code to generate the optimum set time of sense amplifier. RRBD allows calibration after fabrication and recalibration due to device aging degradation.

One solution to low-voltage SRAM is designing new SRAM bitcells with great performance and stability in small supply voltages. Chapter 6 introduced a subthreshold 12T SRAM bitcell. Proposed 12T bitcell shows great improvement in noise margin during read and write modes at small voltages below

Table 8.2: Summary of proposed designs and techniques.

Chp	Proposed Design	Features
3	OpenRAM compiler	Open-source, Python-based Technology-independent Provides timing/power characterizer Provides reference libraries for FreePDK45 and SCMOS technologies
4	Novel 8T SRAM Bitcell	Differential-ended and single-port Improves read stability and writability
5	MRBD technique	Generates sense amplifier enable signal with less deviation Reduces access time and power consumption
5	RRBD technique	Generates accurate sense amplifier enable signal Allows calibration after fabrication using a control code Allows recalibration due to device aging degradation
6	Novel 12T SRAM Bitcell	Differential-ended and single-port Half-select disturbance free and allows efficient bit-interleaving Controls the bitline leakage current in read mode Works in subthreshold regions (suitable for low-power applications) Improves read stability and writability at low voltages
7	Approximate SRAM	Triple-supply voltage No extra area penalty or design difficulty Suitable for low-power multi-media applications Allows dynamic power-quality management

threshold voltage, where 6T and 8T cannot work and some 10T bitcells have limitation in writability. Proposed 12T has fully differential structure, therefore shows better noise and mismatch immunity compared to single ended designs. Its half-select and read disturb free features allows efficient bit-interleaving structure to solve multi soft errors with conventional ECC techniques.

Chapter 7 introduced an error tolerant SRAM architecture for video applications. Proposed design uses lower cell-supply in write mode and higher cell-supplies in read mode respect to wordline and bitline voltage level. In proposed video memory, higher order bits of pixel data are stored in protected bitcells (bitcells with larger cell-supply during read and smaller cell-supply in write mode) while lower order bits are stored in unprotected bitcells (same cell-supply voltage as wordline and bitlines voltage level). Proposed design improves the image quality and at the same time keeps the power consumption low. The proposed design has three supply voltages and allows low-power SRAM implementation with minimum design changes and negligible area overhead compared to conventional 6T SRAM array and also provides a dynamic power-quality trade-off at run time.

Table 8.2 summarizes the proposed techniques and designs of this dissertation.

8.2 Future Work

1. Improving the runtime of OpenRAM characterizer:

   As explained in Chapter 3, OpenRAM includes a memory characterizer that measures the timing and power characteristics through Spice simulation. This characterizer generates the Spice stimulus and runs the circuit simulations to produce the characteristics in a Liberty (.lib) file. In the first release of OpenRAM, characterizer uses the entire Spice netlist of SRAM macro for simulation, therefore the runtime of large SRAM blocks is very slow. Besides, to simulate the full range of process, voltage and temperature corners and also the effects of process variation, the number of characterization runs and the data processing time per run grow exponentially. In order to improve the run time it is possible to partition a SRAM macro into sub blocks each with a few hundred transistors and then submit each partition for characterization. Next, characterizer should be modified to assemble and compress all the characterized library data for each partition into a single output Liberty file. By automatic partitioning and Spice stimulus generation, this technique can greatly improve the characterization speed.

2. Adding Approximate SRAM design to OpenRAM:

   Approximate SRAM offers the opportunity to improve the power consumption of computer systems for specific applications. Approximate SRAM designs like dual-voltage SRAMs (where high voltage cells can be used to store most important bits of data and low voltage cells store less important bits) can be added to OpenRAM. In fact, OpenRAM can be modified to generate approximate SRAMs based on the level of precision. Quality of the output signal, technology parameters and power constraint can be used to determine the number of high and low voltage SRAM cells. Statistical calculations can determine the failure rate of SRAM cells at different supply voltages or SRAM cells in different sizes and results can be added as an input table to compiler. Based on the data of this table and desired accuracy of output signal, best configuration of approximate SRAM that leads to less power consumption can be generated.

3. Utilizing Reverse Short Channel Effect (RSCE) to optimize the 12T SRAM bitcell:

   Short channel devices have been optimized for regular super-threshold circuits to meet various device objectives such as high mobility, DIBL, low leakage current, and minimal V_{TH} roll-off [123].

However, a transistor that is optimized for super-threshold may not be optimal for achieving high performance and low power in the subthreshold region. Short Channel Effect (SCE) is an undesirable phenomenon in short channel devices where V_{TH} decreases as the channel length is reduced. Variation in device critical dimensions translates into a larger variation in the threshold voltage as SCE worsens with increasing DIBL [124]. Traditionally, nonuniform doping was used to mitigate this problem by making the depletion widths narrow and hence reducing the DIBL effect. As a byproduct of this technique, a short channel device shows, reverse short channel behavior where the V_{TH} decreases as the channel length is increased [125] [126]. In subthreshold circuits, the SCE mechanism is not as strong as in super-threshold circuits. On the other hand, RSCE is still significant enough to affect the subthreshold performance due to the reduced DIBL and the exponential dependency of current on threshold voltage. Applying an effecting device sizing for subthreshold circuits utilizing RSCE helps to achieve high drive current, low device capacitance, less sensitivity to random dopant fluctuations, and better subthreshold swing. RSCE can be utilized in re-designing the proposed 12T SRAM bitcell and find the optimum transistor sizes for subthreshold regions.

Chapter 9

Appendix A: MATLAB Codes for SNM Calculation

```
1
2  function SNM_MonteCarlo()
3  clear all;
4
5  % This is the only input file which is the extracted waveforms from the
       spice simulations. To get this file, you can simply select the
       waveforms in WaveformViewer and extract them in a .txt file.
6  Y1 = importdata ('6T_read.txt');
7
8  Step = 0.0001; % Define the step of extracted simulation results
9  VDD = 0.3;  % Define the supply voltage
10
11 X1_read = 0:Step:VDD;
12 Y1_read = Y1;
13 X2_read = Y1_read;
14 Y2_read = X1_read;
15
16 Spline_Derivative(X1_read, X2_read, Y1_read, Y2_read, VDD, Step)
17 end
18
19 % This function reads the input file, plots the VTC curves and calculates
       the sid of biggest square can be fitted between the lobe of VTC curve
       and plots the Guassian of all SNM values.
20 function Spline_Derivative(X1, X2, Y1, Y2, VDD, Step)
21
22 [x1, index1] = sort(X1);      % sort the X and Y values
23 sx1 = 0:Step:VDD;
24 sy2 = min(Y2):Step:max(Y2);
25 SNM=[];
26
27 for nline=1:size(Y1, 2) % Loop is based on the size of MC Simulation
```

```
28
29 [x2, index2] = sort(X2(:, nline).');
30 y2 = Y2(index2);
31 y1 = Y1(index1, nline).';
32
33 % SPline function uses cubic spline interpolation to find cs1 at the points
       in the vector x1
34 cs1 = spline(x1,[0 y1 0]);
35 cs2 = spline(y2,[0 x2 0]);
36
37 % ppval function evaluates the piecewise polynomial cs at the query points
      sx.
38 sy1 = ppval(cs1,sx1);
39 sx2 = ppval(cs2,sy2);
40
41 [sy2, index] = sort(sy2,'descend');
42 sx2 = sx2(index);
43
44  %dorderth derivative of the cs function
45 p_der1=fnder(cs1);
46
47 % find the intersection between two VTC curves
48 midPos2 = find(abs(sx2-sy2)<0.001);
49 midPos2 = midPos2(end);
50
51 % find the points where slope is close to -1
52 slop1 = find (fnval(p_der1,sx1) <= -0.999 & fnval(p_der1,sx1) >= -1.001);
53
54 mX1 = sx1(slop1(1));
55 mY1 = sy1(slop1(1));
56
57 mx2=[];
58 my2=[];
59
60 for i = 1 : midPos2
61 xclosing = find (abs(mX1-sx2(i)) - abs(mY1-sy2(i)) < 0.001);
62 if (xclosing)
63 mx2=[mx2 sx2(i)];
64 my2=[my2 sy2(i)];
65 end
66 end
67
68 % find the point where slope is closest to -1
69 mX2 = mx2(1);
70 mY2 = my2(1);
71
72
73 % calulate the SNM as side of largest square
74 biggest_square_side = abs(mX1-mX2) ;
75 SNM = [SNM biggest_square_side];
76
77 % Plot VTC1
78 plot(x1, y1, 'b-');
79 hold on;
```

```
80
81 % Plot VTC2
82 plot(x2, y2, 'b-');
83
84 % Plot the biggest square between VTC1 and VTC2
85 rectangle('Position',[mX2, mY2, biggest_square_side, biggest_square_side],
       'LineWidth', 1, 'EdgeColor', 'r');
86 end
87
88
89 % Round the SNM values to 4 digit
90 SNM = sort (round (SNM ,4));
91
92 % Calculate the occurance of SNM values
93 occur = [];
94 for i= 1: 1: size(SNM,2)
95 occur = [occur sum(SNM(:) == SNM(i))];
96 end
97
98 % Plot the SNM distribution
99 figure;
100 plot (SNM, occur, 'LineWidth',2,'MarkerSize',8);
101 hold on;
102 xlabel('Noise Margin(V)','FontSize',14, 'Color','k');
103 ylabel('No. of Occurrences','FontSize',14, 'Color','k');
104 set(gca,'FontSize',14, 'LineWidth',1.2);
105 hold on;
106 grid on;
107 drawnow;
108 end
```

Chapter 10

Appendix B: MATLAB Codes for PSNR Calculation

```
clear;

clock = imread('clock.tiff'); %Read the sample image e.g. clock.tiff

% Errors equal to calculated failure-rate values (extracted from noise
    margin distribution) is added to the image. e.g. failure-rate1 =
    0.00001 and failure-rate2 = 0.99
A11= imnoise(clock,'salt & pepper', 0.00001);
A12= imnoise(clock,'salt & pepper', 0.99);

% Create file to write the degraded images
vec1 = fopen('vector1.txt', 'wt');
vec2 = fopen('vector2.txt', 'wt');
vec3 = fopen('vector3.txt', 'wt');

for i=1:size(A11,1)
for j=1:size(A11,2)

% Read the pixels of degraded images
str1= (num2str(dec2bin(A11(i,j), 8)));
str2= (num2str(dec2bin(A12(i,j), 8)));

% Write the pixels of degraded image into new files
fprintf(vec3, '%d ', bin2dec(num2str(str1)));

% Concatenate bits of pixels of degraded images. e.g. first bit of str1/A11
    and seven bits of str2/A12 form the pixel of degraded image
CON = strcat(str1(1), str2(2:8));

% Write the concatenated pixels of degraded image into new files
fprintf(vec1, '%d ', bin2dec(num2str(CON)));
```

```
29 fprintf(vec2, '%s ', num2str(CON));
30 end
31
32 fprintf(vec1, '\n');
33 fprintf(vec2, '\n');
34 fprintf(vec3, '\n');
35 end
36
37 fclose(vec1);
38 fclose(vec2);
39 fclose(vec3);
40
41 % Read back the concatenated pixels of degraded image to show image and
42 % calculate PSNR
43 A1= importdata('vector3.txt');
44 figure; imshow(uint8(A1));
45
46 A2= importdata('vector1.txt');
47 figure; imshow(uint8(A2));
48
49 PeakSNR1 = psnr (uint8(A1), clock);
50 PeakSNR2 = psnr (uint8(A2), clock);
51
52 fprintf('\n The PSNR1 value is %f', PeakSNR1);
53 fprintf('\n The PSNR2 value is %f', PeakSNR2);
```

Chapter 11

Appendix C: Acronyms

6T : Six-Transistor SRAM Bitcell

8T : Eight-Transistor SRAM Bitcell

10T : Ten-Transistor SRAM Bitcell

12T : Twelve-Transistor SRAM Bitcell

ASIC : Application Specific Integrated Circuits

BER : Bit Error Rate

BL : Bitline

BTI : Bias Temperature Instability

CMOS : Complementary Metal-Oxide-Semiconductor

CRBL : Configurable Replica Bitline

DIBL : Drain Induced Barrier Lowering

DRC : Design Rule Check

DSM : Deep Sub-Micron

FoM : Figure of Merit

GMC : Generic Memory Compiler

HNM : Hold Noise Margin

IC : Integrated Circuit

LVS : Layout Versus Schematic

MC : Monte Carlo

MRBD : Multi Replica Bitline Delay

NMOS : N-channel metal-oxide semiconductor

PDK : Process Design Kits

PMOS : P-channel metal-oxide semiconductor

PSNR : Peak Signal to Noise Ratio

PVT : Process, Temperature, Voltage

RBL : Replica Bitline

RC : Replica Cell

RRBD : Reconfigurable Replica Bitline Delay

SA : Sense Amplifier

SAE : Sense Amplifier Enable

SCMOS : Scalable CMOS

SINM : Static Current Noise Margin

SNM : Static Noise Margin

SOC : System-on-Chip

SOI : Silicon On Insulator

SPNM : Static Power Noise Margin

SRAM : Static Random Access Memory

SVNM : Static Voltage Noise Margin

VLSI : Very Large Scale Integration

VTC : Voltage Transfer Characteristic

WL : WordLine

WNM : Write Noise Margin

WTI : Write Trip Current

WTP : Write Trip Power

WTV : Write Trip Voltage

ZBT : Zero Bus Turnaround

Bibliography

[1] K. Nii, Y. Tsukamoto, T. Yoshizawa, and H. Makino, "A 90 nm dual-port SRAM with $2.04\ um^2$ 8T-thin cell using dynamically-controlled column bias scheme," in *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 508 – 543, February 2004.

[2] L. Chang, R. Montoye, Y. Nakamura, K. Batson, R. Eickemeyer, R. Dennard, W. Haensch, and D. Jamsek, "An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 956 – 963, April 2008.

[3] S. Ataei and J. E. Stine, "Multi replica bitline delay technique for variation tolerant timing of SRAM sense amplifiers," in *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 173 – 178, May 2015.

[4] Y.-W. Chiu, Y.-H. Hu, M.-H. Tu, J.-K. Zhao, Y.-H. Chu, S.-J. Jou, and C.-T. Chuang, "40 nm bit-interleaving 12T subthreshold SRAM with data-aware write-assist," *IEEE Transactions on Circuits and Systems-I*, vol. 61, pp. 2578 – 688, September 2014.

[5] B. Wang, J. Zhou, and T. T. Kim, "Ultra-low power 12T dual port SRAM for hardware accelerators," in *International SoC Design Conference (ISOCC)*, pp. 274 – 275, November 2014.

[6] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, April 1965.

[7] A. Khakifirooze and D. A. Antoniadis, "The future of high-performance CMOS: Trends and requirements," pp. 30 – 37, September 2003.

[8] H. Yamauchi, "Embedded SRAM trend in nano-scale CMOS," pp. 19 – 22, December 2007.

[9] K. Zhang, F. Seigneret, H. Yamauchi, H. Pilo, H. Shiral, and M. Hatanaka, "Embedded memory design for nano-scale VLSI systems," February 2008.

[10] E. J. Marinissen, B. Prince, D. Keltel-Schulz, and Y. Zorian, "Challenges in embedded memory design and test," pp. 722 – 727, March 2005.

[11] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. USA: Addison-Wesley Publishing, 4th ed., 2010.

[12] S. Ataei and J. E. Stine, "A differential single-port 8T SRAM bitcell for variability tolerance and low voltage operation," in *International Green and Sustainable Comuting Conference (IGSC)*, pp. 1 – 6, December 2015.

[13] S. Okumura, Y. Iguchi, S. Yoshimoto, H. Fujiwara, H. Noguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto, "A 0.56v 128-kb 10T SRAM using column line assist (CLA) scheme," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 659 – 663, March 2009.

[14] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy, "A 32 Kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 650 – 658, February 2009.

[15] S. Ataei, J. E. Stine, and M. R. Guthaus, "A 64 kb differential single-port 12T SRAM design with a bit-interleaving scheme for low-voltage operation in 32 nm SOI CMOS," in *IEEE International Conference on Computer Design (ICCD)*, pp. 1 – 8, October 2016.

[16] J. T. Pawlowski, "Synchronous SRAM having global write enable." `https://www.google.com/patents/US6205514`, March 2001.

[17] J. T. Pawlowski, "Synchronous SRAMs having multiple chip select inputs and a standby chip enable input." `https://www.google.com/patents/US6009494`, December 1999.

[18] H. Sato and S. Ohbayashi, "Synchronous semiconductor memory device operable in a snooze mode." `https://www.google.com/patents/US5602798`, February 1997.

[19] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, pp. 256 – 268, October 1974.

[20] M. H. A. Rahma and M. Anis, *Nanometer Variation-Tolerant SRAM: Circuits and Statistical Design for Yield*. Springer, 2013.

[21] K. Kuhn, "CMOS transistor scaling past 32nm and implications on variation," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pp. 241 – 246, July 2010.

[22] T. chiang Chen, "Where CMOS is going: trendy hype vs. real technology," in *EEE International Solid-State Circuits Conference (ISSCC)*, pp. 1 – 18, February 2006.

[23] T. Mizuno, J. ichi Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2216 – 2221, November 1994.

[24] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 748 – 754, October 1987.

[25] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read stability and write-ability analysis of SRAM cells for nanometer technologies," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 2577 – 2588, November 2006.

[26] C. Wann, R. Wong, D. J. Frank, R. Mann, S. B. Ko, P. Croce, D. Lea, D. Hoyniak, Y. M. Lee, J. Toomey, M. Weybright, and J. Sudijono, "SRAM cell design for stability methodology," in *International Symposium on VLSI Technology (VLSI-TSA)*, pp. 21 – 22, April 2005.

[27] J. Wang, S. Nalam, and B. H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs," in *International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 129 – 134, August 2008.

[28] R. W. Brodersen, *Anatomy of a Silicon Compiler*. Springer, 1992.

[29] D. Johannsen, "Bristle blocks: A silicon compiler," in *ACM/IEEE Design Automation Conference (DAC)*, pp. 310 – 313, June 1979.

[30] A. Cabe, Z. Qi, W. Huang, Y. Zhang, M. Stan, and G. Rose, "A flexible, technology adaptive memory generation tool," *Cadence CDNLive*, 2006.

[31] T.-H. Huang, C.-M. Liu, and C.-W. Jen, "A high-level synthesizer for VLSI array architectures dedicated to digital signal processing," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1221–1224, May 1991.

[32] P. Poechmueller, G. K. Sharma, and M. Glesner, "A CAD tool for designing large, fault-tolerant VLSI arrays," in *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 132 – 137, March 1991.

[33] Y. Xu, Z. Gao, and X. He, "A flexible embedded SRAM IP compiler," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3756 – 3759, May 2007.

[34] Global Foundries, "Memory IP." `http://www.globalfoundries.com/design/memory_ip.aspx`, 2015.

[35] Virage Logic, "SiWare memory." `http://www.viragelogic.com`, 2015.

[36] Faraday Technologies, "Memory compiler architecture." `http://www.faraday-tech.com/html/Product/IPProduct/LibraryMemoryCompiler/index.htm`, 2015.

[37] Dolphin Technology, "Memory products." `http://www.dolphin-ic.com/memory-products.html`, 2015.

[38] T. Shah, "FabMem: A multiported RAM and CAM compiler for superscalar design space exploration," Master's thesis, North Carolina State University, 2010.

[39] R. Goldman, K. Bartleson, T. Wood, V. Melikyan, and E. Babayan, "Synopsys' educational generic memory compiler," in *European Workshop on Microelectronics Education (EWME)*, pp. 89 – 92, May 2014.

[40] C. Ming and B. Na, "An efficient and flexible embedded memory IP compiler," in *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp. 268 – 273, October 2012.

[41] S. Wu, X. Zheng, Z. Gao, and X. He, "A 65nm embedded low power SRAM compiler," in *International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, pp. 123 – 124, April 2010.

[42] Micron, "4 Mb ZBT SRAM databook." `http://www.datasheetarchive.com/dl/Datasheets-IS23/DSA00446412.pdf`, 2000.

[43] B. Amrutur and M. Horowitz, "A replica technique for wordline and sense control in low-power SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 1208 – 1219, August 1998.

[44] J. Stine, I. Castellanos, M. Wood, J. Henson, F. Love, W. Davis, P. Franzon, M. Bucher, S. Basavarajaiah, J. Oh, and R. Jenkal, "FreePDK: An open-source variation-aware design kit," in *IEEE International Conference on Microelectronic Systems Education (MSE)*, pp. 173 – 174, June 2007.

[45] MOSIS, "MOSIS scalable CMOS (SCMOS)." `https://www.mosis.com/files/scmos/scmos.pdf`, 2015.

[46] M. Wieckowski, *GDS Mill User Manual*, 2010.

[47] N. Shibata, H. Morimura, and M. Watanabe, "A 1-V, 10-MHz, 3.5-mw, 1-Mb MTCMOS SRAM with charge-recycling input/output buffers," *IEEE Journal of Solid-State Circuits*, vol. 34, pp. 866 – 877, June 1999.

[48] K. Kushida, A. Suzuki, G. Fukano, A. Kawasumi, O. Hirabayashi, Y. Takeyama, T. Sasaki, A. Katayama, Y. Fujimura, and T. Yabe, "A 0.7v single-supply SRAM with 0.495 um^2 cell in 65nm technology utilizing self-write-back sense amplifier and cascaded bit line scheme," in *IEEE Symposium on VLSI Circuits*, pp. 46 – 47, June 2008.

[49] S. O. Toh, Z. Guo, T. K. Liu, and B. Nikolic, "Characterization of dynamic SRAM stability in 45 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 2702 – 2712, November 2011.

[50] S. Miyano, S. Moriwaki, Y. Yamamoto, A. Kawasumi, T. Suzuki, T. Sakurai, and H. Shinohara, "Highly energy-efficient SRAM with hierarchical bit line charge-sharing method using non-selected bit line charges," *IEEE Journal of Solid-State Circuits*, vol. 48, pp. 924 – 931, April 2013.

[51] K. Yamaguchi, H. Nambu, K. Kanetani, Y. Idei, N. Homma, T. Hiramoto, N. Tamba, K. Watanabe, M. Odaka, T. Ikeda, K. Ohhata, and Y. Sakurai, "A 1.5-ns access time, 78 um^2 memory-cell size, 64-kb ECL-CMOS SRAM," *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 167 – 174, February 1992.

[52] N. Tamba, A. Anzai, K. Akimoto, M. Ohayashi, T. Hiramoto, T. Kokubu, S. Ohmori, T. Muraya, A. Kishimoto, S. Tsuji, H. Hayashi, N. Handa, T. Igarashi, H. Nambu, M. Yoshida, T. Fujiwara, K. Watanabe, A. Uchida, M. Odaka, K. Yamaguchi, and T. Ikeda, "A 1.5-ns 256-kb BiCMOS SRAM with 60-ps 11-k logic gates," *IEEE Journal of Solid-State Circuits*, vol. 48, pp. 1344 – 1352, November 1994.

[53] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. K. Liu, and B. Nikolic, "Large-scale SRAM variability characterization in 45 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 3174 – 3192, November 2009.

[54] H. Yamauchi, "A discussion on SRAM circuit design trend in deeper nanometer-scale technologies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, pp. 763 – 774, June 2009.

[55] K. Nii, Y. Tsukamoto, M. Yabuuchi, Y. Masuda, S. Imaoka, K. Usui, S. Ohbayashi, H. Makino, and H. Shinohara, "Synchronous ultra-high-density 2RW dual-port 8T-SRAM with circumvention of simultaneous common-row-access," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 977 – 986, March 2009.

[56] S. Ishikura, M. Kurumada, T. Terano, Y. Yamagami, N. Kotani, K. Satomi, K. Nii, M. Yabuuchi, Y. Tsukamoto, S. O. andToshiyuki Oashi, H. Makino, H. Shinohara, and H. Akamatsu, "A 45 nm 2-port 8T-SRAM using hierarchical replica bitline technique with immunity from simultaneous R/W access issues," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 938 – 944, October 2008.

[57] J.-J. Wu, Y.-H. Chen, M.-F. Chang, P.-W. Chou, C.-Y. Chen, H.-J. Liao, M.-B. Chen, Y.-H. Chu, W.-C. Wu, and H. Yamauchi, "A large σ VTH /VDD tolerant zigzag 8T SRAM with area-efficient decoupled differential sensing and fast write-back scheme," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 815 – 827, April 2011.

[58] M.-F. Chang, J.-J. Wu, K.-T. Chen, and H. Yamauchi, "A differential data aware power-supplied (D^2AP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin applications," *International SoC Design Conference (ISOCC)*, vol. 45, pp. 1234 – 1245, June 2010.

[59] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, Y. Nakase, and H. Shinohara, "A 45nm 0.6V cross-point 8T SRAM with negative biased read/write assist," in *Symposium on VLSI Circuits*, pp. 158 – 159, November 2009.

[60] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 680 – 688, March 2007.

[61] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV robust schmitt trigger based subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 2303 – 2313, October 2007.

[62] C.-H. Lo and S.-Y. Huang, "P-P-N based 10T SRAM cell for low-leakage and resilient subthreshold operation," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 695 – 704, March 2011.

[63] Y. Ishii, Y. Tsukamoto, K. Nii, H. Fujiwara, M. Yabuuchi, K. Tanaka, S. Tanaka, and Y. Shimazaki, "A 28 nm 360ps-access-time two-port SRAM with a time-sharing scheme to circumvent read disturbs," in *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 236 – 238, February 2012.

[64] M. R. Guthaus, J. E. Stine, S. Ataei, B. Chen, B. Wu, and M. Sarwar, "OpenRAM: An open-source memory compiler," in *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, November 2016.

[65] J. P. Kulkarni, A. Goel, P. Ndai, and K. Roy, "A read-disturb-free, differential sensing 1R/1W port, 8T bitcell array," *IEEE Transaction on Circuits and Systems-I*, vol. 19, pp. 1727 – 1730, September 2011.

[66] C. H. Kim and K. Roy, "Dynamic Vt SRAM: A leakage tolerant cache memory for low voltage microprocessors," in *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 251 – 254, August 2002.

[67] S. Heo, K. K. Barr, M. Hampton, and K. Asanovic, "Dynamic fine-grain leakage reduction using leakage-biased bitlines," in *ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pp. 137 – 147, May 2002.

[68] K. Itoh, A. R. Fridi, A. Bellaouar, and M. I. Elmasry, "A deep sub-V single power supply SRAM cell with multi-v_T, boosted storage node and dynamic load," in *Proceeding of symposium of VLSI Circuits*, pp. 132 – 133, June 1996.

[69] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa, and K. Osada, "A 300-MHz 25-uA/Mb-leakage on-chip SRAM module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 186 – 194, January 2005.

[70] H. Zhu and V. Kursun, "Symmetrical triple-threshold-voltage nine-transistor SRAM circuit with superior noise immunity and overall electrical quality," in *International SoC Design Conference (ISOCC)*, pp. 333 – 336, November 2011.

[71] J. Lee and A. Davoodi, "Comparison of dual-Vt configurations of SRAM cell considering process-induced Vt variations," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3018 – 3021, May 2007.

[72] I. T. R. for Semiconductors, "2012 itrs report: System drivers," 2012.

[73] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 473 – 484, April 1992.

[74] H. Nambu, K. Kanetani, K. Yamasaki, K. Higeta, M. Usami, Y. Fujimura, K. Ando, T. Kusunoki, K. Yamaguchi, and N. Homma, "A 1.8-ns access, 550-MHz, 4.5-Mb CMOS SRAM," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 1650 – 1658, February 1998.

[75] D. Sekar and J. Meindl, "The impact of multi-core architectures on design of chip-level interconnect networks," in *IEEE International Interconnect Technology Conference (IITC)*, pp. 123 – 125, June 2007.

[76] T. Mizuno, J. Okumtura, and A. Toriumi, "Experimental study of threshold Voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2216 – 2221, November 1994.

[77] R. Houle, "Simple statistical analysis techniques to determine optimum sense amp set times," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 1816 – 1825, August 2008.

[78] V. Sharma, S. Cosemans, M. Ashouei, J. Huisken, F. Catthoor, and W. Dehaene, "A 4.4 pJ/access 80 mhz, 128 kbit variability resilient SRAM with multi-sized sense amplifier redundancy," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 2416 – 2430, October 2011.

[79] N. Verma and A. Chandrakasan, "A 65nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 328 – 606, February 2007.

[80] M. H. Abu-Rahma, K. Chowdhury, J. Wang, Z. Chen, S. S. Yoon, and M. Anis, "A methodology for statistical estimation of read access yield in SRAMs," in *ACM/IEEE Design Automation Conference (DAC)*, pp. 205 – 210, June 2008.

[81] A. Neale and M. Sachdev, "Digitally programmable SRAM timing for nano-scale technologies," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 1 – 7, March 2011.

[82] U. Arslan, M. McCartney, M. Bhargava, X. Li, K. Mai, and L. Pileggi, "Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines," in *IEEE Custom Integrated Circuits Conference (CICC)*, pp. 415 – 418, September 2008.

[83] S. Komatsu, M. Yamaoka, M. Morimoto, N. Maeda, Y. Shimazaki, and K. Osada, "A 40-nm low-power SRAM with multi-stage replica-bitline technique for reducing timing variation," in *IEEE Custom Integrated Circuits Conference (CICC)*, pp. 701 – 704, September 2009.

[84] Y. Niki, A. Kawasumi, A. Suzuki, Y. Takeyama, O. Hirabayashi, K. Kushida, F. Tachibana, Y. Fujimura, and T. Yabe, "A digitized replica bitline delay technique for random-variation-tolerant timing generation of SRAM sense amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 2545 – 2551, November 2011.

[85] K. Osada, J.-U. Shin, M. Khan, Y.-D. Liou, K. Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi, "Universal-vdd 0.65-2.0v 32 kb cache using voltage-adapted timing-generation scheme and a lithographical-symmetric cell," in *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 168 – 169, February 2001.

[86] J. Wu, J. Zhu, Y. Xia, and N. Bai, "A multiple-stage parallel replica-bitline delay addition technique for reducing timing variation of SRAM sense amplifiers," *IEEE Transactions on Circuits and Systems-II*, vol. 61, pp. 264 – 268, April 2014.

[87] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "An analytical model for negative bias temperature instability," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 493 – 496, November 2006.

[88] A. Kawasumi, Y. Takeyama, O. Hirabayashi, K. Kushida, F. Tachibana, Y. Niki, S. Sasaki, and T. Yabe, "A 47% access time reduction with a worst-case timing-generation scheme utilizing a statistical method for ultra low voltage SRAMs," in *IEEE Symposium on VLSI Circuits*, June 2012.

[89] Y.-C. Lai and S.-Y. Huang, "Robust SRAM design via bist-assisted timing-tracking (BATT)," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 642 – 649, February 2009.

[90] J. Mathews and R. L. Walker, *Mathematical Methods of Physics*. USA: Addison-Wesley Publishing, 2nd ed., 1970.

[91] B.-D. Yang and L.-S. Kim, "A low-power SRAM using hierarchical bit line and local sense amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1366 – 1376, June 2005.

[92] R. Joshi, R. Houle, K. Batson, D. Rodko, P. Patel, W. Huott, R. Franch, Y. Chan, D. Plass, S. Wilson, and P. Wang, "6.6+ GHz low Vmin, read and half select disturb-free 1.2 Mb SRAM," in *IEEE Symposium on VLSI Circuits*, pp. 250 – 251, June 2007.

[93] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "A 0.2 V, 480 kb subthreshold SRAM with 1 k cells per bitline for ultra-low-voltage computing," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 518 – 529, February 2008.

[94] K. Osada, K. Yamaguchi, Y. Saitoh, and T. Kawahara, "SRAM immunity to cosmic-ray-induced multierrors based on analysis of an induced parasitic bipolar effect," *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 827 – 833, May 2004.

[95] D. Anh-Tuan, J. Yung Shern Low, J. Yung Lih Low, Z. Kong, X. Tan, and K. Yeo, "An 8T differential SRAM with improved noise margin for bit-interleaving in 65 nm CMOS," *IEEE Transaction on Circuits and Systems-I*, vol. 58, pp. 1252 – 1263, June 2011.

[96] I. J. Chang, D. Mohapatra, and K. Roy, "A priority-based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications," *IEEE Transaction on Circuits and Systems for video technology*, vol. 21, pp. 101 – 112, February 2011.

[97] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Transaction on Circuits and Systems-I*, vol. 59, pp. 3 – 29, January 2012.

[98] S. Ataei and J. E. Stine, "A 64 kb multi-threshold SRAM array with novel differential 8T bitcell in 32 nm SOI CMOS technology," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1 – 6, October 2016.

[99] D. Mohapatra, G. Karakonstantis, and K. Roy, "Significance driven computation: A voltage-scalable, variation-aware, quality-tuning motion estimator," in *ACM/IEEE international symposium on Low power electronics and design (ISLPED)*, pp. 195 – 200, August 2009.

[100] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," in *IEEE European Test Symposium (ETS)*, pp. 1 – 6, May 2013.

[101] V. Chippa, A. Raghunathan, K. Roy, and S. Chakradhar, "Dynamic effort scaling: Managing the quality-efficiency tradeoff," in *ACM/IEEE Design Automation Conference (DAC)*, pp. 603 – 608, June 2011.

[102] F. Frustaci, D. Blaauw, D. Sylvester, and M. Alioto, "Approximate SRAMs with dynamic energy-quality management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, pp. 2128 – 2141, June 2016.

[103] N. Gong, S. Jiang, A. Challapalli, S. Fernandes, and R. Sridhar, "Ultra-low voltage split-data-aware embedded SRAM for mobile video applications," *IEEE Transaction on Circuits and Systems-II*, vol. 59, pp. 883 – 887, December 2012.

[104] J. Kwon, I. J. Chang, I. Lee, H. Park, and J. Park, "Heterogeneous SRAM cell sizing for low-power H.264 applications," *IEEE Transaction on Circuits and Systems-I*, vol. 59, pp. 2275 – 2284, October 2012.

[105] K. Yi, S.-Y. Cheng, F. Kurdahi, and A. Eltawil, "A partial memory protection scheme for higher effective yield of embedded memory for video data," in *International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pp. 1 – 6, August 2008.

[106] M. Cho, J. Schlessman, W. Wolf, and S. Mukhopadhyay, "Reconfigurable SRAM architecture with spatial voltage scaling for low power mobile multimedia applications," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, pp. 161 – 165, January 2011.

[107] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, and M. Alioto, "SRAM for error-tolerant applications with dynamic energy-quality management in 28 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 50, pp. 1310 – 1323, May 2015.

[108] F. Frustaci, D. Blaauw, D. Sylvester, and M. Alioto, "Better-than-voltage scaling energy reduction in approximate SRAMs via bit dropping and bit reuse," in *International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pp. 132 – 139, September 2015.

[109] I. Lee, J. Kwon, J. Park, and J. Park, "Priority based error correction code (ECC) for the embedded SRAM memories in H.264 system," *Journal of Signal Processing Systems*, vol. 73, pp. 123 – 136, March 2013.

[110] M. E. Sinangil and A. P. Chandrakasan, "An SRAM using output prediction to reduce BL-switching activity and statistically-gated SA for up to 1.9x reduction in energy/access," in *International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pp. 318 – 319, February 2013.

[111] H. Fujiwara, K. Nii, H. Noguchi, J. Miyakoshi, Y. Murachi, Y. Morita, H. Kawaguchi, and M. Yoshimoto, "Novel video memory reduces 45% of bitline power using majority logic and data-bit reordering," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, pp. 620 – 627, June 2008.

[112] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, and S. Iwade, "A 90-nm low-power 32-kB embedded SRAM with gate leakage suppression circuit for mobile applications," *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 684 – 693, April 2004.

[113] K. Kanda, T. Miyazaki, M. K. Sik, H. Kawaguchi, and T. Sakurai, "Two orders of magnitude leakage power reduction of low voltage SRAM's by row-by-row dynamic VDD control (RRVD) scheme," pp. 381 – 385, September 2002.

[114] Y. Tsividis and C. McAndrew, *Operation and modeling of the MOS transistor, 3rd edition*. Oxford series in electrical and computer engineering, New York, NY: Oxford University Press, 2011.

[115] M. Powell, S.-H. Yang, B. Falsafi, K. Roy, and T. N. Vijaykumar, "Gated-Vdd: A circuit technique to reduce leakage in deep-submicron cache memories," pp. 90 – 95, July 2000.

[116] H. Kawaguchi, Y. Itaka, and T. Sakurai, "Dynamic leakage cut-off scheme for low-voltage SRAM's," in *Symposium on VLSI Circuits*, pp. 140 – 141, June 1998.

[117] A. Agarwal, H. Li, and K. Roy, "A single-Vt low-leakage gated-ground cache for deep submicron," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 319 – 328, February 2003.

[118] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: Simple techniques for reducing leakage power," pp. 148 – 157, May 2002.

[119] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," pp. 55 – 60, March 2004.

[120] T. Wiegand, G. Sullivan, and A. Luthra, "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T rec. H.264 — ISO/IEC 14496-10 AVC)," Tech. Rep. 14496-10, ISO/IEC, Geneva, Switzerland, May 2002. Available: `http://www.hlevkin.com/Standards/h264.pdf`.

[121] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Transaction on Circuits and Systems-I*, vol. 54, pp. 660 – 668, September 2008.

[122] "Video Test Sequence Database." `http://sipi.usc.edu/database/?volume=misc`.

[123] T.-H. Kim, J. Keane, H. Eom, and C. H. Kim, "Utilizing reverse short-channel effect for optimal subthreshold circuit design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, pp. 821 – 829, July 2007.

[124] R. R. Troutman, "VLSI limitations from drain-induced-barrier-lowering," *IEEE Transactions on Electron Devices*, vol. 26, pp. 461 – 469, April 1979.

[125] C. Subramanian, J. Hayden, W. Taylor, M. Orlowski, and T. McNelly, "Reverse short channel effect and channel length dependence of boron penetration in PMOSFETs," in *International Electron Devices Meeting (IEDM)*, pp. 423 – 426, December 1995.

[126] C. Y. Lu and J. M. Sung, "Reverse short channel effects on threshold voltage in submicrometer salicide devices," *IEEE Electron Device Letters*, vol. 10, pp. 446 – 448, October 1989.

VITA

Fatemeh Ataei

Candidate for the Degree of

Doctor of Philosophy

Dissertation:

LOW-POWER AND HIGH-PERFORMANCE SRAM DESIGN IN HIGH VARIABILITY ADVANCED CMOS TECHNOLOGY

Major Field:

Electrical Engineering

Biographical:

Personal Data: Born in Isfahan, IRAN on December 12, 1984.

Education:

Received the B.S. degree from Shahid Beheshti University, Tehran, Iran, 2007, in Electrical Engineering

Received the M.S. degree from Amirkabir University, Tehran, Iran, 2011, in Electrical Engineering

Completed the requirements for the degree of Doctor of Philosophy with a major in Electrical Engineering from Oklahoma State University in May, 2017.

Experience:

Research Assistant in Oklahoma State University, January 2013 - May 2017

Developed OpenRAM; An Educational, Portable, Flexible and Open-Source Memory Compiler

Proposed Low-power and high-performance SRAM Architectures

Mentored master students in VLSI Computer Architecture Research Group at OSU

Name: Fatemeh Ataei **Date of Degree:** May, 2017

Institution: Oklahoma State University **Location:** Stillwater, Oklahoma

Title of Study: LOW-POWER AND HIGH-PERFORMANCE SRAM DESIGN IN HIGH VARIABILITY ADVANCED CMOS TECHNOLOGY

Pages in Study: 125 **Candidate for the Degree of Doctor of Philosophy**

Major Field: Electrical Engineering

Abstract: As process technologies shrink, the size and number of memories on a chip are exponentially increasing. Embedded SRAMs are a critical component in modern digital systems, and they strongly impact the overall power, performance, and area. To promote memory-related research in academia, this dissertation introduces OpenRAM, a flexible, portable and open-source memory compiler and characterization methodology for generating and verifying memory designs across different technologies.

In addition, SRAM designs, focusing on improving power consumption, access time and bitcell stability are explored in high variability advanced CMOS technologies. To have a stable read/write operation for SRAM in high variability process nodes, a differential-ended single-port 8T bitcell is proposed that improves the read noise margin, write noise margin and readout bitcell current by 45%, 48% and 21%, respectively, compared to a conventional 6T bitcell. Also, a differential-ended single-port 12T bitcell for subthreshold operation is proposed that solves the half-select disturbance and allows efficient bit-interleaving. 12T bitcell has a leakage control mechanism which helps to reduce the power consumption and provides operation down to 0.3 V. Both 8T and 12T bitcells are analyzed in a 64 kb SRAM array using 32 nm technology. Besides, to further improve the access time and power consumption, two tracking circuits (multi replica bitline delay and reconfigurable replica bitline delay techniques) are proposed to aid the generation of accurate and optimum sense amplifier set time.

An error tolerant SRAM architecture suitable for low voltage video application with dynamic power-quality management is also proposed in this dissertation. This memory uses three power supplies to improve the SRAM stability in low voltages. The proposed triple-supply approach achieves 63% improvement in image quality and 69% reduction in power consumption compared to a single-supply 64 kb SRAM array at 0.70 V.

ADVISOR'S APPROVAL: Professor James E. Stine