Digital Timing Control in SRAMs for Yield Enhancement and Graceful Aging Degradation

by

Adam Neale

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

Waterloo, Ontario, Canada, 2010

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

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

Adam Neale

Abstract

Embedded SRAMs can occupy the majority of the chip area in SOCs. The increase in process variation and aging degradation due to technology scaling can severely compromise the integrity of SRAM memory cells, hence resulting in cell failures. Enough cell failures in a memory can lead to it being rejected during initial testing, and hence decrease the manufacturing yield. Or, as a result of long-term applied stress, lead to in-field system failures. Certain types of cell failures can be mitigated through improved timing control. Post-fabrication programmable timing can allow for after-the-fact calibration of timing signals on a per die basis. This allows for a SRAM's timing signals to be generated based on the characteristics specific to the individual chip, thus allowing for an increase in yield and reduction in in-field system failures.

In this thesis, a delay line based SRAM timing block with digitally programmable timing signals has been implemented in a 180 nm CMOS technology. Various timing-related cell failure mechanisms including: 1). Operational Read Failures, 2). Cell Stability Failures, and 3). Power Envelope Failures are investigated. Additionally, the major contributing factors for process variation and device aging degradation are discussed in the context of SRAMs. Simulations show that programmable timing can be used to reduce cell failure rates by over 50%.

Acknowledgements

I would like to take this oppourtunity to express my gratitude and thanks to my supervisor **Professor Manoj Sachdev** at the University of Waterloo. Without his guidance, knowledge, kindness, (and patience), this work would not have been possible. I would also like to thank **Dr. Bill Bishop**, and **Dr. Andrew Kennings**. Thank you for your valuable comments and suggestions on my thesis.

It has been a great pleasure to work as a part of the CMOS Design and Reliability (CDR) Group. The dedication and talent within this group has always inspired me to achieve more than I could have ever imagined during my M.A.Sc. experience. My sincere appreciation goes out to all of the past and current members of this group. I am extremely grateful for the immense support of **Dr. David Rennie** and **Tahseen Shakir**. And, for the bond shared between **Pierce Chuang**, **David Li**, **Jaspal Singh Shah**, and myself, as we spent many long days and late nights designing and laying out test chips together in confined quarters.

I would also like to thank **Dr. Bill Bishop** for being a mentor for me since my first year of undergrad, and inspiring me to go into graduate studies at the University of Waterloo. Over the years I've had the chance to get to know Dr. Bishop as my lecturer, co-op supervisor, senior design project consultant, teaching assistant supervisor, and ultimately as my friend.

Lastly, I would like to acknowledge the funding that I received through out the duration of my M.A.Sc. degree from the **Ontario Graduate Scholarship** (OGS), the **Natural Sciences and Engineering Research Council** (NSERC), and the **University of Waterloo**. Thank you for believing in my ability to do research.

To my friends and family.

“To give anything less than your best is to sacrifice the gift.”

- Steve Prefontaine

Contents

List of Tables **ix**

List of Figures **xii**

1 Introduction **1**

1.1 Research Contributions 4

1.2 Thesis Organization 4

2 SRAM Design & Operation **6**

2.1 High-Level SRAM Operation 6

2.2 Operation of the 6T SRAM Cell 9

2.3 Peripheral Circuitry 12

2.3.1 Row & Column Address Decoders 12

2.3.2 Precharge & Equalization Circuitry 13

2.3.3 Write Driver 13

2.3.4 Sense Amplifier 14

2.4 Modern Timing Control Schemes 17

2.5 Programmable Delay Calibration 20

2.6 Figures of Merit 21

2.6.1 Area 21

2.6.2 Current Leakage 21

2.6.3 Static Noise Margin 22

2.6.4 Dynamic Noise Margin 24

3 Process Variability & Aging Degradation Mechanisms 26

3.1 Mechanisms for Transistor Variability 29

3.1.1 Random Dopant Fluctuation 29

3.1.2 Line Edge Roughness 30

3.2 SNM Variability in the 6T SRAM Cell 31

3.3 Aging Mechanism 36

3.3.1 Gate-Oxide Breakdown 36

3.3.2 Hot Carrier Injection 37

3.3.3 Bias Temperature Instability 39

3.4 Aging in SRAM 42

4 SRAM Timing Failures 45

4.1 Operational Read Failure 45

4.2 Cell Stability Failure 47

4.3 Power Envelope Failure 50

4.4 Timing Related Cell Failure Reduction 52

5 Flexible SRAM Timing Control Architecture 54

5.1 Delay Line 55

5.2 Digitally Controlled Delay Element 57

5.3 Extended Range Delay Element 60

5.4 Timing Block 61

6 Simulation Results & Test Chip 63

7 Conclusion 70

Bibliography 71

List of Tables

3.1 Standard Deviation Across Multiple σ and Defects per Million 29

5.1 Binary and Thermometer Code Example 59

6.1 Test Chip Characteristics 64

6.2 Simulated Performance Data Under Process and Temperature Variation . . 69

List of Figures

1.1 6T SRAM cell area as a function of CMOS technology scaling [35] 2

1.2 Hard and soft fail predications versus technology node [32] 3

2.1 SRAM High-level Block Diagram 7

2.2 SRAM with multiple words per row 8

2.3 Schematic of the 6T SRAM Cell 10

2.4 SRAM Read and Write Operations 12

2.5 Latch-Type Sense Amplifier 15

2.6 Control Signal Timing Schemes 18

2.7 Operation flow of a calibration controller during power-on self-test [21] . . 20

2.8 SNM measurement testbench and butterfly curves 23

2.9 Dynamic Noise Margin 24

3.1 Normal Distribution 28

3.2 V_{TH} variability as a function of channel area for both a 90 nm and a 65 nm process. The line is a guide to the eye and not necessarily a fit to the data [28]. 30

3.3 SRAM VTC curves under both ideal and non-ideal conditions due to transistor mismatch [13] 31

3.4 6T SRAM cell SNM deviation vs. threshold voltage deviation on one of the transistors [30] 33

3.5 SRAM cell SNM vs. threshold voltage deviation of more than one transistor [30] 34

3.6 SRAM cell SNM deviation vs. transistor Length (L), and Width (W) [30] . 35

3.7 A conductive path in the gate stack due to gate-oxide breakdown stress [18] 37

3.8 Hot carrier injection stress mechanism [18] 38

3.9 Conditions for negative and positive bias temperature instability stress . . 39

3.10 NBTI Stress and Recovery States [18] 40

3.11 ΔV_{TH} for PMOS devices under NBTI stress and recovery conditions [44] . 42

3.12 BTI susceptible transistors within the SRAM cell 43

3.13 ΔV_{TH} for BTI Stress in both SiO_2 and high-κ gate stacks [49] 44

4.1 Effect of process and voltage variations on required cell access time 46

4.2 Schematic of a Weak 6T SRAM Cell 48

4.3 Weakened 6T SRAM dynamic cell stability for variable cell access time at a reduced supply voltage, $V_{DD} = 0.7$ V 49

4.4 Measured DNM of a 6T SRAM cell [39] 49

4.5 Example timing configurations for both nominal and reduced supply voltage, V_{DD} 51

4.6 Cell Failure Reduction Using Programmable Timing 53

5.1 Variable wordline access time and sense amplifier enable windows 55

5.2 Pulse generator based variable delay line architecture 56

5.3 Pulse generator delay line timing diagram 56

5.4 Digitally Controlled Delay Element 58

5.5 A comparison between binary and thermometer digital control codes applied to the same DCDE that exhibits a monotonicity error 60

5.6 The extended range delay element uses a two stage delay element to select the delay, the first stage uses a two-bit binary code to select the coarse delay, and a four-bit thermometer code to select the fine delay 61

5.7 Programmable SRAM timing block 62

6.1 Test chip layout in 180 nm CMOS 64

6.2 Simulated SRAM Read Operation 66

6.3 Simulated Wordline Access Time Programmability 67

6.4 Monotonic WL to SAE propagation delays versus increasing control codes under process and temperature variation 68

Chapter 1

Introduction

A system-on-chip (SOC) is an integration of all the components for a computer or other electronic system into a single integrated circuit (IC). Embedded memories can occupy up to 70% of the total die area of modern SOCs [17]. As Complementary Metal Oxide Semiconductor (CMOS) technology scales deep into the sub-100 nm regime, the density of memory bitcells has significantly increased, resulting in larger embedded memories for the same die area. This allows for much more memory intensive applications to be performed on an SOC of a fixed area.

Due to its superior performance capabilities and compatibility with the CMOS logic process, the six transistor (6T) Static Random Access Memory (SRAM) has been adopted as the workhorse for many SOC embedded memories. The cell has scaled well with CMOS processes, and has even become a method for characterizing and comparing processes against one another. The general industry standard for the SRAM cell in terms of area scaling has been relatively constant at 0.5x / generation. This trend is shown in Figure 1.1. Shown in the inset is the layout for a state-of-the-art 0.171 μm^2 bitcell designed in a 32 nm process [46].

Since memories consume the vast majority of SOC die area, and are predominately

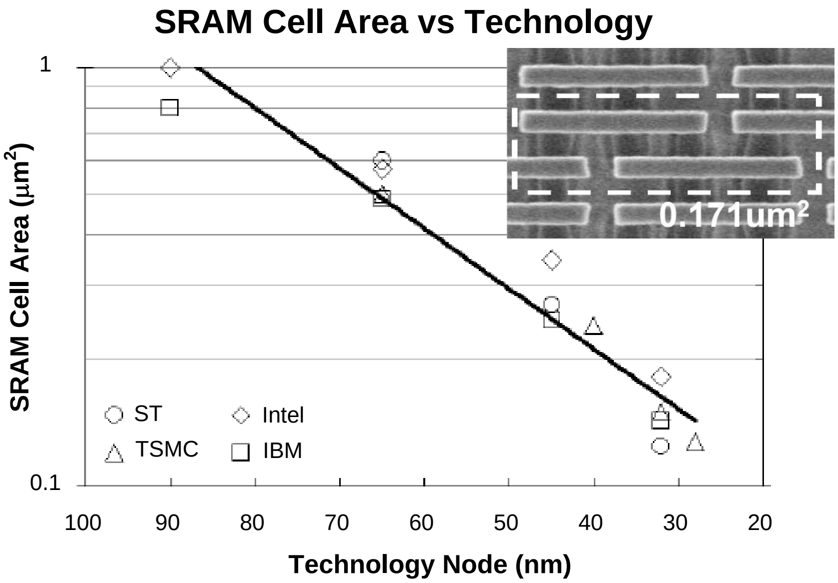


Figure 1.1: 6T SRAM cell area as a function of CMOS technology scaling [35]

comprised of minimum, or near minimum, sized transistors, proper functionality of the SOC is heavily influenced by the functional correctness of its memory array. As device dimensions continue to shrink however, memory cells become more susceptible to process variation and aging effects, and hence increased failure rates [31, 2, 29, 48, 19]. Additionally, for power saving purposes, circuits are typically operated at low voltages. Cell failure is significantly more noticeable when the device is operating at these lower voltages, particularly its minimum operating voltage, VDD_{MIN}. The main failure mechanisms include: inability to write to or read from the cell, signal or power margin failures, read stability failures, and retention failures [32].

Figure 1.2 shows the growing failure rate as a result of voltage and device scaling with shrinking technology nodes. The vertical axis shows the fail count per n-Mbits of an SRAM array and the horizontal axis shows four technology nodes. The figure shows that as CMOS technology scales, the amount of hard failures (unwanted open or short circuits) decrease

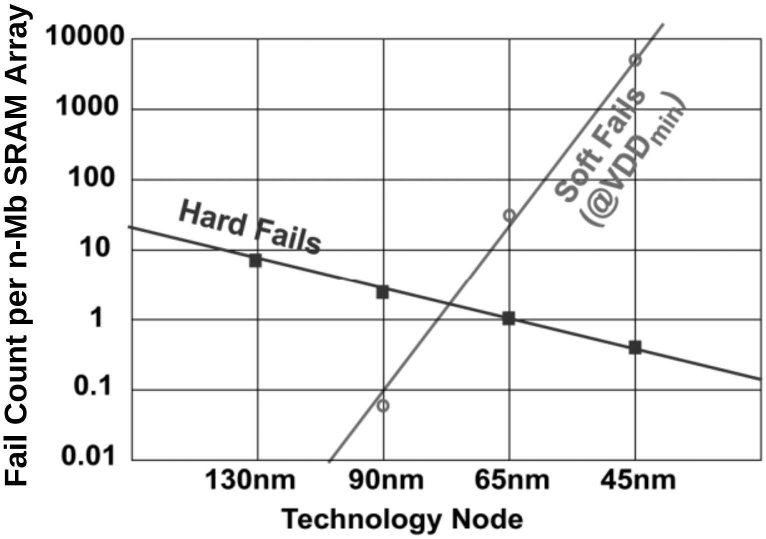


Figure 1.2: Hard and soft fail predications versus technology node [32]

within new processes; however, the amount of soft failures (those listed above) are on the rise. As a mitigation technique, large SRAM arrays often include redundant columns for replacing those that contain failing cells. This is an effective technique against defect based failures; however, the soft failure rate can exceed the maximum repair capacity of the SRAM, and lead to incorrect memory functionality. This in turn results in manufacturing yield loss.

Although the issues caused by technology scaling predominately stem from the device-level, it is in part the responsibility of the circuit designer to cope with these difficulties at the circuit-level. It has been shown that certain types of SRAM soft failures can be mitigated through improved timing control [3]. Various timing schemes have been implemented to better track activity inside the memory array to allow for tighter timing margins [32, 9, 13]. Additionally, the implementation of post-fabrication programmable timing has allowed for after-the-fact timing adjustment to reduce failure rates, and in turn maximize yield [6, 21].

Soft failures in SRAMs are not to be confused with soft errors. Soft errors are caused by external sources of radiation interacting with the silicon substrate leading to the corruption of stored data [5]. Where as, soft failures are caused by the weakening of particular memory cells due to variability in the manufacturing process and device aging.

1.1 Research Contributions

In this work, the impact of control signal timing on several SRAM figures of merit is investigated with the goal of reducing the soft failure rate, and in turn improving the overall manufacturing yield. It is also shown that post fabrication signal timing control can be used to aid in extending the lifetime of SRAMs by allowing for more graceful aging degradation. Additionally, a delay line based SRAM timing block with programmable timing signals has been implemented in a 180 nm bulk CMOS technology. The timing block has been designed to operate at a maximum frequency of 500 MHz, and is capable of full-speed operation while using a low-speed test clock. The cell access and sensing times (two of the most critical timing parameters) can each be varied by over 400 ps under typical operating conditions over a set of 20 digital control codes. Implementation was done at the 180 nm node due to its availability, low cost relative to other technology nodes, and the fact that the timing block was designed in isolation as a functional proof of concept rather than as a component of a full SRAM. Fabrication will be done at a later date.

1.2 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 provides an overview of the basic operation of SRAMs. Chapter 3 discusses process variation and aging mechanisms, and how they affect SRAMs. Chapter 4 discusses timing related failure mechanisms.

Chapter 5 describes the implementation of the programmable timing block. Chapter 6 provides simulation results, and describes the design of the test chip. And finally, Chapter 7 concludes the thesis.

Chapter 2

SRAM Design & Operation

A typical SRAM configuration consists of: an array of addressable storage cells, an address decoder for determining which set of cells to access for a particular address, peripheral circuitry for accessing the cells in the array, and a timing block for generating any necessary control signals. The 6T SRAM cell is currently the defacto standard data storage cell [35]. The following chapter provides a brief overview of its operation, and how it interfaces with the other components of the SRAM.

2.1 High-Level SRAM Operation

Figure 2.1 provides an example of the basic SRAM memory structure. The size of the memory is defined by the number of bits stored within the array. A bit is the elemental piece of binary data stored in a single memory cell. Cells, or bits, are organized into a set of N horizontal rows each containing M bits of data. Values for each of these are typically powers of two (e.g. 64, 128, 256, or 512) to maximize address space usage. The size of the array is then given by $N \times M$ bits. Each row of data is selected one at time by means of an address decoder. The address decoder takes a K bit address and uses it to select the

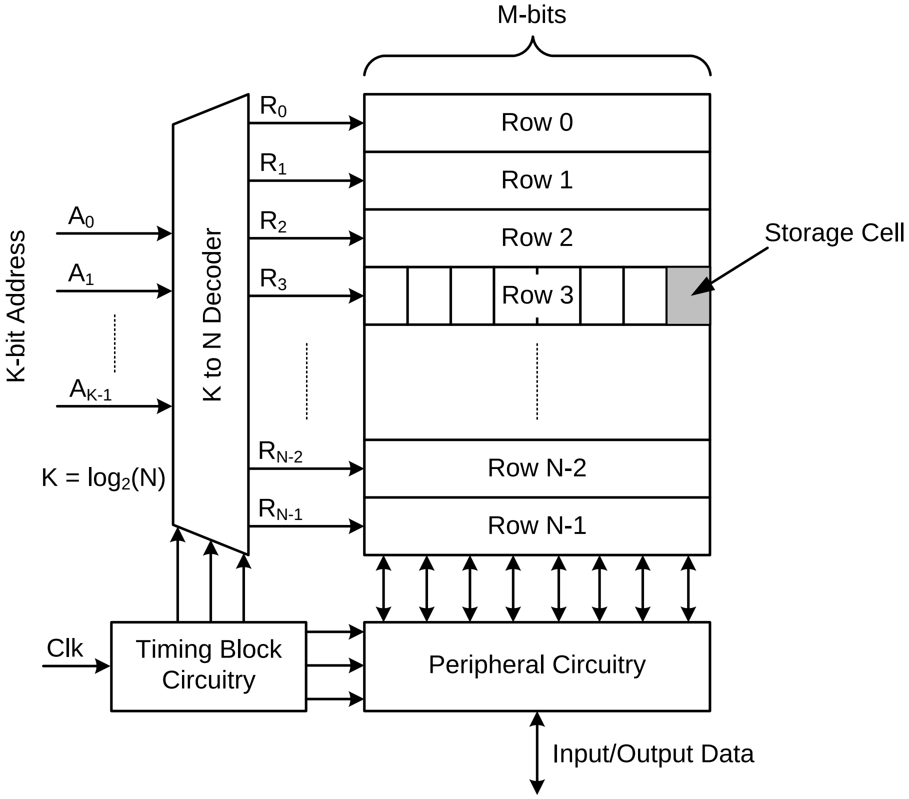


Figure 2.1: SRAM High-level Block Diagram

access control signal of one of the $N = 2^K$ horizontal rows. The row's access control signal is known as the wordline (WL). Once a row has been selected, it can be either read from or written to by the peripheral circuitry. Each column has a complementary set of bitlines (BL/BLB) for access into the selected row's storage cells.

Often times, each row will contain multiple words of data. A word consists of W bits and represents the logical data size for the SRAM. Having multiple words on a single row can lead to physically more compact designs since the SRAM can take on a more square shape. Furthermore, interleaving the bits of multiple words within a single row allows for

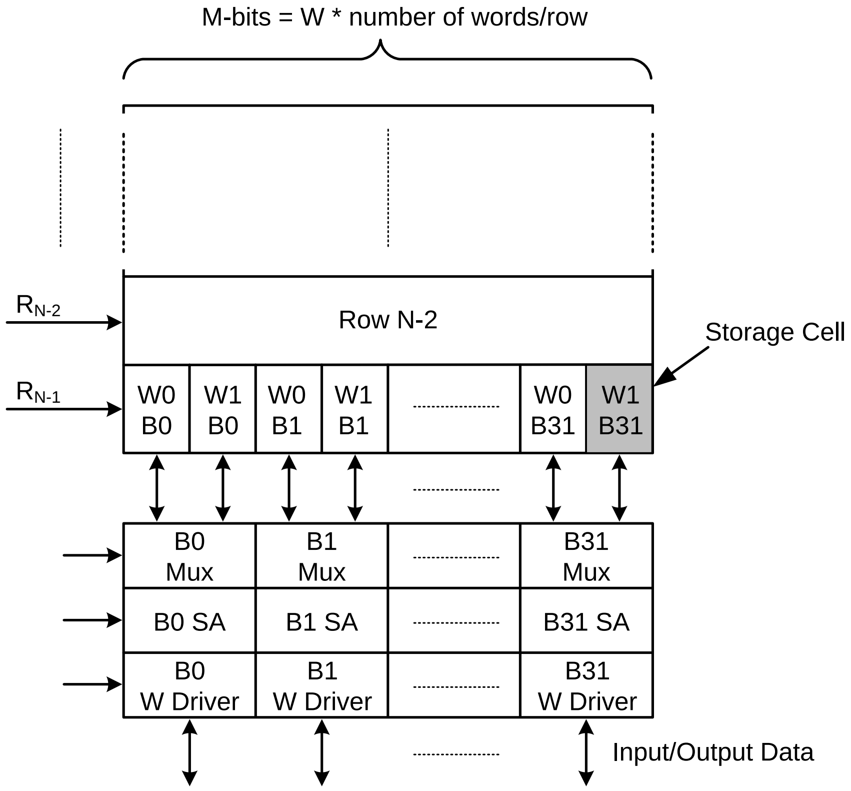


Figure 2.2: SRAM with multiple words per row

the sharing of peripheral circuitry across multiple columns of the array. This is shown in Figure 2.2 where two 32-bit words are interleaved within a single 64-bit row. This allows for a reduction in the amount of column peripheral circuitry by a factor of two. The figure shows the sharing of the bitline multiplexers, sense amplifiers, and write drivers.

Both of these optimizations can lead to lower-power, more dense, and potentially higher speed designs depending on the details of the particular SRAM implementation. They do come at the cost of more complex address decoding however, since a particular word must be selected from the row being accessed. Cells in a non-selected column of a selected row are known as half-selected cells. Half-selected cells can lead to data stability issues, and are discussed in Section 2.6.3 when considering the concept of read access static noise margin.

Both reading and writing operations require a sophisticated timing sequence. Most modern SRAMs are self-timed, meaning that all of their internal timing is generated by a timing block within the SRAM itself. The generation of each of these timing signals is critical to the successful operation of the SRAM. Any shortcomings in the generation of these signals can cripple an otherwise fully functional SRAM, hence rendering it unusable. Each of these components are described in more detail in the following sections.

2.2 Operation of the 6T SRAM Cell

Every memory cell consists of two essential components: a storage cell and a transfer gate. The storage cell holds the data and determines the ability of the circuit to withstand noise. The transfer gate allows data to be written into and read from the storage cell. Figure 2.3 shows the schematic of a 6T SRAM cell. The storage cell is composed of two back-to-back inverters (P1 and N1, P2 and N2). NMOS transistors N1 and N2 are known as the drive transistors, and PMOS transistors P1 and P2 are known as the load transistors. The transfer gate is formed by transistors N3 and N4. These are known as the access transistors. The 6T SRAM cell has three modes of operation: read, write, and retention.

Since an SRAM array contains many thousands (sometimes millions) of cells, and only one word can be accessed at a given time, a SRAM cell will typically be in the unaccessed retention mode for the vast majority of time. In this operating condition the wordline (WL) is turned off, isolating the complementary bitlines (BL/BLB) from the storage cell. Moreover, the bitlines are held at V_{DD}, minimizing leakage and maintaining the bitlines in a precharged state in preparation for a read or write operation.

To read from or write to the storage cell, it must first be accessed. The timing diagrams

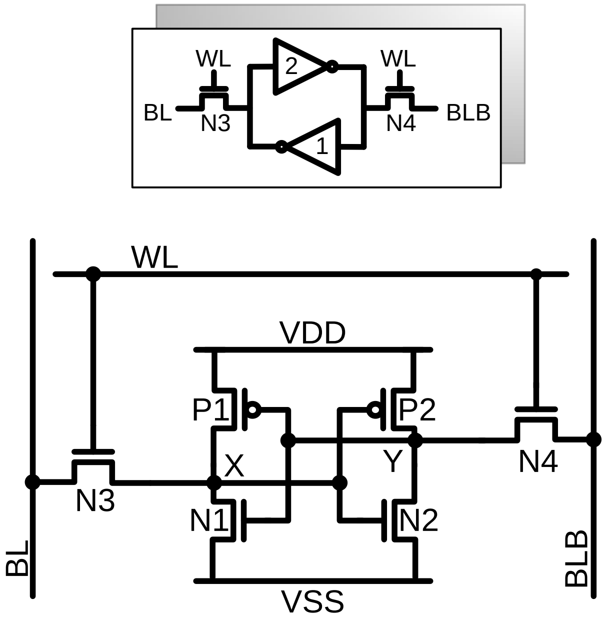


Figure 2.3: Schematic of the 6T SRAM Cell

for the read and write operations are shown in Figures 2.4(a) and 2.4(b) respectively. To access the storage cell, the precharge signal (PRE), not shown in Figure 2.3, is set to evaluate. This allows the bitlines to float at V_{DD}. The WL is then turned on. This connects the bitlines to the storage cell via the access transistors. For the read operation, since the bitlines are precharged high, one access transistor will have zero voltage across it while the other will be have a potential difference across it equal to V_{DD}. Current flows from the bitlines through the access transistor to the node that is storing a ‘0’, and down to ground through the drive transistor. In this way, one bitline will begin discharging, and can be read out as a ‘0’ by the peripheral circuitry. Without loss of generality, assuming that Node X in Figure 2.3 is initially ‘0’ (and hence Node Y is a ‘1’), the bitline BL will discharge through the access transistor N3 and drive transistor N1. At the same time BL is being discharged however, Node X will tend to rise due to the current flowing into the node via the access transistor. Hence, N1 must be stronger than N3 to prevent Node X from rising above the switching threshold of the P2/N2 inverter to prevent the cell from flipping.

This constraint determines the read stability of the cell. The voltage rise inside the cell depends upon the strength of the driver transistor relative that of the access transistor. This ratio is known as the cell ratio (CR), and is given by

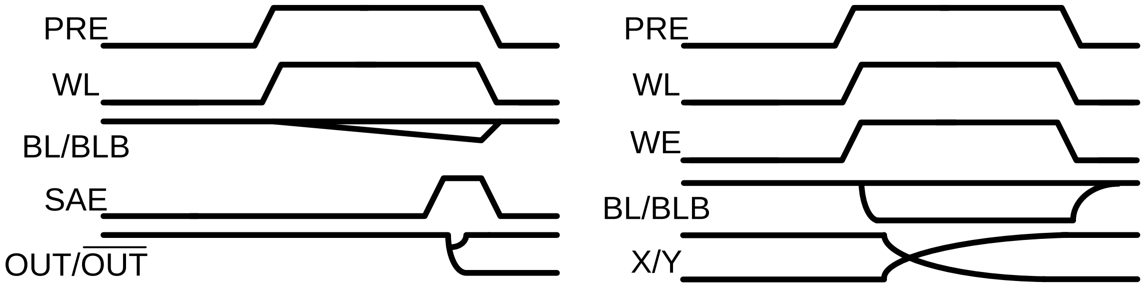
$$CR = \frac{W_{N1}/L_{N1}}{W_{N3}/L_{N3}} \tag{2.1}$$

where W_{N1}, L_{N1}, W_{N3}, and L_{N3} are the width and length of the driver and access transistors respectively. The CR should be greater than 1.2 to prevent the internal node voltage of the cell from rising above the threshold of the complementary inverter [33].

For a write operation, the bitlines are driven to complementary values by a write driver accessed via the write enable signal (WE). Due to read stability restrictions, and the fact that NMOS access transistors are not able pass V_{DD}, the write operation is not completely symmetric. The write operation essentially writes a '0' into one node of the storage cell by discharging the stored '1' value, and the internal feedback of the cell writes the other node. For example, if Node X is initially '0', and Node Y is a '1', then BLB will be pulled down to '0' to write into the cell. The load transistor P2 will oppose this operation. Hence, P2 must be weaker than the access transistor N4 so that BLB can be pulled low enough. This constraint determines the writeability of the cell. Once Node Y has been pulled low enough, N1 will turn off, P1 will turn on, and Node X will be pulled high. Once Node X is high, it will turn off P2, turn on N2, and hence latch the new data into the cell. The strength of load transistor relative to the access transistor is known as the pull-up ratio (PR), and is defined by

$$PR = \frac{W_{P2}/L_{P2}}{W_{N4}/L_{N4}} \tag{2.2}$$

where W_{P2}, L_{P2}, W_{N4}, and L_{N4} are the width and length of the load and access transistors respectively. The condition for a successful write operation can typically be performed



(a) Read Operation

(b) Write Operation

Figure 2.4: SRAM Read and Write Operations

using minimum sized load and access transistors for the given technology node. The intrinsic weakness of PMOS transistors relative to NMOS transistors will ensure the load transistor is weaker than the access transistor, and allow for writeability of the cell.

To ensure both read stability and writeability, the drive transistors (N1 & N2) must be strongest, access transistors (N3 & N4) of intermediate strength, and load transistors (P1 & P2) weak. Additionally, for high array densities, all the transistors must be close to minimum size for the given technology, and the SRAM cells must be designed to operate correctly under all process corners at all voltage and temperature variations.

2.3 Peripheral Circuitry

2.3.1 Row & Column Address Decoders

Row and column decoders are used within an SRAM to reduce the required number of select signals and additionally to reduce the capacitive load on the word- and bit-lines. The row decoder is able to reduce the number of select signals used to address the memory rows by $\log_2 N$, where N is the number of rows in the memory array. Column decoders are used to select a particular word from a multi-word row in the memory. This is typically done using a pass gate style multiplexer. The total number of addressing bits used to access a particular word in the memory can be divided into three separate segments. For instance, in one particular arrangement, the least significant bits are used for column select addressing, the middle bits for the row selection, and the most significant bits, if there are multiple memory arrays on the chip, for the page or bank addressing. Segmenting the addressing bits in this fashion aids in the facilitation of spatial locality when the SRAM is being used as a cache [14]. As an example, a 64-kbit array partitioned into two pages ($1 = \log_2(2)$), each containing 256 rows ($8 = \log_2(256)$) and four 32-bit words per row ($2 = \log_2(4)$) requires 11 address bits ($11 = 1 + 8 + 2$) to address each 32-bit word.

2.3.2 Precharge & Equalization Circuitry

To help reduce read and write cycle time, the precharge and equalization phase can be done while the address is being decoded. During this time, all the bitlines within the memory array are set to a predetermined voltage level, and each BL/BLB pair is equalized to help minimize any asymmetrical behaviour between the two as a result of device mismatch. Once the bitlines have been precharged and the address has been decoded, the bitlines are allowed to float. At this point, either a read or write operation may take place. Common precharge voltage levels include V_{DD}, $V_{DD}/2$, $V_{DD} - V_{TH}$, or ground. A common precharge and equalize circuit is illustrated in Figure 2.5.

2.3.3 Write Driver

When writing into the array, the write driver is responsible for quickly discharging one of the precharged bitlines below the write margin from each BL/BLB pair being used for writing. Considering Figure 2.3 as an example, in the event that a ‘0’ is being written into node X, the BL will be discharged. Contrarily, if a ‘0’ is being written into node Y, then the BLB will be discharged. Typically, the write driver will be activated by the write enable (WE) signal generated by the timing block.

2.3.4 Sense Amplifier

The read operation is typically the slowest memory operation, and as such defines the minimum delay of the SRAM cell [35]. Bitlines experience a large capacitance due to their physical metal length and large number of cell access transistors connected to them. As such, a significant amount of time is required for a bitline to fully discharge. Rather than waiting for this to occur on its own, a sense amplifier is used to detect a small differential voltage on the bitlines, and quickly generate a full-swing output. The timing control of the sense amplifier is critical for the correct functionality of the SRAM. If the sense amplifier enable signal (SAE) is enabled before a sufficient amount of differential voltage is generated, the output may resolve incorrectly. If the sense amplifier is turned on too late however, the read time will be longer than necessary and excessive power will be dissipated. Power dissipation during a read cycle is further discussed in Section 4.3.

There are many sense amplifier variants. Figure 2.5 shows an example of a latch-type sense amplifier implemented in a SRAM column. This particular implementation is based off a pair of cross-coupled inverters, similar to that of the 6T SRAM cell. The forward feedback action of the inverters is used to accelerate the discharging of one of the bitlines. Before reading can begin, precharge and equalization circuitry is used to bias and equalize the bitlines at V_{DD}, and put the inputs of the sense amplifier into a metastable region. Here, two separate sets of precharge and equalization circuitry is used (one for the bitcell column, and another for the sense amplifier). This is done so that the sense amplifier can be isolated from the bitlines (through the YMUX PMOSs), and full-swing can be generated on the sense amplifier while only a small differential voltage is developed on the bitlines. This saves the extra time and energy cost of fully discharging and then precharging the

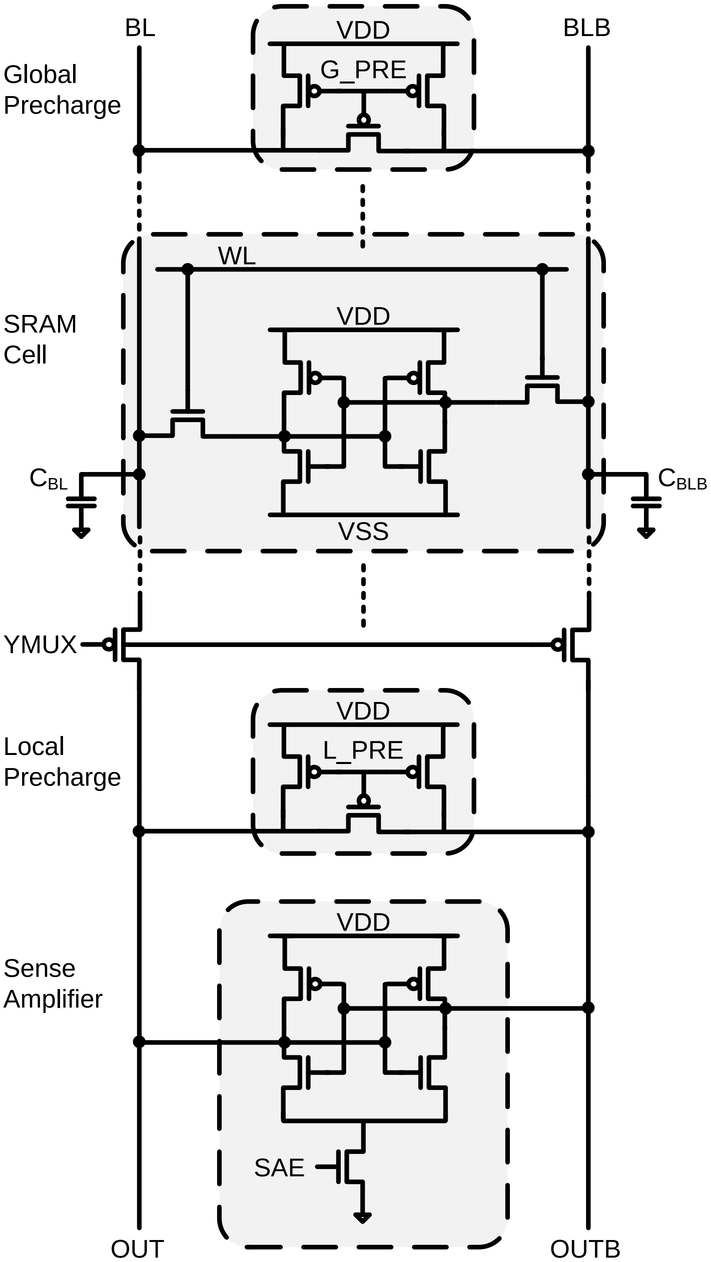


Figure 2.5: Latch-Type Sense Amplifier

entire bitline capacitance. The reading process begins when the precharge and equalize circuitry is turned off, allowing the bitlines and sense amplifier inputs to float. The WL signal is then turned on. One of the bitlines will begin discharging through the storage cell. Once a sufficient differential voltage has been developed on the complementary bitlines, the WL and isolating YMUX transistors are turned off. The SAE signal is then quickly turned on. This isolates the operation of the sense amplifier from the bitlines, and allows the forward feedback action of the sense amplifier to quickly resolve its input/output to a full-swing differential signal.

For the sense amplifier to resolve correctly, the differential input voltage must be greater than some minimum detectable signal. To ensure reliable sensing, this minimum signal should be large enough to overcome any process or environment fluctuations, as discussed in Chapter 3 within the sense amplifier, but should be small enough to prevent excess delay and power dissipation spent unnecessarily discharging and precharging the bitlines.

Differential voltage is developed on the bitlines by exposing them through the access transistors to the storage cell. The wordline access time necessary to develop a given differential voltage is derived as follows:

Beginning with the cell current,

$$I_{Cell} = \frac{\Delta Q}{\Delta t} \tag{2.3}$$

where, I_{Cell} is the cell current sunk during a read operation, ΔQ is the charge draw from the bitline load capacitance, and Δt_{WL} is the wordline access time, the charge, ΔQ, is related to the bitline capacitance, C_{BL}, and differential voltage, ΔV by,

$$\Delta Q = C_{BL} \times \Delta V \tag{2.4}$$

Substituting and rearranging equation 2.4 into equation 2.3 gives:

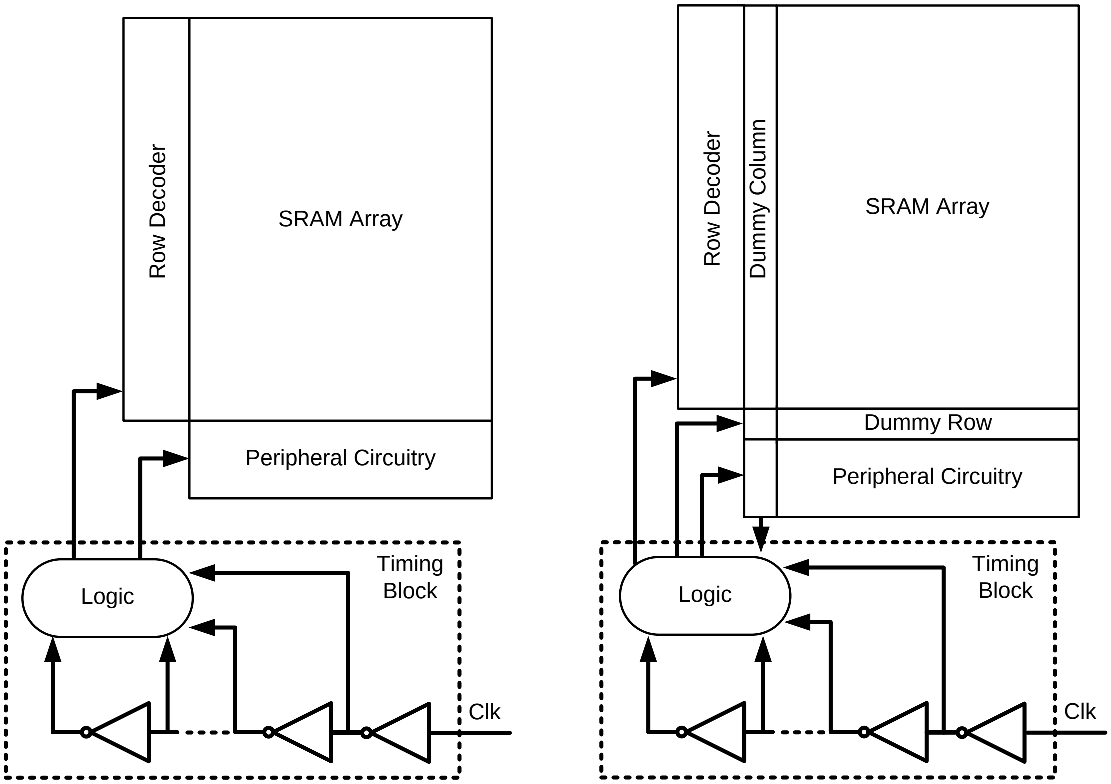
$$\Delta t_{WL} = \frac{C_{BL} \times \Delta V}{I_{Cell}} \tag{2.5}$$

Both ΔV and I_{Cell} are heavily influenced by process variation and mismatch within the sense amplifier and memory cells. As will be discussed in later sections, any fluctuation due to process variation and mismatch can lead to weaker cells, or reduced I_{cell}. If less current is drawn through the cell during reading, then the wordline access time must be increased to develop the necessary differential voltage required for the sense amplifier. Additionally, variation in the parameters of the sense amplifier transistors can lead to a higher required ΔV to resolve data correctly. This can also be corrected by increasing the wordline access time window. This identifies the wordline access time and sense amplifier enable signal as critical for correct operation of the SRAM, and quickly lend themselves as potential candidates to significantly benefit from controllability.

2.4 Modern Timing Control Schemes

There are four different timing control methods typically used in SRAM design. These include: direct clocking [43], delay line timing [37], self-timed replica control [3], and pipelined timing [40]. Direct clocking applies the clock signal directly to the word line and the sense amplifier. This method is limited in that it requires large timing margins for reliable operations, and hence has been superseded by the other methods. Delay line based timing, shown in Figure 2.6(a), uses a chain of inverters to create the required timing intervals. Signals are then "tapped" off of the delay line and passed through logic elements to create the necessary signaling. This allows for tighter margins relative to direct clocking, however it is intrinsically an open loop system, and hence only loosely tracks global process variations. Delay line based timing is investigated in more detail in Chapter 5.

Self-timed replica control, on the other hand, shown in Figure 2.6(b), adds to the delay



(a) Delay Line Timing Scheme

(b) Replica Delay Timing Scheme

Figure 2.6: Control Signal Timing Schemes

line scheme by using a dummy row and column each containing the same number of SRAM cells as the main array to mimic the load capacitances within the array. This allows the timing mechanism to mimic the delays in the SRAM array, leading to better tracking of the global and local process variations, and thus tighter timing margins and performance. Once the dummy column's bitlines have discharged below the switching threshold of the dummy column's sense amplifier, this is fed back into the control logic to turn off the WL signal and turn on the SAE signal allowing output data to be resolved. The dummy column can discharge its bitlines through multiple cells to account for any additional logic delay before the sense amplifiers are enabled. This timing scheme is common in many SRAM implementations [3, 27, 4, 25].

Finally, pipelined timing places a series of registers between the sense amplifier and the data output buffers. This spreads the read delay across multiple clock cycles, and allows the SRAM to be clocked at speeds much higher than the other timing methods. This method is very attractive because it allows the SRAM cycle time to match that of the processor cycle time. The synchronous data buses in large SRAM arrays such as L2 and L3 caches are usually pipelined in modern microprocessor designs [36, 46].

Each of these methods provide their own set of trade-offs in terms of complexity, area overhead, and potential for performance improvements. Although delay line timing provides the least tracking for process variation relative to the self-timed replica control and pipelined timing, it requires much less area overhead and complexity of design. To accommodate for the limitation in process variation tracking, adjustable programmable delay elements can be used to tune the timing characteristics of the timing block.

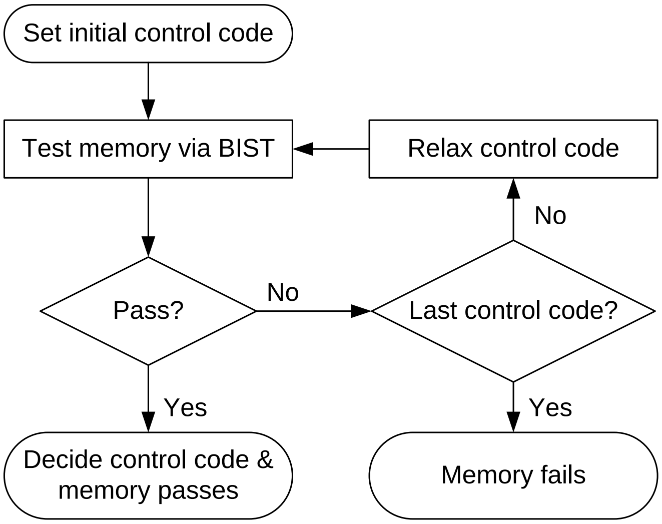


Figure 2.7: Operation flow of a calibration controller during power-on self-test [21]

2.5 Programmable Delay Calibration

Previous work has been done that integrates programmable controllability of the SAE signal into an SRAM's built-in self-test (BIST) unit [6, 21]. Although this work is limited to adjusting only the SAE signal, and does not go into depth regarding the timing related failure mechanisms, it provides a BIST-based calibration procedure for its programmable elements during the power-on self-test (POST). This methodology can be used to determine the proper control code for each individual chip. This is shown in Figure 2.7.

The procedure begins by testing the array with the most aggressive timing setting. If there are failures, the algorithm will incrementally relax the timing via digital control code until failures no longer appear. If elements in the array are still unable to pass functional testing even with the most relaxed timing, then it is deemed to have failed and the die is rejected. Since the controller is embedded inside the memory BIST, the area overhead associated with the controller is almost negligible [21]. While this system only calibrates the array during start-up, it could easily be extended to run periodically to recalibrate the memory in the event of additional device degradation over time.

2.6 Figures of Merit

Many figures of merit (FOM) are used to characterize the standard 6T SRAM cell. These FOM include those relating to the traditional delay, area, and power metrics, as well as memory specific metrics. These are discussed in the following subsections.

2.6.1 Area

The area of the SRAM cell is one of the most significant driving factors for all SRAM design. As SOC's continually demand more memory, the size of the bitcell must decrease in order to increase the amount of memory for a fixed package size. This leads to an increase in memory density. To achieve this, most SRAMs use minimum, or near minimum, sized transistors for their bitcells. This minimum size is dictated by the technology node. As is shown in Figure 1.1, SRAM cell area goes hand-in-hand with technology node scaling, and hence has led the SRAM cell size to become a key metric used by companies to publicize and promote their technology. The general industry standard is roughly a 0.5x area shrink per technology generation. Although, continually scaling the bitcell and increasing the memory density can lead to significant system-level benefits, it comes a substantial penalty in terms of the other FOM.

2.6.2 Current Leakage

Current leakage occurs when there is an unwanted path for charge to flow from the voltage supply down to ground. In deep sub-micron technologies transistor current leakage is a constant issue. Since the devices are never fully off, there is always some sub-threshold leakage. In addition, leakage is more pronounced in smaller devices. This is an issue in SRAMs since the bitcells are made using minimum sized devices. In large SRAM arrays, as transistor counts can be on the order of millions, unwanted leakage accumulates and can lead to substantial power dissipation.

2.6.3 Static Noise Margin

The static noise margin (SNM) is the most common metric of SRAM cell stability [38]. It is defined by the amount of noise voltage a SRAM cell can tolerate before flipping [38]. It can be measured in simulation by applying DC noise sources to the internal nodes of the 6T storage cell and observing the voltage transfer characteristic (VTC) response between the two internal nodes. Figure 2.8 shows the VTC response under both the accessed and retention conditions. The schematic testbench for measuring the SNM is shown in the inset of the figure.

Due to the shape of the curve in Figure 2.8, it is commonly referred to as a butterfly curve. The size of the eye opening within the curve provides a visual representation of the cell's stability. Once the curves have been plotted, the largest possible box is drawn within each of the eye openings. Ideally, the boxes should be identical; however, one may be smaller than the other due to mismatch or process variation within the cell. The SNM of the SRAM is defined as the length of the side of the smaller of the two boxes. SNM measurements can be performed under either access or retention conditions. This is done by having the WL either on or off respectively during simulation. Under access conditions, the additional contribution of the bitline capacitance weakens the feedback action of the storage cell, and hence substantially reduces the access mode SNM as compared to the retention mode. For this reason, worst-case SNM cell robustness is typically measured during the access mode. This measure is also known as the read margin. When a cell is half-selected or being read from, the internal node holding the '0' value must remain below

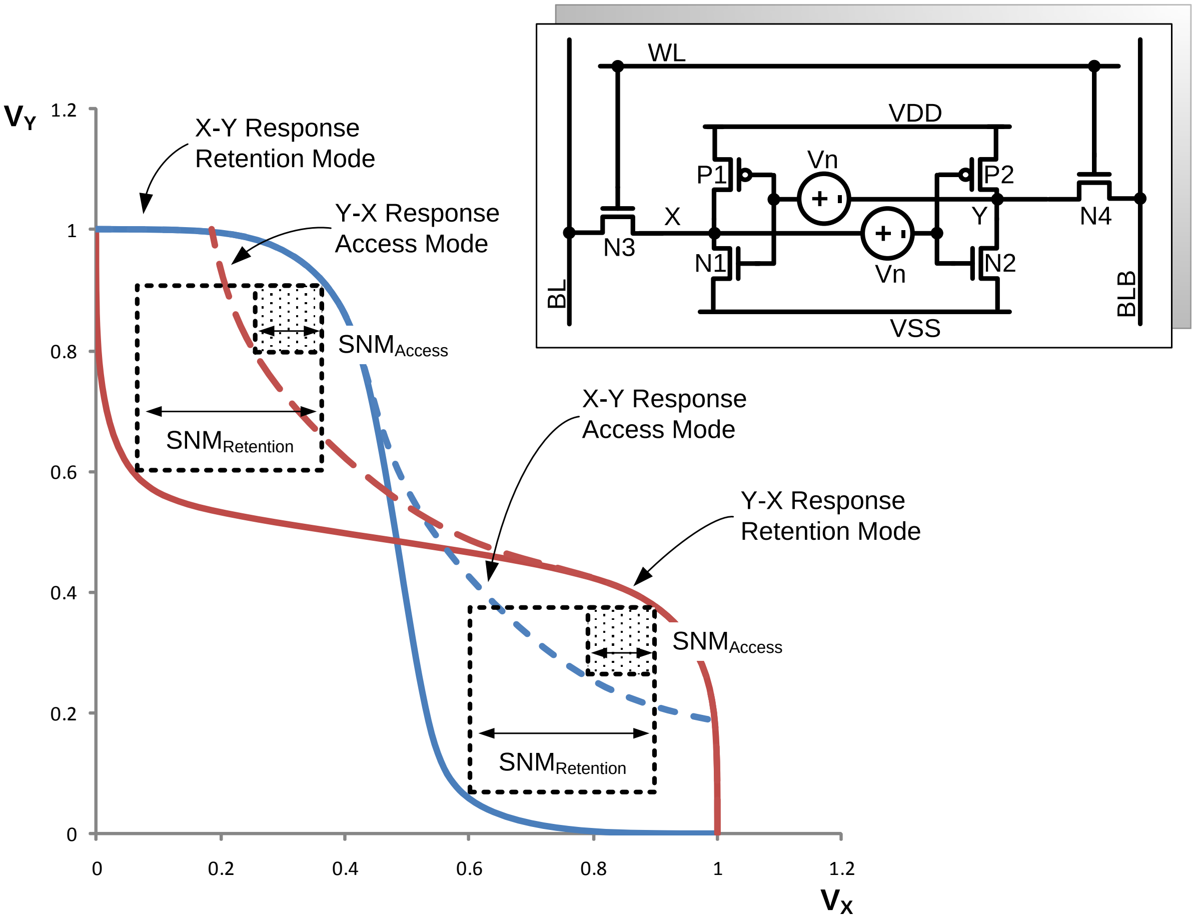


Figure 2.8: SNM measurement testbench and butterfly curves

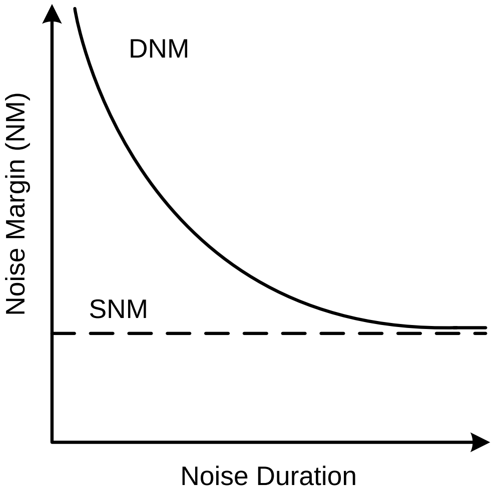


Figure 2.9: Dynamic Noise Margin

the read margin to prevent the read operation from corrupting the data within the cell.

2.6.4 Dynamic Noise Margin

Traditionally, noise margin metrics are static measurements based upon the assumption that the amount of time required for a read or write operation is much larger than the transient time of noise (i.e., SNM). In deep-nanometric SRAM circuits operating at very high frequencies however, this assumption does not always hold [50]. The premise behind dynamic noise margin (DNM) is that noise must be applied to the SRAM cell for a period of time for the cell to become unstable. In fact, an SRAM cell has a time constant which represents the amount of time it takes for a noise source to propagate through the storage cell and flip the data. SRAM cell stability will be maintained so long as the access time is kept below the time constant. This concept is illustrated in Figure 2.9.

When considered as a function of time, the noise margin begins very high. As noise accumulates on the given node, the noise margin gradually decays until it reaches a steady state value. The steady state value is defined as the SNM and the transitionary noise margin as the DNM [39].

Chapter 3

Process Variability & Aging Degradation Mechanisms

When designing an SRAM, process variability and aging degradation are two major concerns that must be taken into account. Both non-idealities influence transistor performance, and in turn the SRAM behaviour. Manufacturing process variability is the first major concern, it produces an initial offset from nominal design values, and then device aging degradation adds on additional variation over time. To account for these variabilities, designers must ensure SRAMs operate correctly within a certain amount of tolerance or variation. These guard bands are characterized in terms of the number of standard deviations, σ's, from the mean, or nominal design value, μ.

Systematic variability causes circuits to vary from die-to-die or wafer-to-wafer, while random variability can cause variations in the properties of adjacent transistors [45]. Variability used to be primarily systematic. As feature sizes scale below 100 nm however, random variability has begun to become increasingly problematic [2].

With continued scaling, the density of SRAM bitcells are able to increase, allowing for more memory to be packed into a given area. The reduction in transistor size however, comes at an increase in variation of transistor process parameters from one device to the other. Transistors are mainly susceptible to deviation from their nominal threshold voltage (V_{TH}), device length and width, as well as oxide thickness. Issues such as random dopant fluctuation can lead to a variation in a transistor's V_{TH}, whereas line edge roughness can vary a transistor's length or width. The measurable effect of process variation can lead to substantial deviation in circuit behaviour from that which is expected. In an SRAM cell, variations may affect the SNM, writeability, or access time. Additionally, the symmetric nature of the SRAM cell makes it especially vulnerable to mismatches in the parameters of paired transistors. Although correct functionality can be ensured by assuming the worst case values for all possible device parameters, this level of over-design can be prohibitively conservative, and thus lead to rather uneconomical circuits. Instead, by statistically modeling these variations, designers can make decisions based on the amount of margin to provide.

Device parameter variations are typically modeled using a normal (Gaussian) distribution, as shown in Figure 3.1. Normal distributions are specified with a standard deviation, σ, about the nominal or mean value, μ. A $\pm 1\ \sigma$ deviation about the mean includes 68.27% of the sampled set, $\pm 2\ \sigma$ deviations includes 95.45%, and $\pm 3\ \sigma$ deviations includes 99.73%. These values are summarized in Table 3.1[1].

Deviation can now be considered in multiple applications. It can refer to the variability of a process parameter from its nominal value, yield of operationally correct bitcells on a die, or even yield of passable dies on a wafer or manufacturing run. For example, if 95.45% of transistors tested exhibit a certain amount of V_{TH} shift from their nominal value, μ, then that amount of V_{TH} deviation represents 2 σ of variability. Whereas, if a 1-Mbit (10^6 cell) SRAM is found to have 2 700 failing cells, it exhibits a 99.9937% cell-level yield. Programmable timing attempts to reduce the cell failure rate and increase this

[1] Defects per million are calculated based on short-term, bi-lateral variability (i.e., a two-sided capability study).

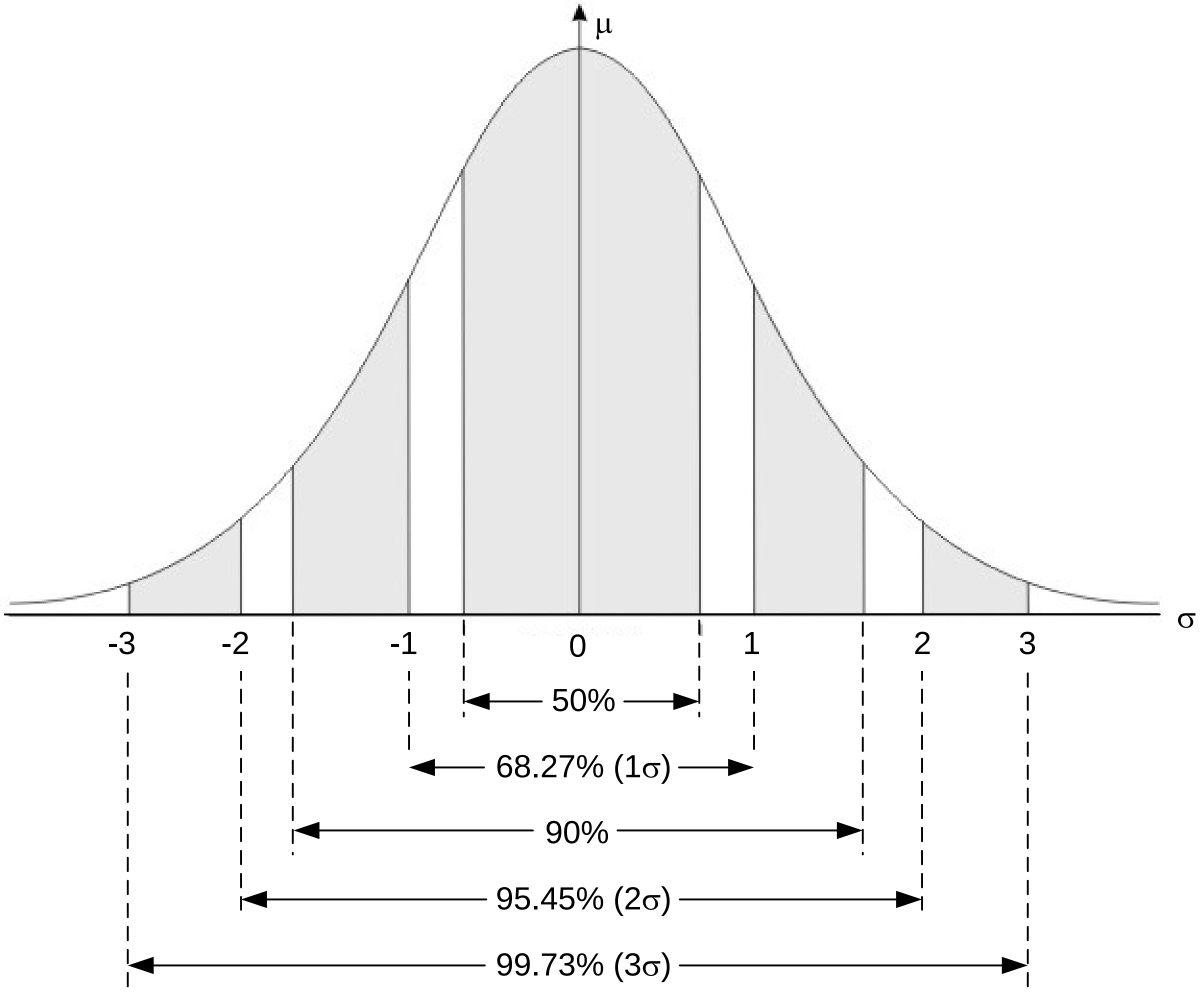


Figure 3.1: Normal Distribution

Table 3.1: Standard Deviation Across Multiple σ and Defects per Million

# of Standard Deviations (σ)	% of Total	Defects/10^6
1	68.27	317 300
2	95.45	45 500
3	99.73	2 700
4	99.9937	63
5	99.99994	0.6
6	99.9999998	0.002

yield. Finally, if a SRAM is considered to be passible if it has less than a certain number of failing cells, then if one million SRAM arrays are manufactured, and 45 500 fail, then it has a 95.45% overall yield. This thesis focuses on variability at the transistor level, with the measurable goal of improving yield at the cell-level by reducing cell failure. This in turn can lead to improved yields at the high-volume manufacturing level.

3.1 Mechanisms for Transistor Variability

3.1.1 Random Dopant Fluctuation

One of the most significant sources for process variability is random dopant fluctuation [7]. Due to the finite number of dopant atoms in the extremely small MOSFET channel area, there exists a fundamental variability in the threshold voltage. To achieve a channel dopant concentration of 10^{19} $atoms/\text{cm}^3$ in a MOSFET with channel length less than 50 nm requires less than 100 dopant atoms. Any absence or addition of only a few dopant atoms will lead to a variation in channel dopant concentration, and thus variation in threshold voltage, V_{TH}. Figure 3.2 shows the standard deviation of the threshold voltage $\sigma_{V_{TH}}$, as a function of one over the square root of channel area ($1/\sqrt{W \times L}$) for both a 90 nm and a

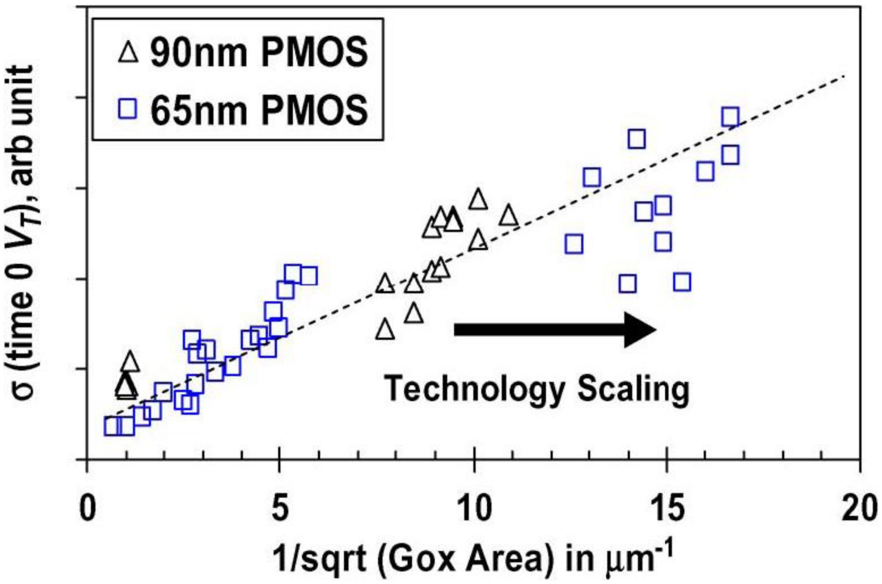


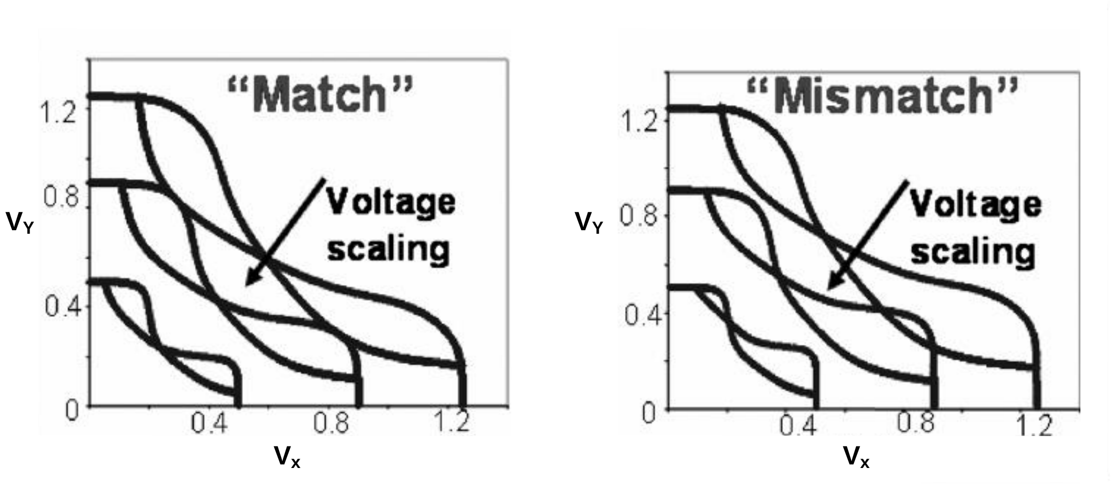
Figure 3.2: V_{TH} variability as a function of channel area for both a 90 nm and a 65 nm process. The line is a guide to the eye and not necessarily a fit to the data [28].

65 nm process [28].

As technology scales, the device's channel area will decrease, and thus lead to an increase in threshold voltage variability. This threshold voltage variation due to random dopant fluctuation increases proportionally with $1/\sqrt{WL}$ as described by Pelgrom [31].

3.1.2 Line Edge Roughness

Line edge roughness arises from a combination of the resolution limit of the lithography process and material characteristics, resulting in non-uniformity in local line widths [34]. This roughness is on the order of a few nanometers and becomes significant for sub-micron technology. Although the absolute variance of the line width decreases as the feature size scales down, the line edge variance relative to the feature will increase. This leads to an increase in device dimension variability for scaled devices.



(a) Ideal Matching

(b) With Mismatch

Figure 3.3: SRAM VTC curves under both ideal and non-ideal conditions due to transistor mismatch [13]

3.2 SNM Variability in the 6T SRAM Cell

For the ideal SRAM cell, shown in Figure 2.3, the voltage transfer characteristic of both halves of the cell is perfectly symmetrical; as can be seen in Figure 2.8, both squares within the eyes of the butterfly curve are of the same size. As the cell is affected by process variability however, the properties of one transistor will vary from its paired transistor. This mismatch between transistor pairs creates an asymmetry in the cell's voltage transfer characteristic. An example of this is shown in Figure 3.3. The measured SNM is the side of the smaller of the two squares that can fit within the eyes of the butterfly curve.

The butterfly curves shown in Figure 3.3, obtained by Hamzaoglu et. al., were measured in a 45 nm 1.2 V process [13]. In addition to showing the effect of transistor mismatch, the plots also show how the SNM scales proportionally with voltage. This is consistent with the work done by Seevinck et. al. [38].

The SNM values for Figures 3.4, 3.5, and 3.6 were obtained by Pavlov and Sachdev using a 6T cell in a 0.13μm CMOS process with $V_{DD} = 1.2V$ using special SRAM transistor models [30]. The data is normalized with respect to the typical case (typical process corners, ambient temperature, typical voltages) by the following equation:

$$SNM_{realative} = \frac{SNM_{measured} - SNM_{typical}}{SNM_{typical}} \times 100\% \tag{3.1}$$

Once the SNM variability is known, it can be correlated to the SRAM yield. It has been shown that the $\mu - 6\sigma$ SNM value must be greater than 4% of V_{DD} to obtain a 90% yield on a 1 MB SRAM [42]. Asymmetries within the cell will lead to a reduction in the SNM and an increase in the number of unstable SRAM cells, thus impacting the yield. This typically translates into a requirement that $SNM_{MIN} \geq 20\%\ SNM_{TYPICAL}$ [30]. The SNM deviation from the mean as a function of threshold voltage deviation from the mean is shown in Figure 3.4. The relationship is shown for slow, fast, and typical process corners, as well as for variations in the driver, pull-up, and access transistors. V_{TH} variation is performed for one transistor at a time, while the other transistors remain at their nominal V_{TH} value. Sweeping the V_{TH} of one transistor, effectively creates a mismatch between that particular transistor and its corresponding pair transistor. This in turn creates an asymmetry within the SRAM cell.

The V_{TH} variation of the driver transistor causes the greatest variation in SNM. This is due to its large W/L ratio compared to the other transistors within the SRAM cell [30]. The SNM variation caused by altering the V_{TH} of the access transistor depends on which way the V_{TH} is altered. Decreasing the access transistor's V_{TH} decreases the SNM of the cell, whereas increasing the V_{TH} has only a marginal impact. Since the SNM is being measured during a read access, lowering the V_{TH} of the access transistor will effectively reduce the cell ratio of one side of the cell, leading to an increase in the logical '0' voltage value, which in turn leads to a decrease in SNM. Finally, varying the V_{TH} of the PMOS

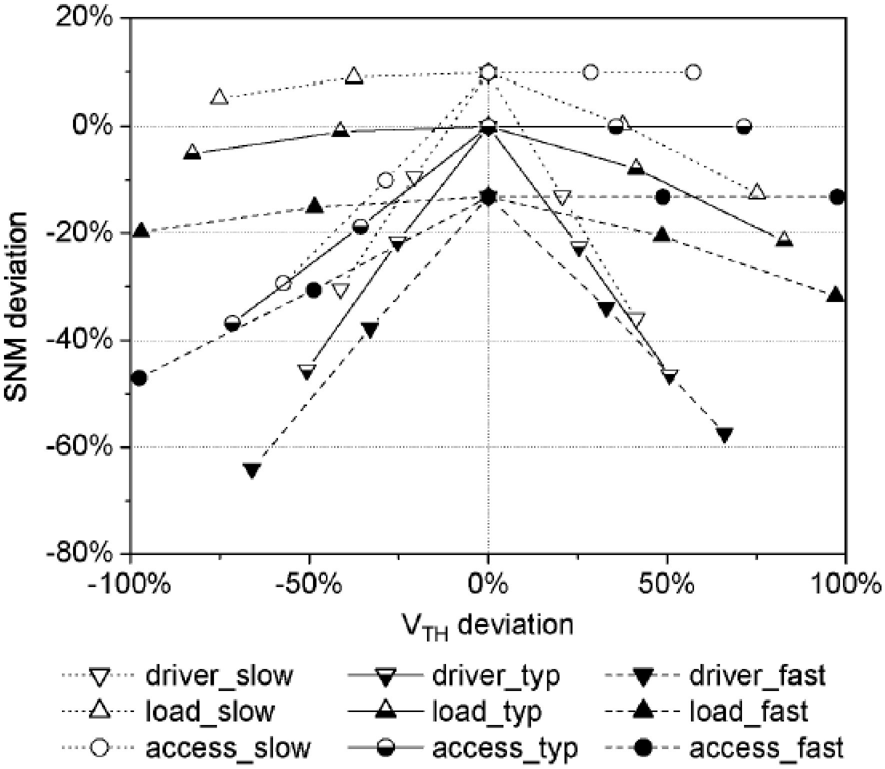


Figure 3.4: 6T SRAM cell SNM deviation vs. threshold voltage deviation on one of the transistors [30]

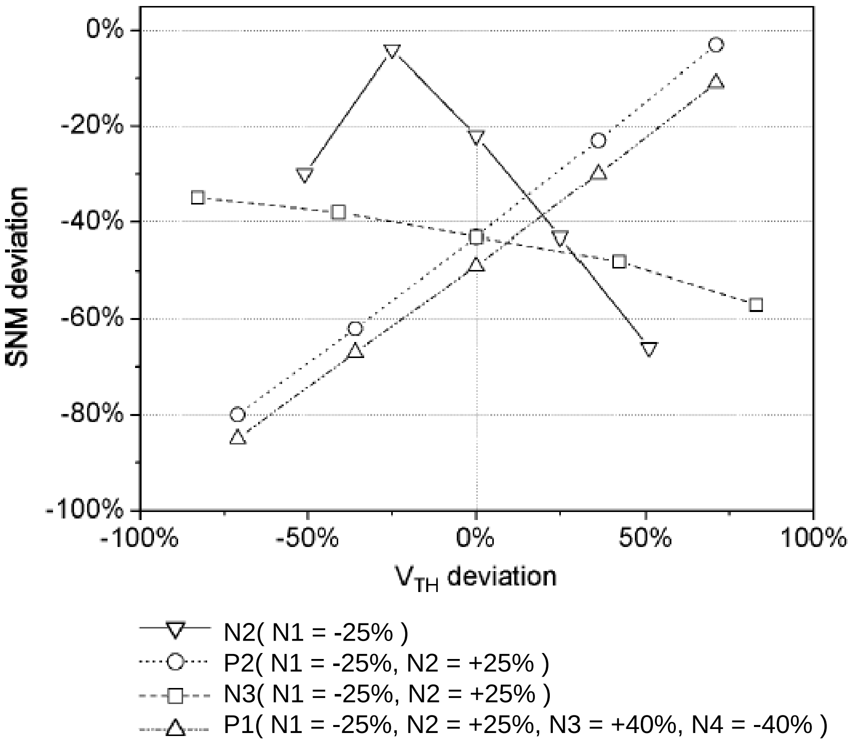


Figure 3.5: SRAM cell SNM vs. threshold voltage deviation of more than one transistor [30]

load transistor has a minimal impact on the SNM. This is due to its intrinsic weaker drive strength and small W/L ratio relative to the NMOS access and driver transistors within the cell. Note that when the V_{TH} deviation is zero this indicates that all transistors are at their nominal V_{TH} values, and the cell is symmetric.

While Figure 3.4 shows the SNM deviation versus V_{TH} deviation for a single transistor within an SRAM cell, if more than one transistor exhibits a V_{TH} deviation from its nominal value, the SNM deviation can be more drastic. Figure 3.5 shows a variety of cases where multiple transistors exhibit a V_{TH} deviation.

N2(N1 = -25%) represents the case where the V_{TH} of transistor N2 is the dependant variable, and transistor N1 has a constant deviation of -25% of its nominal V_{TH} value. Note

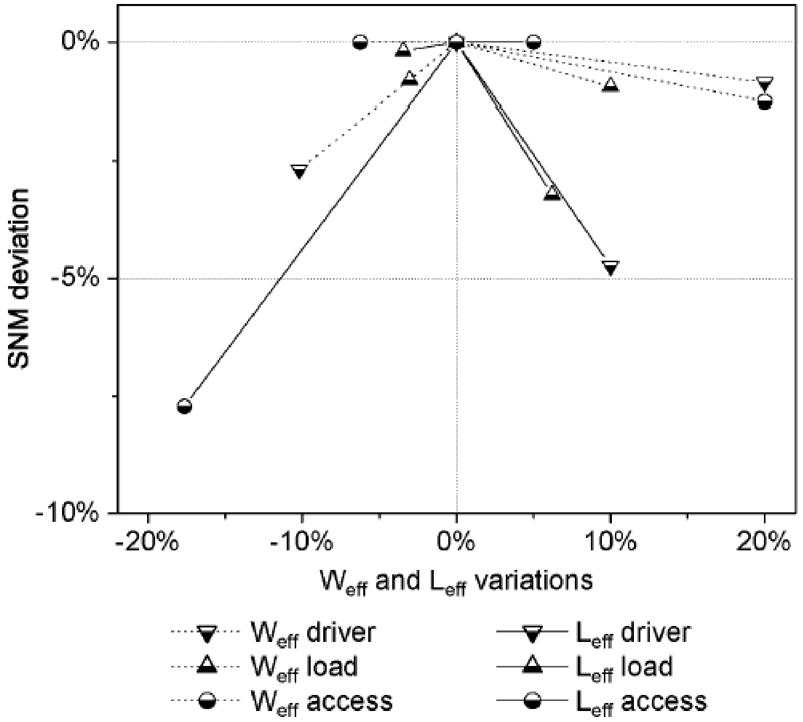


Figure 3.6: SRAM cell SNM deviation vs. transistor Length (L), and Width (W) [30]

that in this case, the cell obtains its maximum SNM (minimum deviation) when the two transistors experience a -25% deviation and the cell is symmetrical. The P1(N1 = -25% , N2 = +25% , N3 = +40% , N4 = -40%) provides one of the worst case SNM degradations due to asymmetry of the transistor's V_{TH}.

Mismatch in the length, L, and width, W, of SRAM cell transistor pairs also contribute to SNM deviation. Their contribution is marginal however, when compared to the V_{TH} deviation contribution. Figure 3.6 shows the SRAM cell's SNM dependence on W and L variation in a single transistor under typical conditions.

Regardless of the direction of the geometry deviation, the optimal SNM occurs at nominal transistor sizing. This is because any deviation causes asymmetries within the cell and hence SNM degradation. The most significant causes of SNM degradation occur for geometry deviations that lead to a decrease in the cell ratio. This includes decreasing the driver transistor width or access transistor length, or increasing the driver transistor length. Decreasing the cell ratio increases the logical ‘0’ voltage level stored within the cell, which leads to a decrease in SNM. Overall, Figure 3.6 shows that a weaker (smaller W/L ratio) driver transistor or a stronger access transistor decreases the SNM, and the deviation in the load transistor has a minimal affect on the SNM.

3.3 Aging Mechanism

Over time, a transistor’s properties have a tendency to degrade and shift from their designed nominal value. There are three mechanisms that are widely recognized in the semiconductor industry as the most prominent lifetime reliability concerns for transistors. These include: gate-oxide breakdown, hot-carrier effects, and bias temperature instability [8].

3.3.1 Gate-Oxide Breakdown

Gate-Oxide breakdown can occur when there is a voltage drop across the gate stack. During this time, traps can be created within the dielectric. Traps are electrically active defects that capture carriers at energy levels within the bandgap. Traps created within the dielectric can reduce the V_{TH} of the device. Additionally, these defects may eventually join together and form a conductive path through the stack, creating a leakage path. This can be seen in Figure 3.7.

Breakdown has become an increasing cause for concern as the gate dielectric thickness has be scaled down to the one nanometer range. By having a thinner gate oxide, a smaller critical trap density is required to tunnel through the oxide, damaging the device, and allowing leakage current to flow [18]. The scaling of the physical dimensions of the gate

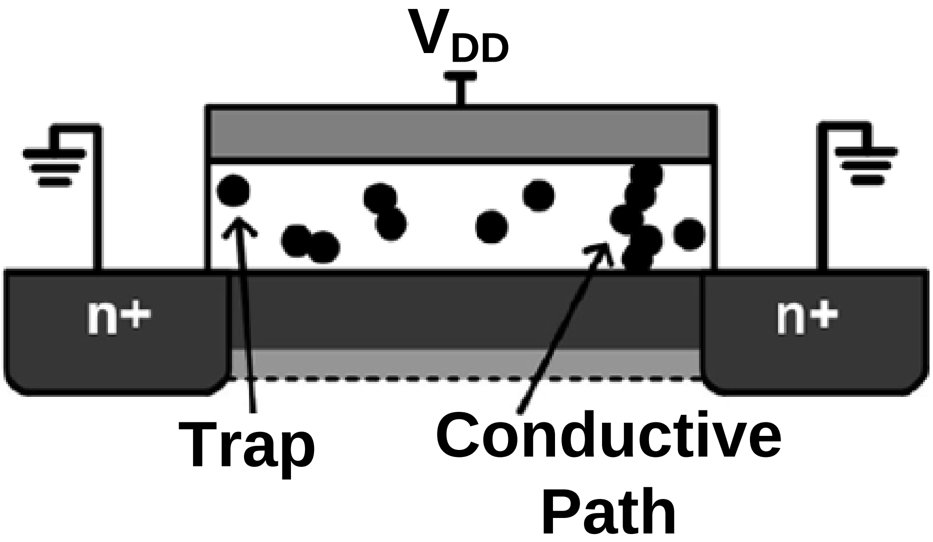


Figure 3.7: A conductive path in the gate stack due to gate-oxide breakdown stress [18]

stack can be slowed or reversed with the introduction of different materials in the stack such as high-κ dielectrics. High-κ dielectrics are those with a high dielectric constant, κ, compared to silicon dioxide, SiO_2. These allow for an oxide capacitance comparable to that of a thin SiO_2 dielectric, while keeping the actual oxide thickness relatively high.

3.3.2 Hot Carrier Injection

Hot carrier injection (HCI), occurs when hot carriers (those with high kinetic energy) are accelerated towards the drain by a lateral electric field across the channel and generate secondary carriers through impact ionization. If either the primary or secondary carrier gains enough energy, it can be injected into the gate stack. Carriers injected into the gate stack can create traps within the oxide that can alter the V_{TH} of the device. This phenomenon is shown in Figure 3.8.

HCI has become less prominent with the reduction of operating voltage, but remains a

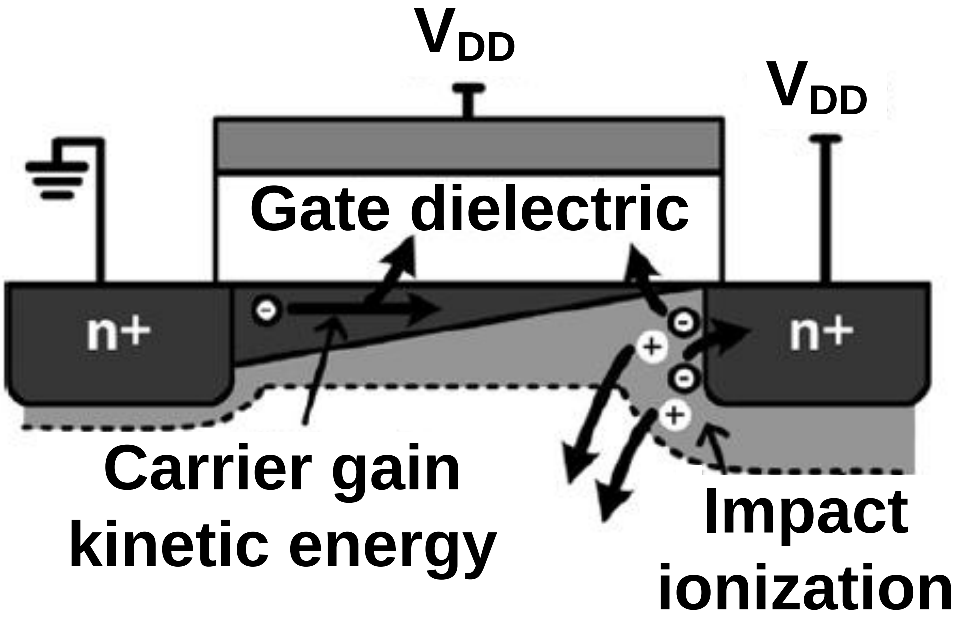
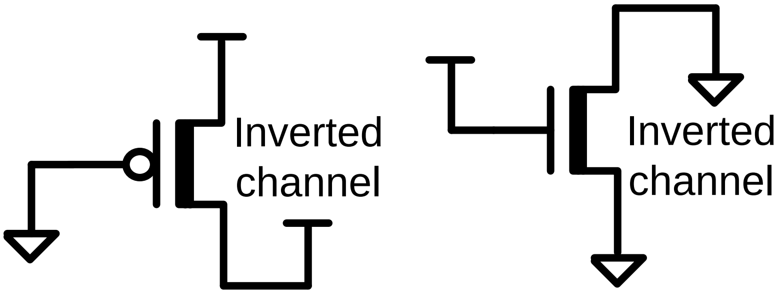


Figure 3.8: Hot carrier injection stress mechanism [18]



(a) Negative Bias Temperature Instability (b) Positive Bias Temperature Instability

Figure 3.9: Conditions for negative and positive bias temperature instability stress

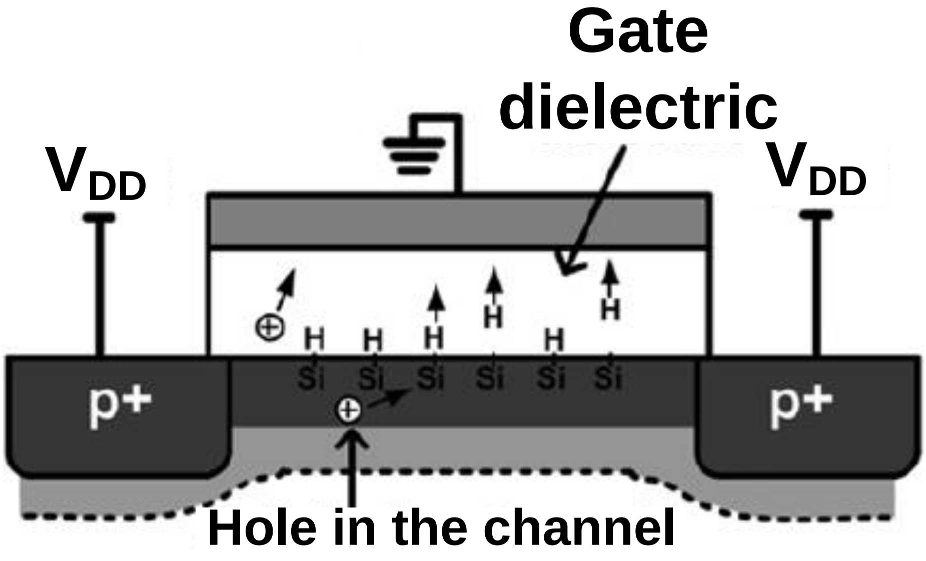
serious concern due to the large local electric fields in scaled devices [18].

3.3.3 Bias Temperature Instability

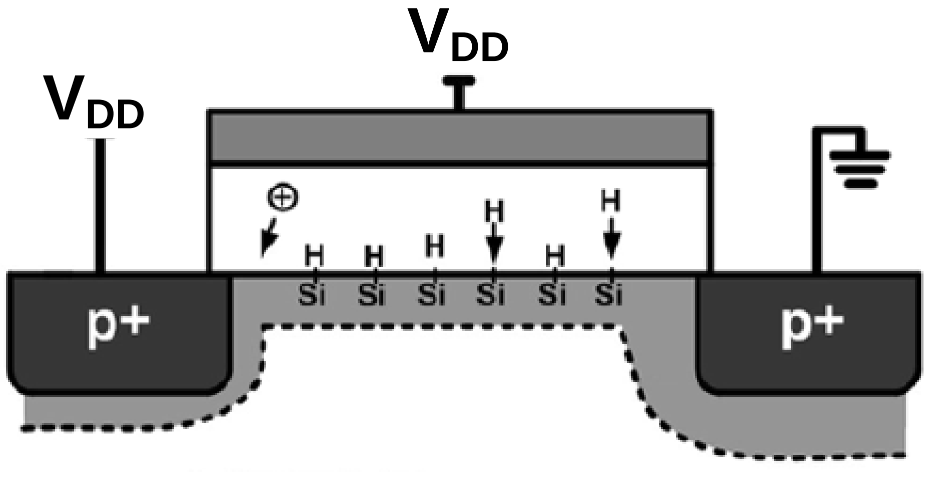
Bias temperature instability (BTI), occurs in two different variants: Negative BTI (NBTI) in PMOS devices and Positive BTI (PBTI) in NMOS devices, shown in Figure 3.9.

NBTI in PMOS transistors is often cited as the primary reliability concern in modern CMOS processes [18]. It is characterized by a positive shift in the V_{TH} of the device occurring when it has been biased in strong inversion, but with a minimal lateral electric field ($V_{DS} \approx 0\ V$) over a period of time. The V_{TH} is generally attributed to hole trapping in the dielectric bulk, and/or to the breaking of Si-H bonds at the gate dielectric interface caused by holes in the inversion layer, and generates positively charged interference traps [12, 16]. This is shown in Figure 3.10.

When a stressed device is turned off (i.e., the bias is removed from the gate) the transistor is able to "recover". During this recovery phase, the trapped holes are released and the free hydrogen diffuses back towards the substrate/dielectric interface, recombining with the silicon to reform the Si-H bonds. This reverses the positive V_{TH} shift to its nominal value. PBTI in NMOS devices, shown in Figure 3.9(b), is similar to NBTI in



(a) NBTI Stress State



(b) NBTI Recovery State

Figure 3.10: NBTI Stress and Recovery States [18]

PMOS devices, only the strong inversion is generated by biasing the gate at V_{DD} and the minimal lateral electric field is maintained by holding the source and drain close to ground. PBTI in NMOS transistors has been found to be non-critical in silicon dioxide dielectrics, however it does contribute to the aging of high-κ dielectric gate stacks that are now being seen in newer technology nodes [11].

A comprehensive model for NBTI V_{TH} shift is given in [44]. It is summarized here. Interface traps, N_{it}, formed between the channel and the gate result in an increase in charge in the gate stack. This causes a shift in V_{TH} as follows:

$$\Delta V_{TH} = \frac{qN_{it}}{C_{ox}} \text{ , where } C_{ox} = \frac{\epsilon_{ox}}{T_{ox}} \tag{3.2}$$

where C_{ox} is the gate oxide capacitance per unit area, q is the electron charge, ϵ_{ox} is the dielectric constant, and T_{ox} is the oxide thickness. The total number of interface traps N_{it} is dependent on whether or not the transistor is in the stress or recovery state, and is calculated as follows:

Stress:

$$N_{it} = \sqrt{K^2 \cdot (t - t_o)^{0.5} + N_{it0}^2} + \delta \tag{3.3}$$

$$K = A \cdot t_{ox} \cdot \sqrt{C_{ox}(V_{gs} - V_{TH})} \left(1 - \frac{V_{ds}}{\alpha(V_{gs} - V_{TH})}\right) \cdot \exp\left(\frac{E_{ox}}{E_o}\right) \cdot \exp\left(\frac{-E_a}{kT}\right) \tag{3.4}$$

Recovery:

$$N_{it} = (N_{it0} - \delta) \cdot \lfloor 1 - \sqrt{\eta(t - t_o)/t} \rfloor \tag{3.5}$$

where t is the time elapsed in seconds, N_{it0} is the amount of interface traps at initial time, t_o, δ is a constant representing non-H based oxide traps and other charged residues, t_{ox} is the oxide thickness, C_{ox} is the oxide capacitance, E_{ox} is the electric field across the

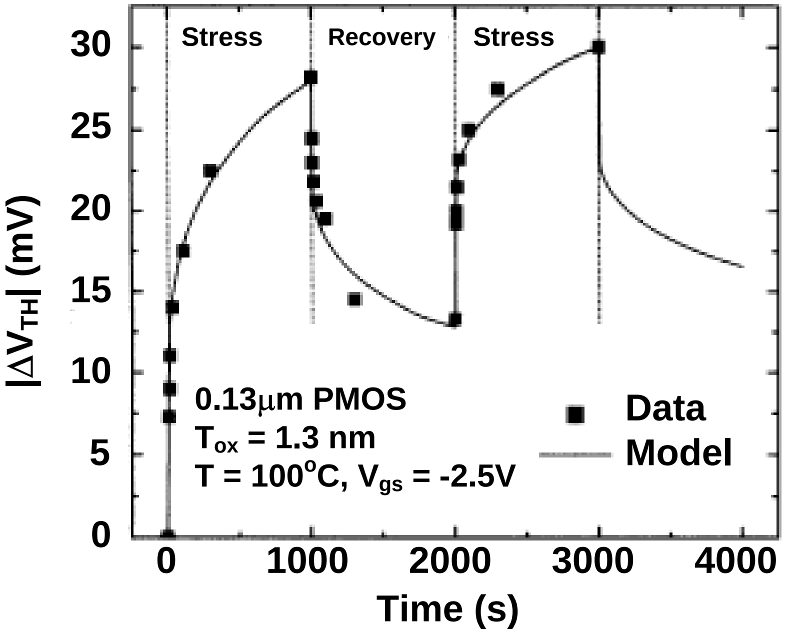


Figure 3.11: ΔV_{TH} for PMOS devices under NBTI stress and recovery conditions [44]

oxide, k is the Boltzmann constant, T is the temperature, and α E_o, E_a, and η are fitting parameters.

Figure 3.11 shows ΔV_{TH} due to NBTI for a PMOS transistor under both stressed and recovery conditions [44]. In the stressed state, the PMOS first undergoes a rapid increase in V_{TH} and then the rate of increase begins to taper off. Once the stress is removed, and the device is allowed to recover, V_{TH} begin to decreases. The figure shows alternating stress and recovery times of approximately 15 minutes over the period of one hour.

3.4 Aging in SRAM

Aging affects SRAM performance in much the same way as process variation. When transistors experience an applied electrical stress, their parameters, most notably V_{TH}, have a

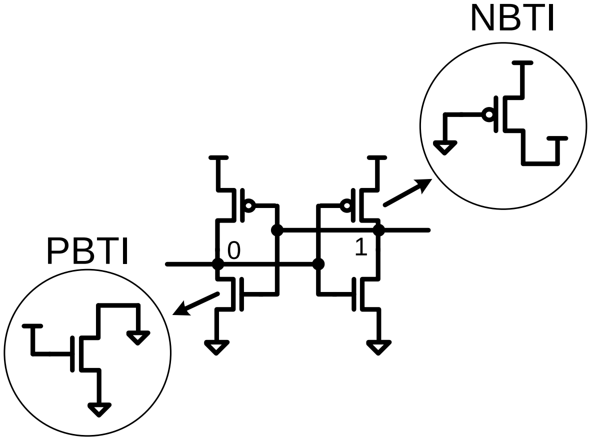


Figure 3.12: BTI susceptible transistors within the SRAM cell

tendency to shift from their nominal value. When these stresses are applied asymmetrically on the SRAM cell, they create a mismatch between the cell's transistor pairs, and cause a reduction in the cell's SNM. This SNM degradation leads to cell failure.

NTBI is the most significant aging mechanism present within SRAMs [41]. During the SRAM's retention mode, one PMOS load transistor and one NMOS driver transistor in every memory cell will be subject to NBTI and PBTI stress respectively at any given period of time. This can be seen in Figure 3.12.

The PMOS transistor responsible for retaining the '1' has a $V_{DS} \approx 0$ V and a stress on the transistor being applied by the grounded gate. This causes the PMOS to undergo NBTI degradation, and cause a positive shift in that transistor's V_{TH}. Additionally, the NMOS responsible for retaining the '0' will undergo PBTI stress. This effect will be minimal in silicon dioxide gate stacks; however, the effect on SRAM's using high-κ dielectric gate stacks will become significant. This can be seen in Figure 3.13.

As technology advances, and new high-κ materials are being used for the gate, BTI aging effects become more severe for both the NMOS and PMOS devices. Additionally, since PBTI stress affects the driver transistor, it has the potential to significantly impact

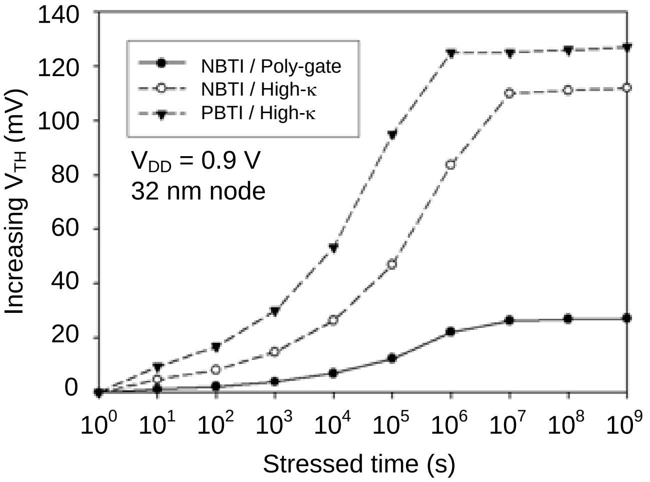


Figure 3.13: ΔV_{TH} for BTI Stress in both SiO_2 and high-κ gate stacks [49]

the SNM of the cell. This is due to the fact that, as was seen in Figure 3.4, mismatch in the driver transistor has the most significant impact on SNM of any of the 6T SRAM cell transistor pairs.

Since memory arrays have a relatively low switching activity (since switching only occurs when new data is written into a cell, and data can only be written one word/port at a time in an array of potentially millions of data words), memory bitcells can be exposed to BTI stress for extended periods of time. As this stress is only applied to one side of the bitcell at any given time, asymmetries arise in the V_{TH}'s of the cell's transistors, leading to mismatch and a degraded SNM for the cell. With continued stress, this mismatch gets worse over time, leading to a further degradation in SRAM SNM.

Chapter 4

SRAM Timing Failures

The timing control block is a critical component in any SRAM design. It is responsible for generating all of the internal signals for the correct read and write operation of the SRAM. These signals include control for the precharge, word line, sense amplifier clocking, and write driver activation. Several SRAM cell failure mechanisms are heavily influenced by the cell's control signal timing. These failures are 1) operational, when an operation is not completed successfully, 2) stability related, if the cell's data gets corrupted, or 3) power related, if it causes the SRAM array to consume an excessive amount of power. These are a subset of those failure mechanisms listed in Chapter 1. Variable timing circuitry allows these failures to be corrected or at least reduced. Each of these failure mechanisms are discussed below.

4.1 Operational Read Failure

Since a read is typically the slowest memory operation, its timing is the most vulnerable to failure [35]. During a read operation, the amount of differential voltage generated on the bitlines is directly proportional to two parameters: the width of the wordline signal

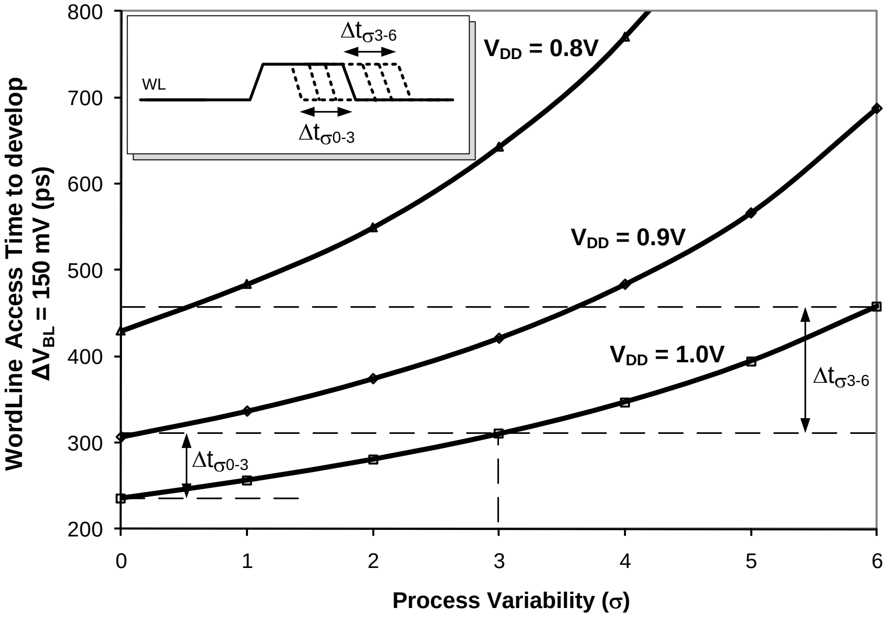


Figure 4.1: Effect of process and voltage variations on required cell access time

and the strength of the SRAM cell. The width of the wordline signal is a function of the timing block design; however, the strength of the SRAM cell is a function of process, process variability, aging degradation, and the cell design. Large SRAM arrays can contain hundreds of millions of transistors, all of which can differ from the ideal performance, both systematically and randomly.

To observe the effects of variability on the amount of time required to generate the required differential voltage on the bitlines for a successful read operation, Monte Carlo simulations were performed on a 6T SRAM cell in a 65 nm standard CMOS process. The results are presented in Figure 4.1. These simulations were repeated for reduced supply voltages. Looking at the response when the supply voltage is at the full 1 V, it can be seen that the required wordline width increases from approximately 240 ps for 0 σ to 450 ps for 6 σ of variability. Using variable timing, the control signal of an SRAM array can be optimized in silicon.

An array designed to cover 3 σ of variation using a static timing would have a wordline pulse width set to 310 ps. This would cover 99.73% of the variability cases. A flexible timing scheme would have three benefits. It could increase the yield by providing extra time for the read operation to complete in the cases of variability beyond 3 σ. For the majority of dies whose variability is less than 3 σ, a flexible timing scheme would create more optimal timing signals, allowing those dies to be operated at with a higher DNM and reduced power dissipation because the cell is being accessed for a short period of time. Moreover, the supply voltage can be reduced, while still maintaining a guard band of a given number of σ.

Additionally, a fabricated array will have an unknown amount of variability. By using flexible timing, the edges of the control signals can be moved to not only correct failures, but also to characterize the array's variability. By starting with the most aggressive timing setting, and relaxing that timing until the SRAM performs correctly, or visa versa, with the most relaxed timing, and pushing the timing until failure, the residual difference between nominal timing setting and those of the chip-under-test can be characterized. This can lead to "binning" of chips based on their amount of variability.

4.2 Cell Stability Failure

In an SRAM array containing multiple words per row, a cell is said to be half-selected when it is accessed via the wordline, but its bitlines are not routed to the sense amplifier. In the case of a half-selected cell, the dynamic noise margin is determined by the width of the wordline access time window. Cells weakened due to process variation and aging experience a lower DNM.

To illustrate this response, simulations were performed on a 6T SRAM cell in a 65 nm standard CMOS process. Resistors are used to symmetrically weaken the cell, as shown in Figure 4.2. If the resistance is relatively low, it models the effects of process variability,

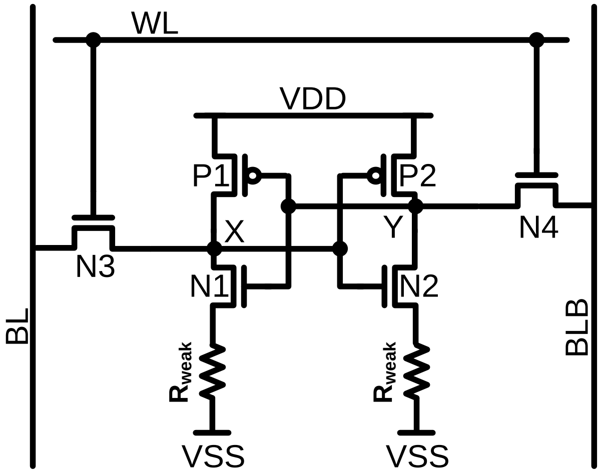


Figure 4.2: Schematic of a Weak 6T SRAM Cell

whereas if the resistance is large it models the effect of defects, such as high-resistance contacts. As can be seen in Figure 4.3, failures are the result of both the resistance and the wordline timing.

When the value of R$_{weak}$ is low, or when the access time is low, the cell is stable; however, if the resistance is large enough, and the access time is sufficiently long, the cell can become unstable. This behavior shows a strong dependence on the supply voltage. For example, a weakened SRAM cell with R_{weak} = 10 kΩ is stable with a supply voltage of 1 V. If the supply voltage is reduced to 0.7 V however, the width of the wordline signal must be kept to less than 100 ps or else the cell will become unstable. These results are similar to those of Sharifkani and Sachdev [39]. In their work, they show measured results that illustrate the relationship between cell stability and access time, as can be seen in Figure 4.4.

Care must be taken when designing the timing for the SRAM array so that enough time is available for the selected cells to develop the required differential voltage on the bitlines for the sense amplifier to resolve the data; however, not so much time as to upset the half-selected cells.

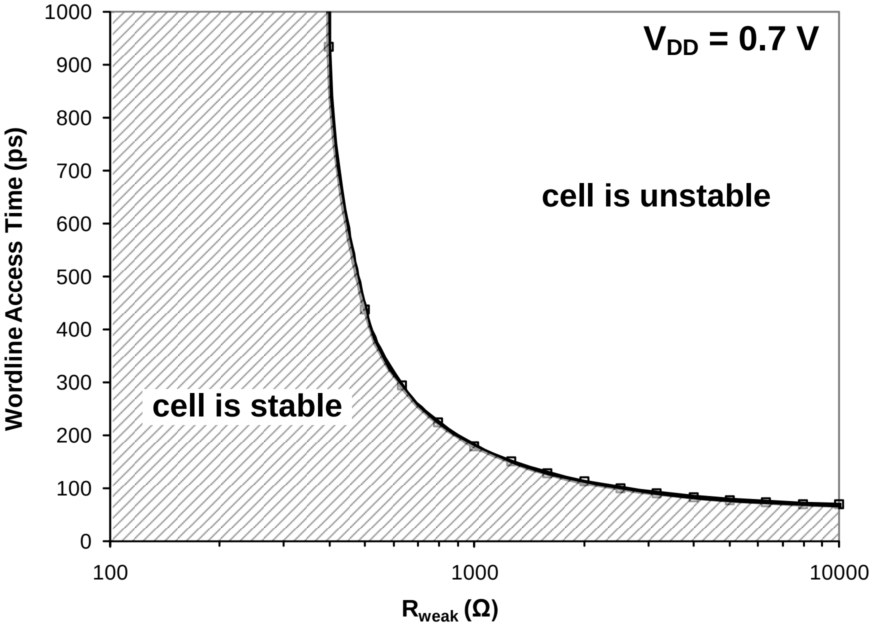


Figure 4.3: Weakened 6T SRAM dynamic cell stability for variable cell access time at a reduced supply voltage, $V_{DD} = 0.7$ V

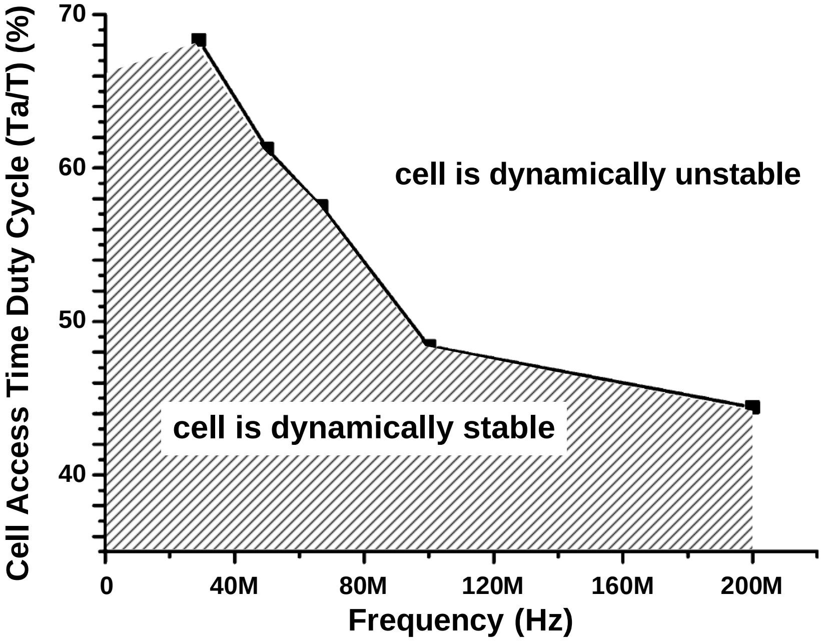


Figure 4.4: Measured DNM of a 6T SRAM cell [39]

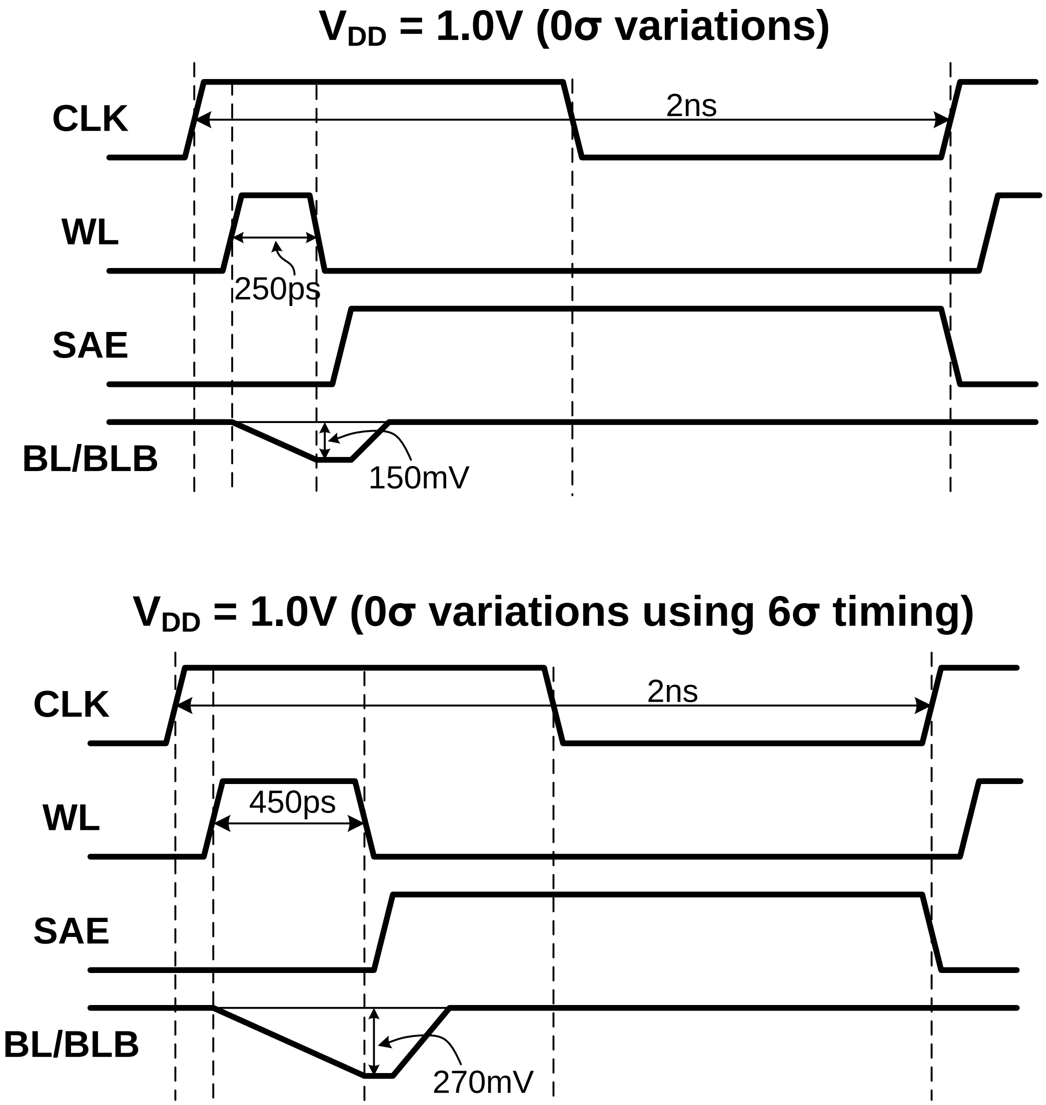
4.3 Power Envelope Failure

Changing the SRAM control timing can have a large effect on the power dissipation of an SRAM; this is especially true during a read operation. During a read operation, one of the bitlines is discharged; however, it only needs to be sufficiently discharged for the sense amplifier to be able to resolve the correct data value. Earlier, in Section 4.1, it was shown that process, aging, and voltage can affect the required timing for an SRAM array. It was shown that for 6 σ of variation at 1 V, a wordline width of 450 ps was required to read successfully, compared with 250 ps for typical process conditions. If the wordline width was set to 450 ps to cover the 6 σ variations, all of the dies with lower variability would discharge their bitlines beyond that which was necessary, resulting in larger power dissipation. Figure 4.5 illustrates this situation by showing the SRAM control signals and the bitline voltages. For the situation where there are no variations with a wordline width of 250 ps, a differential bitline voltage of 150 mV is developed. However, if this is increased to 450 ps, the differential voltage developed on the bitlines is 270 mV. The word size in modern SRAMs may be as large as 128-bits, and as such each of these columns will dissipate unnecessary power during each read operation. A flexible timing approach allows each die to have the optimal wordline width to prevent this from happening.

It is common for SRAM arrays to operate on lower supply voltages to reduce power, especially leakage power. Figure 4.1 shows that lower supply voltages require longer access times to generate the necessary differential voltage on the bitlines. With variable timing, the SRAM array could be characterized to determine the wordline width required to generate sufficient differential voltage on the bitlines for a variety of supply voltages.

During a read operation, the array switching power is calculated as

$$P_{switch,\ array} = N_{BL} C_{BL} \Delta V_{BL} V_{DD} f \alpha \tag{4.1}$$



V_{DD} = 1.0V (0σ variations)

V_{DD} = 1.0V (0σ variations using 6σ timing)

Figure 4.5: Example timing configurations for both nominal and reduced supply voltage, V_{DD}

where N_{BL} is the number of bitlines being discharged, C_{BL} is the bitline capacitance, ΔV_{BL} is the developed bitline differential voltage used by the sense amplifier to sense the cell's stored value, V_{DD} is the supply voltage, f is the operating frequency, and α is the switching activity. To a first order, ΔV_{BL} can be approximated by assuming a linear dependence on the wordline width, T_{WL}, where $\Delta V_{BL} < V_{DD}$ (as should always be the case for a differentially sensed SRAM).

Provided that the bitlines do not fully discharge, as shown in [1], the array switching power can be rewritten as

$$P_{switch,\ array} = N_{BL} I_C T_{WL} V_{DD} f \alpha \tag{4.2}$$

where I_C is the bit-cell read current. Therefore, the switching power associated with the SRAM is directly proportional to the wordline width. This provides additional incentive for the designer to limit the wordline access time to only what is necessary to sense the cell.

4.4 Timing Related Cell Failure Reduction

To measure the degree of cell failure reduction through programmable timing, Monte Carlo simulations were run on a 6T SRAM cell in a 1.2 V, 65 nm standard CMOS process. The results are shown in Figure 4.6. For a static wordline access time of 375 ps, 96% of cells were able to develop a differential bitline voltage greater than 50 mV. As the sense amplifier undergoes process variation or device aging, the required differential bitline voltage for the sense amplifier to correctly resolve data increases. For a fixed wordline access time, process variation and device aging within the 6T memory cells prohibits the necessary differential bitline voltage from being developed. As the wordline access time is progressively increased to 500 ps, 52.1% more cells can produce over 120 mV of differential bitline voltage than for

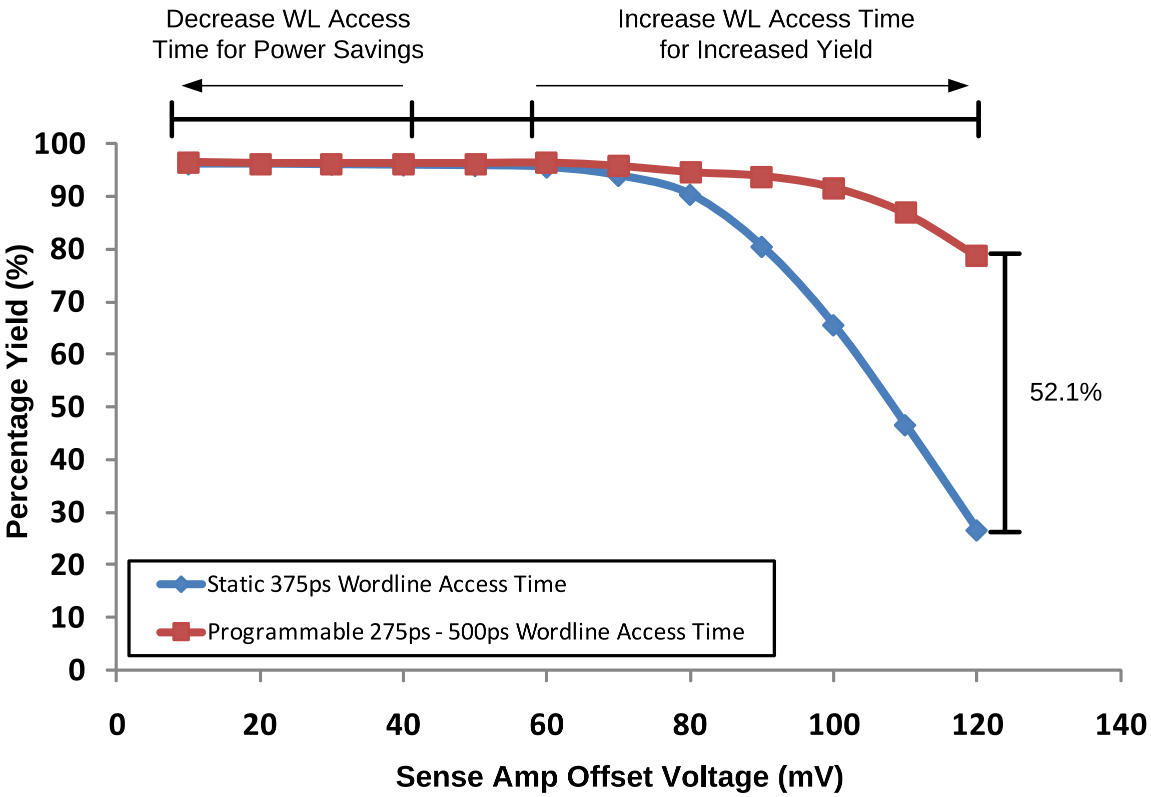


Figure 4.6: Cell Failure Reduction Using Programmable Timing

the case of the static wordline access time. Additionally, if there is less variability within the sense amplifier, and hence less differential bitline voltage is required for correctly sensing the cell, the SRAM can reduce its wordline access time to save power and increase DNM.

Chapter 5

Flexible SRAM Timing Control Architecture

A delay line based SRAM timing block has been implemented to show the ease of controllability of the SRAM's timing signals. Four signals are generated based off of the rising edge of an external input clock signal. These are the: Precharge (PRE), Wordline Enable (WLE), Sense Amplifier Enable (SAE), and Write Enable (WE) signals. PRE determines the duration of the precharge and evaluation phases within the SRAM, and ultimately the maximum clock frequency. WLE is used by the address decoder to enable the actual Wordline signal, WL. It is timed such that the WL is active inside PRE's evaluation phase. SAE is responsible for triggering the bitline's sense amplifier after a sufficient bitline differential voltage has been generated. SAE is only triggered on a read operation. Finally, WE is responsible for allowing the write driver access to the bitlines for discharging them when necessary. This is only available on a write operation.

As discussed in Chapter 4, the two most crucial timings are the wordline access time and sense amplifier enable window. As shown in Figure 5.1, the wordline access time, also known as wordline width, can be controlled by varying the arrival of the falling edge of the

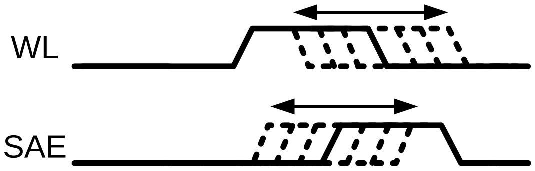


Figure 5.1: Variable wordline access time and sense amplifier enable windows

WL signal, and the sense amplifier enable window can be varied by the arrival time of the rising edge of the SAE signal. Since the WL signal must be contained within the signal PRE, only the falling edge of WL can be adjusted to increase the wordline access time. The SAE signal's falling edge has a constant arrival rate so it stays within the precharge's evaluation window. These concepts can be better understood with reference to the read operation timing signals shown in Figure 2.4(a). The main focus of the timing block implementation, as discussed in the remainder of this chapter, is on the controllability of the delay of these two edges.

5.1 Delay Line

Each of the timing block's output signals is constructed using a variable delay line based on a pulse generator [47]. The delay line structure is shown in Figure 5.2. The input signal is a common clock used for generating all of the timing block's outputs. The common clock signal is fed into a static delay line. For each output signal, the static delay line is branched off or "tapped" at two separate locations. These tapped signals are then fed through a variable delay element and then "AND"ed together to form the specific output signal. Figure 5.3 illustrates the functionality of the delay line.

Signal IN represents the common input clock signal. The delay from this point to node 'A' (t_{IN-A}) determines the low phase, or delay to, t_D, the output signal. Notice that since

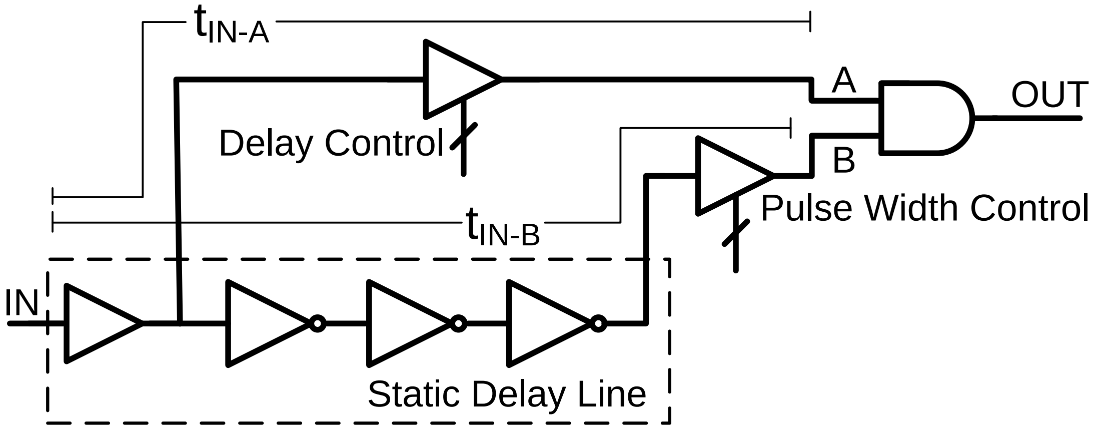


Figure 5.2: Pulse generator based variable delay line architecture

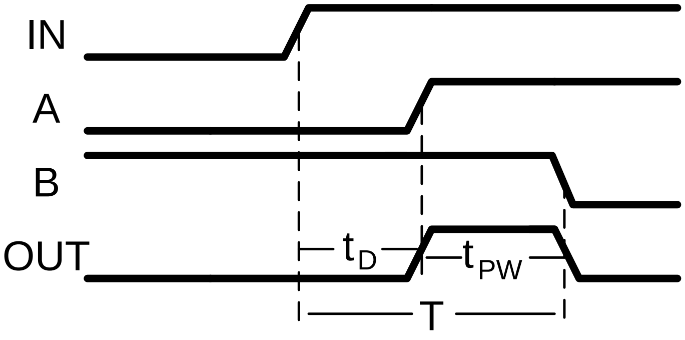


Figure 5.3: Pulse generator delay line timing diagram

an odd number of inverters is used to separate the tapping locations of node ‘A’ and node ‘B’, node ‘B’ is a delayed and inverted copy of node ‘A’. The delay from the input IN to node ‘B’ is designated by t_{IN-B}. The difference between these two delays is used to generate the high phase, or pulse of the output signal, t_{PW}. These are then fed into an AND gate to generate the output signal OUT. This is summarized in Equations 5.1, 5.2, and 5.3.

$$t_D = t_{IN-A} \tag{5.1}$$

$$t_{PW} = t_{IN-B}\ -\ t_{IN-A} \tag{5.2}$$

$$OUT = A\ AND\ B \tag{5.3}$$

By varying the t_{IN-A} delay, the low phase of the output signal can by varied, and by varying the t_{IN-B} delay, the high phase of the output signal can be varied. Since the delay signal is generated using only the rising edge of the input signal, the output signal is independent of the input signal’s frequency. This condition is valid while the period of the input signal is greater than period of the output signal generated by the delay line, $T_{in} > T_{out}$. This implementation strategy allows for full-speed testing while using a low-speed external input clock.

5.2 Digitally Controlled Delay Element

Variable delay is achieved with the digitally controlled delay element (DCDE) shown in Figure 5.4. It is able to achieve a given delay time varied by a fine-grain, sub-gate-delay step size based on a digital code.

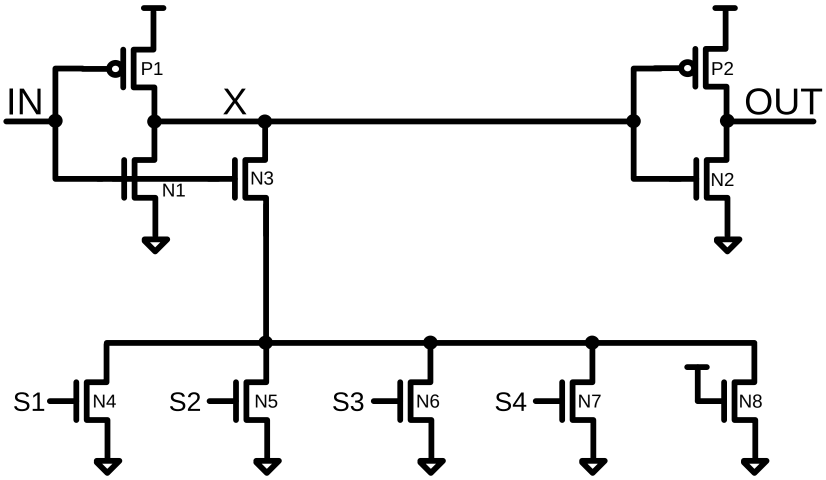


Figure 5.4: Digitally Controlled Delay Element

Transistors N1/P1, N2/P2 form two inverters to make up a standard delay element or buffer, and N3 to N8 provide the variable delay functionality by modulating the discharge resistance in the circuit's pulldown path. When the input signal, IN, is logic high, two path are available to discharge the charge stored at node 'X'. There is a fixed path through N1, and a current-starving variable path through N3. The gate of N8 is pulled up to V_{DD} to ensure that there is always a discharge path available through N3 to ground, and a digital code ($S_4S_3S_2S_1$) is applied to the gates of N7 down to N4 determining which transistors are turned on or off. This works to vary the effective resistance of the controlling transistors, and thereby determine the delay of the pulldown path. A drawback to using a DCDE that obtains its delay through a variable resistive network is that for a binary encoding scheme, it is susceptible to monotonicity errors [24]. A monotonicity error occurs when an incremental input code change results in an increase in delay rather than an decrease in delay, or visa versa. The issue of non-monotonicities can be avoided however, by using a thermometer encoding scheme rather than a binary one for issuing successive codes. An example comparison between successive binary and thermometer codes is shown in Table 5.1.

Table 5.1: Binary and Thermometer Code Example

Decimal Code	Binary	Thermometer
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0111
4	0100	1111

For a thermometer coding scheme, successive input codes are created by turning on one additional transistor at a time, where as the binary scheme uses a weighing scheme based upon bit position. This provides the added benefit of being able to size transistors N3 to N7, to provide a uniform step size between codes, as opposed to having a 1/x relationship between step sizes for a binary encoding scheme [26]. These benefits come at the cost of a reduction in available codes that can be applied to the DCDE (this will be addressed in the next section). Figure 5.5 provides a plot of the delay element's delay versus applied digital code. When V_{DD} is applied to all four control transistors, (Code = 1111), all of the transistors are on, and the delay element produces its smallest delay. Conversely, when GND is applied to the control transistors (Code = 0000), all of the control transistors are off, and the delay element produces its largest delay. Additionally, this DCDE is not susceptible to static power consumption, since there is never a static path directly connecting V_{DD} to GND. This is one of the significant drawbacks to the monotonic DCDE presented in [24] and [26]. The static power consumption for each of these DCDE is 340 μW and 79.2 μW respectively. Whereas, the maximum static power consumption for the presented DCDE is 3.3 nW. This is a reduction by five orders of magnitude.

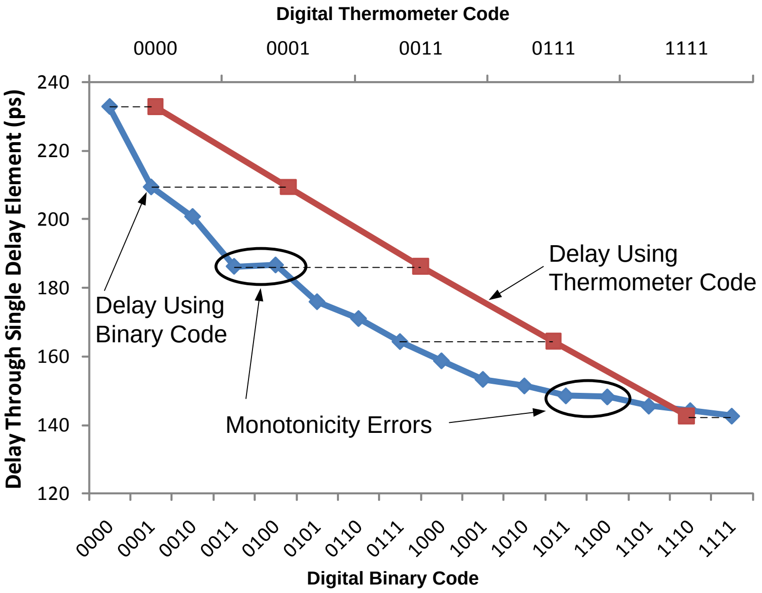


Figure 5.5: A comparison between binary and thermometer digital control codes applied to the same DCDE that exhibits a monotonicity error

5.3 Extended Range Delay Element

The DCDE discussed in the previous section is capable of providing a fine, sub-gate-delay step size between successive digital control codes. However, there is a limit to the range of its delays. By adding an additional control code transistor in the pull-down path, a binary encoding scheme would allow twice as many control codes; however, this would come at the cost of increasing the probability of monotonicity errors between successive codes. By using a thermometer encoding scheme, one additional transistor is required for each additional code, leading to a significant area overhead. The scheme shown in Figure 5.6 provides a coarse binary control scheme to supplement the fine thermometer control scheme of the DCDE.

The binary encoded control signal COARSE SELECT is used by a multiplexer to select either the original input signal, IN, or a copy of it delayed by a selected number

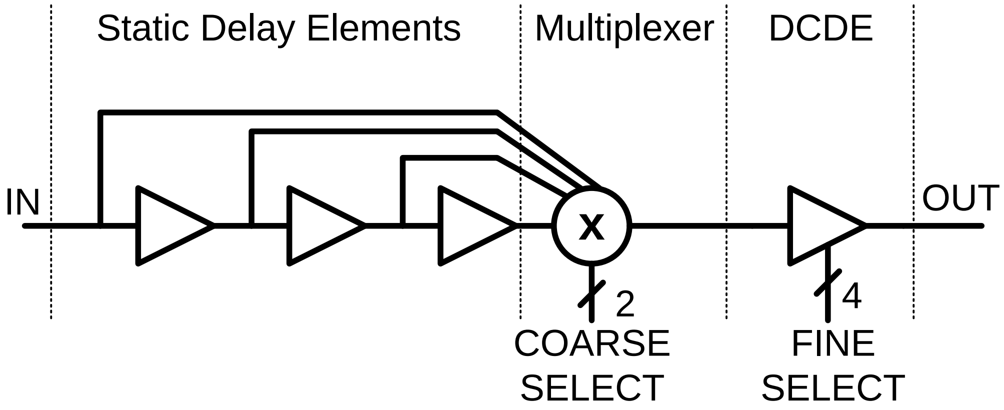


Figure 5.6: The extended range delay element uses a two stage delay element to select the delay, the first stage uses a two-bit binary code to select the coarse delay, and a four-bit thermometer code to select the fine delay

of static buffers. The signal is then fed into the DCDE from the previous section where the thermometer encoded control signal, FINE SELECT, determines the fine-granularity delay. This particular implementation uses a two-bit binary code coarse control signal in conjunction with a four-bit thermometer code fine control signal, yielding a total of 20 control codes for each extended range delay element.

5.4 Timing Block

The techniques of the preceding sections have been combined to create a delay-line based SRAM timing block, as shown in Figure 5.7. The timing block generates Precharge (PRE), Wordline Enable (WLE), Sense Amplifier Enable (SAE), and Write Enable (WE) signals based off of a single rising edge of an external input clock. For clarity, Figure 5.7 shows only the creation of the WLE and SAE signals. The PRE and WE signals are created in a similar manner only without the use of the extended range delay elements. The

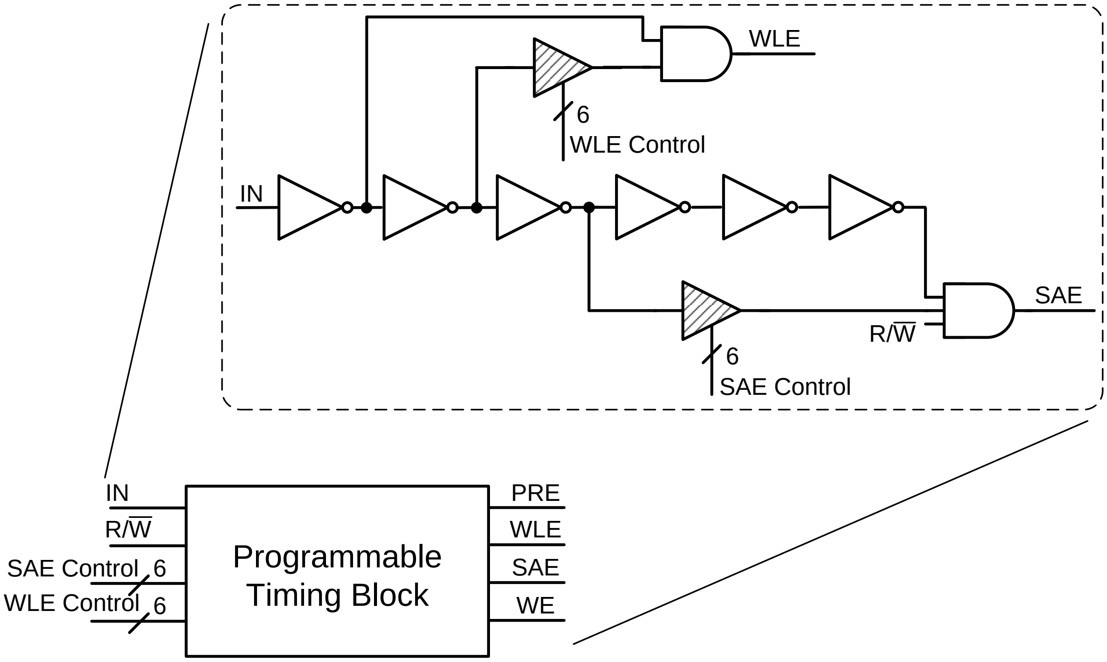


Figure 5.7: Programmable SRAM timing block

timing block could easily be extended to incorporate additional signaling specific to a particular SRAM implementation. The main features of the block include: 1). extended variable access time via variable WLE falling edge control, 2). extended sense amplifier enable window via variable SAE rising edge control, and 3). full-speed testing using a low-speed clock. These features are provided through the use of the extended range DCDE and the pulse generator delay-line architecture respectively. Additionally, fine-tuned digital controllability is provided for the propagation delay of each of the signal's rising and falling edges.

Chapter 6

Simulation Results & Test Chip

The SRAM timing block has been implemented in a 180 nm CMOS process to verify the functionality of the design. The test chip and design layout is shown in Figure 6.1. The timing block and the shift register storing the control codes is highlighted in the figure. In addition to the timing block, three other independent experiments will be conducted on the test chip; however, they are not related to the work described in this dissertation. To save on pins, control code data is shifted-in serially via a shift register. The complete timing block and shift register occupies an area of 185 μm x 160 μm. A 42-bit shift register was used to provide independent, fine-tuned controllability for the propagation delay of both the rising and falling edges of all the signals being generated. If only the SAE and WLE signals using the six-bit extended range delay elements were being controlled, only a 12-bit shift register would be required, resulting in an approximate 4x reduction in area for the shift register. Table 6.1 summarizes the test chip's characteristics.

Figure 6.2 shows the timing block control signals under nominal operating conditions during a read operation. All of the signals are generated based off of the input clock signal's rising edge. First, the wordline enable signal, WLE, rises and is sent to the address decoder triggering the proper wordline signal, WL, for the row in the memory array being accessed.

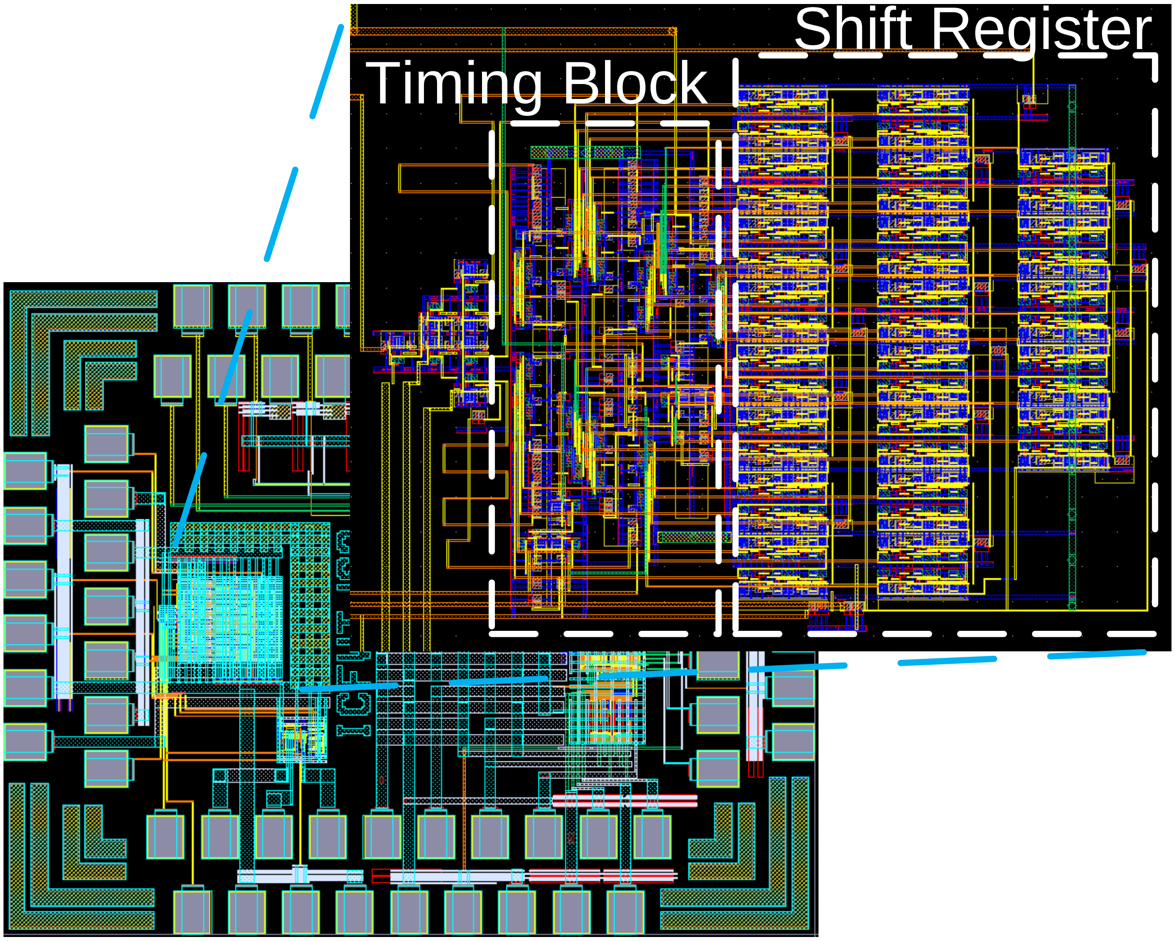


Figure 6.1: Test chip layout in 180 nm CMOS

Table 6.1: Test Chip Characteristics

Feature	Description
Technology	TSMC 180 nm CMOS 1P6M
Package	CFP80
Maximum Frequency	500 MHz
Area	185 μm x 160 μm
Supply Voltage	1.8 V
Special Features	Extended WLE and SAE edge control

The WLE signal is timed, based off of the address decoder delay, so that the WL signal would be generated just after the rising edge of the precharge signal, PRE. The signal PRE is used to differentiate between when the bitlines are pre-charging and when they are being used to evaluate a given memory cell. After the memory cell has been accessed for a sufficient amount of time, such that the sense amplifier's required differential voltage has been developed on the bitlines, the sense amplifier enable signal, SAE, is triggered. Controlling the wordline access time and the arrival time of the SAE rising edge determines the amount of differential voltage sensed by the sense amplifier. Once data has been read by the sense amplifier, the SAE and PRE signal can fall allowing the start of the next cycle. The timing signaling for a write operation is similar except that instead of issuing the SAE signal, the write enable signal, WE, is used, allowing the write driver access to the bitlines to write to the cell. The write enable window is bounded inside PRE's evaluation window.

Figure 6.3 shows a subset of the various WLE access times that can be achieved with the programmable timing. Four of the possible 20 codes are shown. For each of these codes, the four least significant bits (LSB) are set to zero (0000). These four bits represent the thermometer code portion of the control code. By setting them to zero, all of the control transistors in the delay element's pulldown path will be turned off and the fine granularity DCDE will experience the most delay. The two most significant bits (MSB) are stepped in a binary sequence. These are the binary control code bits for the coarse granularity extended range DCDE. As this portion of the code is swept from 00 to 11, the delay of the extended delay element decreases and in turn, the WLE access time decreases. This behavior follows that shown in Figure 5.1.

Figure 6.4 shows the various SAE propagation delays between the rising edge of the WL and the rising edge of the SAE, when the bitlines are developing the necessary differential voltage required to resolve the cell data. Under typical process corners, the SAE rising

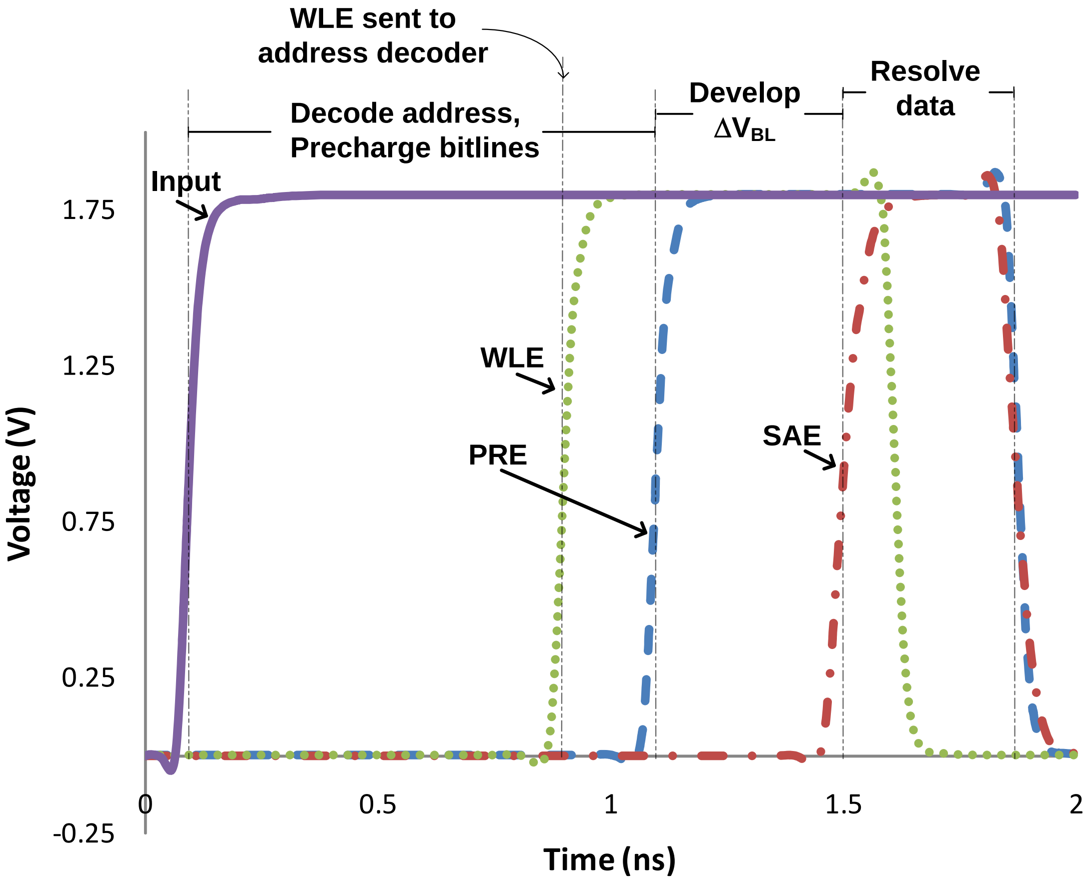


Figure 6.2: Simulated SRAM Read Operation

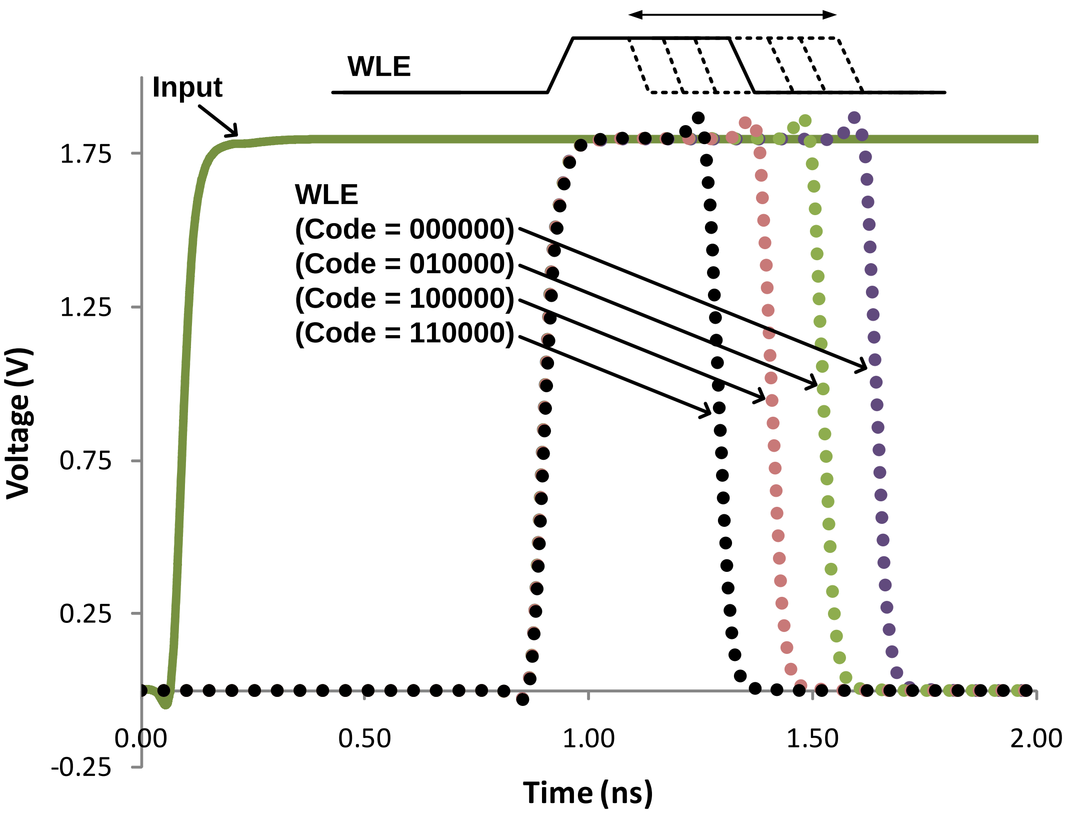


Figure 6.3: Simulated Wordline Access Time Programmability

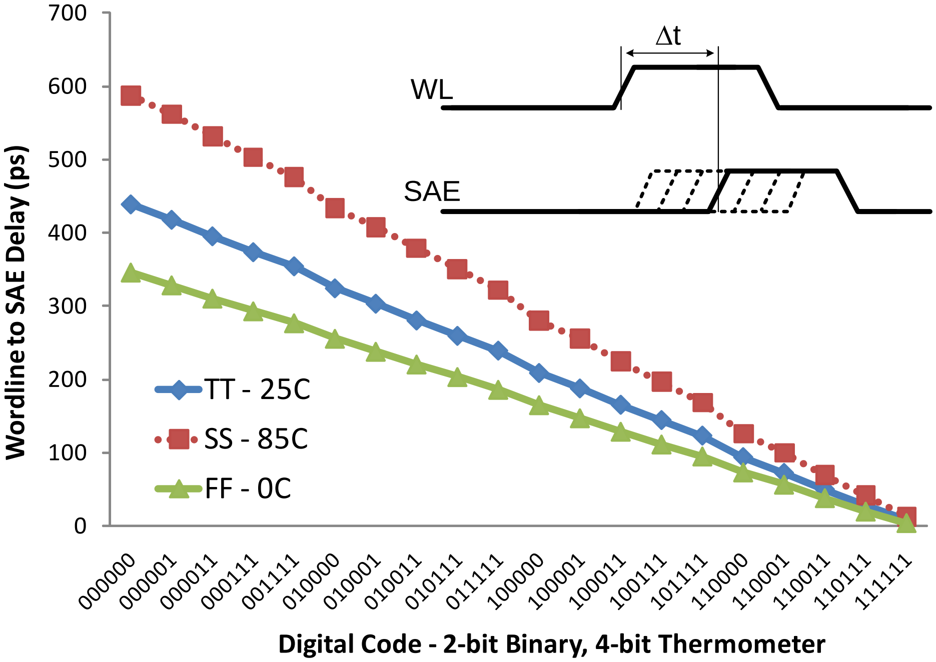


Figure 6.4: Monotonic WL to SAE propagation delays versus increasing control codes under process and temperature variation

edge can be varied by 430 ps across 20 control codes with an average step size of 22.6 ps between successive codes. Similarly, the falling edge of the WLE signal can be varied by 420 ps across 20 control codes with an average step size of 22 ps between successive codes. Table 6.2 presents simulated performance data for the SRAM timing block under different control codes and operating conditions.

Under typical operating conditions both the WLE and SAE signals have a range of over 400 ps with an average step size between codes of approximately 20 ps. Under systematic slow NMOS and PMOS process corners, SS, the range of both signals is over 550 ps with average step size of approximately 30 ps, and under fast corners, FF, the range is smaller, under 350 ps, with an average step size of approximately 18 ps. In both range and step size, these variations track the required read and write access times under the respective

Table 6.2: Simulated Performance Data Under Process and Temperature Variation

Signal Variability				
	SAE Rising Edge		WLE Falling Edge	
Conditions	Range (ps)	Avg. Step (ps)	Range (ps)	Avg. Step (ps)
TT - 25^{o}C	430.4	22.6	416.3	21.9
SS - 85^{o}C	574.8	30.2	553.9	29.1
FF - 0^{o}C	342.3	18.0	328.0	17.2

corners. If the entire SRAM is operating under the SS corner, the NMOS transistors in the SRAM cell will be weaker and hence more time will be required for read and write operations. Since the systematic variation also affects the timing block, in the SS corner condition, the timing block naturally provides more delay in its timing signals. The same is true under the FF corner; the timing block is able to provide smaller delays when less time is required for correct functional operation.

Chapter 7

Conclusion

Embedded memories are fundamental building blocks of modern SOCs. As CMOS processes scale deep into the sub-micron regime, the accompanying increase in process variability and aging leads to a significant increase in the soft failure rate and in turn yield loss. This thesis investigates the ways in which SRAM timing can be used to improve transient based SRAM figures of merit, and reduce the soft failure rate. Timing correctable soft failures include: operational read failures, cell stability failures, and power envelope failures. This work has shown that post-fabrication programmability of the wordline access time and sense amplifier enable window provides the designer with the ability to optimize the SRAM timing to compensate for the process variability on a per die basis. A delay line based SRAM timing block with digitally programmable timing signals has been implemented in a 180 nm standard CMOS process to demonstrate the monotonic controllability of its timing parameters. The wordline access time and sense amplifier enable window can each be varied monotonically by more than 400 ps under typical operating conditions over a set of 20 digital control codes. This timing block implementation can be used to first characterize the specific soft failure rate of an SRAM array, and then optimize the timing so as to maximize the yield.

Bibliography

[1] M. H. Abu-Rahma, M. Anis, and S. S. Yoon. Reducing SRAM power using fine-grained wordline pulsewidth control. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 18(3):356–364, March 2010. 52

[2] K. Agarwal and S. Nassif. The impact of random device variation on SRAM cell stability in sub-90-nm CMOS technologies. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(1):86–97, Jan. 2008. 2, 26

[3] B.S. Amrutur and M.A. Horowitz. A replica technique for wordline and sense control in low-power SRAM's. *Solid-State Circuits, IEEE Journal of*, 33(8):1208–1219, Aug 1998. 3, 17, 19

[4] K. Ando, K. Higeta, Y. Fujimura, K. Mori, M. Nakayama, H. Nambu, K. Miyamoto, and K. Yamaguchi. A 0.9-ns-access, 700-MHz SRAM macro using a configurable organization technique with an automatic timing adjuster. In *VLSI Circuits, 1998. Digest of Technical Papers. 1998 Symposium on*, pages 182 –183, 11-13 1998. 19

[5] R.C. Baumann. Radiation-induced soft errors in advanced semiconductor technologies. *Device and Materials Reliability, IEEE Transactions on*, 5(3):305 – 316, Sept. 2005. 4

[6] C.J. Brennan, S. Eustis, J. Goss, A. Humphrey, M. Ouellette, J. Rowland, and M. Fragano. BIST controlled variable sense amp timing for 90 nm embedded SRAM.

In *Custom Integrated Circuits Conference, 2004. Proceedings of the IEEE 2004*, pages 345–348, Oct. 2004. 3, 20

[7] D. Burnett, K. Erington, C. Subramanian, and K. Baker. Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits. In *VLSI Technology, 1994. Digest of Technical Papers. 1994 Symposium on*, pages 15 –16, 7-9 1994. 29

[8] E.H. Cannon, A. KleinOsowski, R. Kanj, D.D. Reinhardt, and R.V. Joshi. The impact of aging effects and manufacturing variation on SRAM soft-error rate. *Device and Materials Reliability, IEEE Transactions on*, 8(1):145 –152, March 2008. 36

[9] Meng-Fan Chang, Sue-Meng Yang, and Kung-Ting Chen. Wide VDD embedded asynchronous SRAM with dual-mode self-timed technique for dynamic voltage systems. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 56(8):1657 –1667, Aug. 2009. 3

[10] G. Chen, K.Y. Chuah, M.F. Li, D.S.H. Chan, C.H. Ang, J.Z. Zheng, Y. Jin, and D.L. Kwong. Dynamic NBTI of PMOS transistors and its impact on device lifetime. In *Reliability Physics Symposium Proceedings, 2003. 41st Annual. 2003 IEEE International*, pages 196 – 202, 2003.

[11] R. Degraeve, M. Aoulaiche, B. Kaczer, P. Roussel, T. Kauerauf, S. Sahhaf, and G. Groeseneken. Review of reliability issues in high-κ/metal gate stacks. In *Physical and Failure Analysis of Integrated Circuits, 2008. IPFA 2008. 15th International Symposium on the*, pages 1 –6, 7-11 2008. 41

[12] T. Grasser and B. Kaczer. Evidence that two tightly coupled mechanisms are responsible for negative bias temperature instability in oxynitride MOSFETs. *Electron Devices, IEEE Transactions on*, 56(5):1056 –1062, May 2009. 39

[13] F. Hamzaoglu, K. Zhang, Yih Wang, H.J. Ahn, U. Bhattacharya, Zhanping Chen, Yong-Gee Ng, A. Pavlov, K. Smits, and M. Bohr. A 3.8 GHz 153 Mb SRAM design with dynamic stability enhancement and leakage reduction in 45 nm high-κ metal gate CMOS technology. *Solid-State Circuits, IEEE Journal of*, 44(1):148 –154, Jan. 2009. x, 3, 31

[14] J. L. Hennessy and D. A. Patternson. *Computer Architecture - A Quantitative Approach*, page 967. Morgan Kaufmann, 4th edition. 13

[15] Chih-Sheng Hou, Jin-Fu Li, and Che-Wei Chou. Test and repair scheduling for built-in self-repair RAMs in SOCs. In *Electronic Design, Test and Application, 2010. DELTA '10. Fifth IEEE International Symposium on*, pages 3–7, Jan. 2010.

[16] D. Ielmini, M. Manigrasso, F. Gattel, and M.G. Valentini. A new NBTI model based on hole trapping and structural relaxation in MOS dielectrics. *Electron Devices, IEEE Transactions on*, 56(9):1943 –1952, Sept. 2009. 39

[17] J. Jayabalan and J. Povazanec. Integration of SRAM redundancy into production test. In *Test Conference, 2002. Proceedings. International*, pages 187–193, 2002. 1

[18] J. Keane, Xiaofei Wang, D. Persaud, and C.H. Kim. An all-in-one silicon odometer for separately monitoring HCI, BTI, and TDDB. *Solid-State Circuits, IEEE Journal of*, 45(4):817 –829, April 2010. xi, 36, 37, 38, 39, 40

[19] K.J. Kuhn. Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pages 471 –474, 10-12 2007. 2

[20] Y. Kunitake, T. Sato, and H. Yasuura. Signal probability control for relieving NBTI in SRAM cells. In *Quality Electronic Design (ISQED), 2010 11th International Symposium on*, pages 660 –666, 2010.

[21] Ya-Chun Lai and Shi-Yu Huang. Robust SRAM design via BIST-assisted timing-tracking (BATT). *Solid-State Circuits, IEEE Journal of*, 44(2):642–649, Feb. 2009. x, 3, 20

[22] J.-F. Li, T.-W. Tseng, and C.-S. Hou. Reliability-enhancement and self-repair schemes for SRAMs with static and dynamic faults. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, (99):1 –1, Aug. 2009.

[23] Shyue-Kung Lu, Chun-Lin Yang, Yuang-Cheng Hsiao, and Cheng-Wen Wu. Efficient BISR techniques for embedded memories considering cluster faults. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 18(2):184 –193, Feb. 2010.

[24] M. Maymandi-Nejad and M. Sachdev. A monotonic digitally controlled delay element. *Solid-State Circuits, IEEE Journal of*, 40(11):2212–2219, Nov. 2005. 58, 59

[25] 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. *Solid-State Circuits, IEEE Journal of*, 33(11):1650 –1658, Nov 1998. 19

[26] Muhammad A. Nummer. *Precise Timing of Digital Signals: Circuits and Applications*. PhD thesis, University of Waterloo, 2007. 59

[27] K. Osada, Jinuk Luke Shin, M. Khan, Y. Liou, K. Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi. Universal-Vdd 0.65-2.0-V 32-kB cache using a voltage-adapted timing-generation scheme and a lithographically symmetrical cell. *Solid-State Circuits, IEEE Journal of*, 36(11):1738 –1744, Nov 2001. 19

[28] Sangwoo Pae, J. Maiz, C. Prasad, and B. Woolery. Effect of BTI degradation on transistor variability in advanced semiconductor technologies. *Device and Materials Reliability, IEEE Transactions on*, 8(3):519 –525, Sept. 2008. x, 30

[29] Liang-Teck Pang and B. Nikolic. Measurement and analysis of variability in 45 nm strained-Si CMOS technology. In *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, pages 129–132, Sept. 2008. 2

[30] A. Pavlov and M. Sachdev. *CMOS SRAM - Circuit Design and Parametric Test in Nano-scaled Technologies*. Springer Science, 2008. xi, 32, 33, 34, 35

[31] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. Matching properties of MOS transistors. *Solid-State Circuits, IEEE Journal of*, 24(5):1433–1439, Oct 1989. 2, 30

[32] H. Pilo, C. Barwin, G. Braceras, C. Browning, S. Lamphier, and F. Towler. An SRAM design in 65 nm technology node featuring read and write-assist circuits to expand operating voltage. *Solid-State Circuits, IEEE Journal of*, 42(4):813–819, April 2007. x, 2, 3

[33] J.M Rabaey, A. Chandrakasan, and B Nikolic. *Digital Integrated Circuits: A Design Perspective*. Prentice-Hall, Inc., 2003. 11

[34] R. Radojcic, D. Perry, and M. Nakamoto. Design for manufacturability for fabless manufactuers. *Solid-State Circuits Magazine, IEEE*, 1(3):24 –33, Summer 2009. 30

[35] David J. Rennie, Tahseen Shakir, and Manoj Sachdev. Design challenges in nanometric embedded memories. In *Signals, Circuits and Systems (SCS), 2009 3rd International Conference on*, pages 1–8, Nov. 2009. x, 2, 6, 14, 45

[36] S. Rusu, S. Tam, H. Muljono, J. Stinson, D. Ayers, J. Chang, R. Varada, M. Ratta, S. Kottapalli, and S. Vora. A 45 nm 8-core enterprise Xeon processor. *Solid-State Circuits, IEEE Journal of*, 45(1):7 –14, Jan. 2010. 19

[37] S.E. Schuster, B.A. Chappell, R.L. Franch, P.F. Greier, S.P. Klepner, F.J. Lai, P.W. Cook, R.A. Lipa, R.J. Perry, W.F. Pokorny, and M.A. Roberge. A 15 ns CMOS 64K RAM. *Solid-State Circuits, IEEE Journal of*, 21(5):704–712, Oct 1986. 17

[38] E. Seevinck, F.J. List, and J. Lohstroh. Static-noise margin analysis of MOS SRAM cells. *Solid-State Circuits, IEEE Journal of*, 22(5):748–754, Oct 1987. 22, 31

[39] M. Sharifkhani and M. Sachdev. SRAM cell stability: A dynamic perspective. *Solid-State Circuits, IEEE Journal of*, 44(2):609–619, Feb. 2009. xi, 25, 48, 49

[40] A. Sharma. *Advanced Semiconductor Memories (Architectures, Designs, and Applications)*. Wiley, 2002. 17

[41] H. Singh and H. Mahmoodi. Analysis of SRAM reliability under combined effect of NBTI, process and temperature variations in nano-scale CMOS. In *Future Information Technology (FutureTech), 2010 5th International Conference on*, pages 1 –4, 21-23 2010. 43

[42] P.A. Stolk, H.P. Tuinhout, R. Duffy, E. Augendre, L.P. Bellefroid, M.J.B. Bolt, J. Croon, C.J.J. Dachs, F.R.J. Huisman, A.J. Moonen, Y.V. Ponomarev, R.F.M. Roes, M. Da Rold, E. Seevinck, K.N. Sreerambhatla, R. Surdeanu, R.M.D.A. Velghe, M. Vertregt, M.N. Webster, N.K.J. van Winkelhoff, and A.T.A. Zegers-Van Duijnhoven. CMOS device optimization for mixed-signal technologies. In *Electron Devices Meeting, 2001. IEDM Technical Digest. International*, pages 10.2.1 –10.2.4, 2001. 32

[43] S. Tachibana, H. Higuchi, K. Takasugi, K. Sasaki, T. Yamanaka, and Y. Nakagome. A 2.6-ns wave-pipelined CMOS SRAM with dual-sensing-latch circuits. *Solid-State Circuits, IEEE Journal of*, 30(4):487–490, April 1995. 17

[44] R. Vattikonda, Wenping Wang, and Yu Cao. Modeling and minimization of PMOS NBTI effect for robust nanometer design. In *Design Automation Conference, 2006 43rd ACM/IEEE*, pages 1047 –1052, 2006. xi, 41, 42

[45] V. Wang, K. Agarwal, S.R. Nassif, K.J. Nowka, and D. Markovic. A simplified design model for random process variability. *Semiconductor Manufacturing, IEEE Transactions on*, 22(1):12–21, Feb. 2009. 26

[46] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y. Ng, L. Wei, Y. Zhang, K. Zhang, and M. Bohr. A 4.0 GHz 291 Mb voltage-scalable SRAM design in 32 nm high-κ metal-gate CMOS with integrated power management. In *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pages 456 –457,457a, 8-12 2009. 1, 19

[47] Neil H. Weste and David Harris. *CMOS VLSI Design - A Circuits and Systems Perspective*. Addison Wesley, 2005. 55

[48] Fu-Liang Yang, Jiunn-Ren Hwang, and Yiming Li. Electrical characteristic fluctuations in sub 45nm CMOS devices. In *Custom Integrated Circuits Conference, 2006. CICC '06. IEEE*, pages 691 –694, 10-13 2006. 2

[49] Shyh-Chyi Yang, Hao-I Yang, Ching-Te Chuang, and Wei Hwang. Timing control degradation and NBTI/PBTI tolerant design for write-replica circuit in nanoscale CMOS SRAM. In *VLSI Design, Automation and Test, 2009. VLSI-DAT '09. International Symposium on*, pages 162 –165, 2009. xi, 44

[50] Bin Zhang, A. Arapostathis, S. Nassif, and M. Orshansky. Analytical modeling of SRAM dynamic stability. In *Computer-Aided Design, 2006. ICCAD '06. IEEE/ACM International Conference on*, pages 315–322, Nov. 2006. 24