A DUAL COLUMN, REPLICA BITLINE DELAY TECHNIQUE USING STOCHASTIC CURRENT PROCESSING FOR A PROCESS VARIATION TOLERANT, LOW POWER SRAM

By

ANIKET SUHAS KULKARNI

Bachelor of Engineering in Electronics and

Telecommunication

University of Mumbai

Vashi, Maharashtra, INDIA

2011

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2015

A DUAL COLUMN, REPLICA BITLINE DELAY TECHNIQUE USING STOCHASTIC CURRENT PROCESSING FOR A PROCESS VARIATION TOLERANT, LOW POWER SRAM

Thesis Approved:

Dr. Chriswell Hutchens

Thesis Adviser

Dr. James Stine

Dr. R Ramakumar

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my adviser Dr. Chriswell Hutchens for suggesting a topic based on memory design. It was certainly not possible to complete my thesis without his guidance. He has been more of a mentor than an adviser to me. I would also like to thank my committee members Dr. James Stine and Dr. R Ramakumar.

I thank my family members, my mother Mrs. Manisha K., my father Mr. Suhas K., my younger sister Ms. Ankita K. and my girlfriend Ms. Mithila G. for being emotionally and mentally supportive towards me. I dedicate this thesis to my Grandfather Dattatray V Kulkarni, because of his blessings I am what I am today.

I would like to thank my MSVLSI lab members Dr. An Guanglie, Dr. Ran Liao, Mr. Rehan Ahmed, Mr. Kanishka De, Mr. Cheng hao, Mr. Arpit Rao for always helping me with my technical doubts, sharing knowledge and motivating me to achieve my goals. A special thanks to my friends Mr. Heramb J. and Ms. Vandana V. who helped me with Matlab programming. I would like to thank Ms. Samira Ataei for helping me understand few basic and important aspects of SRAM memory design.

I also thank my friends, Subodh G., Arjun K., Vidisha P., Trupti K., and Ankita C. without whom my journey to masters wouldn't have been so exciting. Last but not the least, I thank Oklahoma State University for giving me such a wonderful opportunity to pursue my Master's Degree.

Name: ANIKET S KULKARNI

Date of Degree: DECEMBER, 2015

Title of Study: A DUAL COLUMN, REPLICA BITLINE DELAY TECHNIQUE USING STOCHASTIC CURRENT PROCESSING FOR A PROCESS VARIATION TOLERANT, LOW POWER SRAM

Major Field: ELECTRICAL ENGINEERING

Abstract: SRAM (Static Random Access Memory) design has become the critical and important block in processing ICs with the highest bandwidth power rationed memories taking the business lead. As industry attempts to maintain Moore's law by shrinking the device size, we are facing greater issues with the variability due to random doping fluctuation in devices. This variation compels engineers to design for worst case conditions which leads to inefficient memory model, which make it difficult to stand in the business race. However, a smart design can lead to less variation and "exact" memory parametric prediction to achieve high performance, low power and maximum yield designs. Since, random variation today is more dominant, we consider the application of the central limit theorem to control memory read timing across PVT (Process Voltage Temperature) corners. A statistical read timing is developed for a SRAM memory bank. In the thesis two dummy columns, each at extreme end of the memory bank, are used to implement the statistical memory bank model. By combining Monte-Carlo analysis using cadence virtuoso, and PDK data for the CMOS process (IBM 7RF), an analytically memory timing model is verified. Our major goal is to improve yield across all memory banks in all die across all the wafers; slow-slow (SS), typical-typical (TT) and fast-fast (FF).A smart stochastic/statistical approach is used in the thesis to predict exact parametric yield parameters with less variation to design accurate memory system which gives high performance, low power and maximum yield across all PVT corners to keep you ahead in the memory business. The memory design is compared to the conventional self-timed replica architecture using coefficient of variance of a reference current generated using dummy column. The proposed architecture was able to achieve 62 percent across the process improved accuracy in reference current and sense amplifier firing variation. Proposed architecture looks promising for future node technologies where statistical variability and its impact in subthreshold region is more dominant.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

Table LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION TO SRAM

1.1 INTRODUCTION

Recently, the semiconductor market is showing great interest in smart nodes for infrastructure, health monitoring, smart wearable devices for personal use and also in medical implants. According to the report of Organization for Economic Co-operation and Development, by the year 2022 we would have 25 to 50 personal devices connected to Internet [1] In an Ideal world people expect their devices to require near zero power, high frequency, fast computing, low area and infinite memory. We as engineers strive to achieve these goals, but as of now, it is all but impossible. To develop such applications, we need high performance and low power designs. Processing speed generally depends on how fast cache memory can be accessed [2]. It is challenging to achieve both a fast and low power cache memory, the reason will become clear later as we proceed. Cache is usually composed of Static Random Access Memory (SRAM) designs. RF communication and low power chip uses digital signal processing and then transmits data to save power consumption. Digital Signal processing, DFT and FFT require data to be stored and this means large on chip memory. As a result SRAM design has become a critical and important block or sub block in system design at all levels.

Fig. 1.1 shows the pyramid of memory hierarchy. A processor fetches the data/opcode from the memory to execute the command. This “data” should have fast access time. Mostly, this data is held in L0 and L1. L0 has register files used to save temporary data used for processing. L1 cache provides data to L0 registers where L0, L1 and L2 are volatile memories. SRAM speed defines the processor speed, and a large on-chip size of L1 and L2 with fast access time would set the performance of processor.

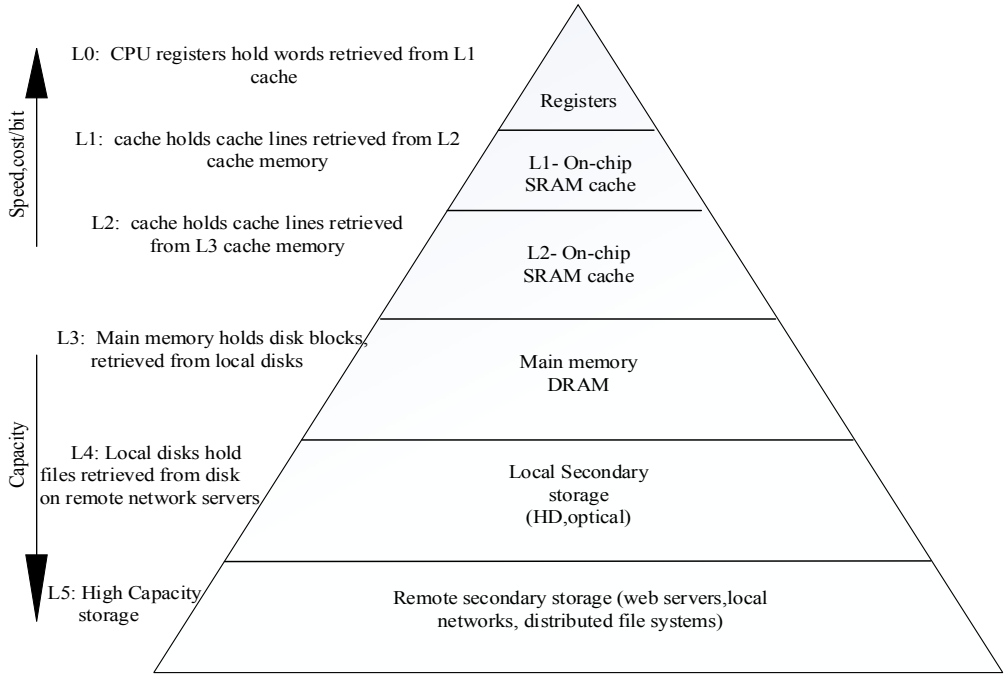


Fig. 1.1 Memory Hierarchy [3]

As shown in above pyramid, the higher level has large capacity and less speed/bits, whereas the lower level has increase in speed/bit at a higher cost. Program is executed at constant bandwidth from high level to low level. The cache is stalled when data is not ready at higher level to transfer. The intent is to maintain a near constant bandwidth across the hierarchy. This is due to data access speed/ bit differing at each stage. When the program is executed, it copies the code from higher level memory and saves it to the lower stage memory. This improves the processor's performance. Overall, designing a fast cache memory results in setting processors/applications performance.

1.2 MOTIVATION

SRAM design has become a critical and important block in processing ICs with the highest bandwidth power rationed memories taking the business lead. As industry attempts to maintain Moore's law by shrinking the device size, we are facing greater issues with the variability due to random doping fluctuation in devices [4]. Variation compels engineers to design either for worst case conditions which leads to a slow memory and good yield or considerate condition which leads to fast memory and poor yield. An inefficient memory model makes it difficult to stand in the business race. This challenging and undefined marginal issues has motivated me to come up with a solution where our major goal is to design accurate memory system which gives high performance, low power and maximum yield across all die and across all the wafers; slow-slow (SS), typical-typical (TT) and fast-fast (FF) to be ahead in the memory business.

1.3 SRAM ARCHITECTURE AND OPERATION

Before going into details of SRAM designing and issues related to design, let us discuss the basic block diagram of an SRAM and its working. Below figure 1.2 shows basic block diagram of an SRAM.

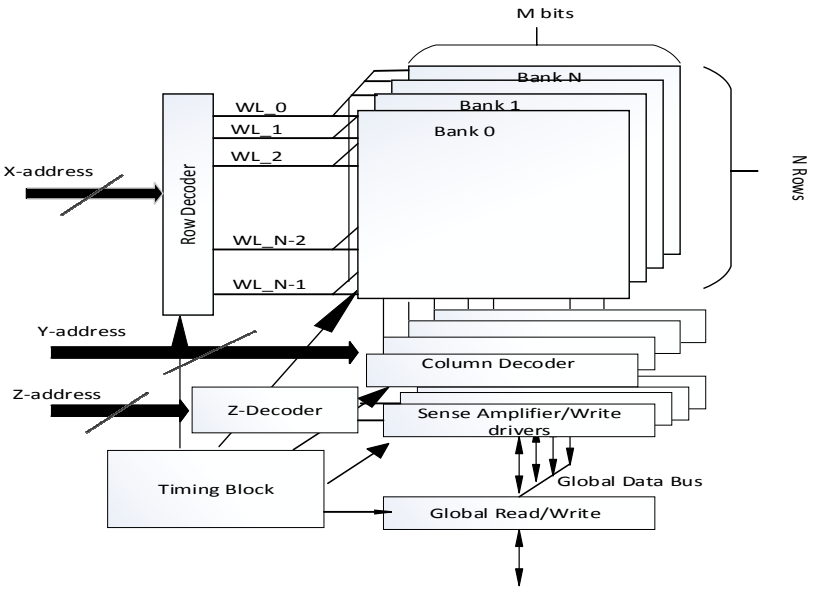


Fig. 1.2 Block diagram of SRAM [3]

An address is generated to access the memory location. A memory is accessed either to read or write depending on the control signals generated while processing. The presented address is decoded by the address decoders. Row is accessed by X-address decoder, column is accessed by Y-address decoder and Bank is selected by Z-address decoder. Write drivers are used to write the data (logic 0 or logic 1) into the memory. Sense Amplifiers are used to sense the bit line voltage difference and amplify it to the extreme ends to VDD and VSS. The faster the sense amplifier, the better the memory performance, provided power remains less than a 1 to 2 watts [5, 6]. Assuming there are 128 rows or 128 cells in a column and 32 cells in a row or 32 columns in a bank, a bank will be 128 x 32 bit memory or a 4096 bits (4Kb) memory. Four such banks to make a cache memory of 16Kb. The decode equation is known by $Y = 2^n$ where Y is number of outputs and n is number of inputs. Thus we need 7:128, 5:32 and 2:4 decoders to decode the address of a 4Kb memory. There are many different ways to access a cache which is beyond the discussion of this thesis.

1.4 CONVENTIONAL 6T SRAM

A conventional 6T SRAM cell is shown in figure 1.3 It consists of 6 transistors (6T) M1 to M6. M1-M3 and M2-M4 forms a cross coupled inverter pair. M5-M6 are pass transistors/access transistors. BL (Bit line) and BLB (Bit Line Bar) are used to read and write data on activation of WL (Word Line) signal. C_{BL} and C_{BLB} are the parasitic capacitances associated with the bit lines. M3 and M4 are pull up (PU) transistors whereas M1 and M2 are pull down (PD) transistors. WL controls the access of the node (Q0 and Q1) voltages to the bit line. Bit line voltages defines if the operation is read or write. As a memory designer, we design the cell to provide a non-destructive read operation and a reliable write operation which always conflicts in transistor size designing [2]. Every memory cell will have either logic ‘0’ or ‘1’ stored in it.

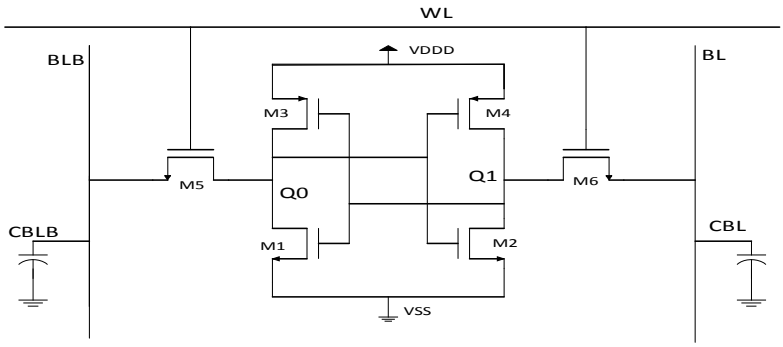


Fig. 1.3 Conventional 6T SRAM CELL

1.5 READ AND WRITE OPERATION

1.5.1 Read Operation

Before initiating a read operation, BL and BLB are pre-charged to VDD. Once the capacitance CBL and CBLB are fully charged, WL is activated and bit lines can access the node data. Bit line connected to logic '0' starts discharging whereas the other bit line does not discharge. With the bit line difference sufficient to take a valid decision as to a '1' and '0', the control unit triggers the sense amplifier. The sense amp in return amplifies the difference, i.e. 50mV to extreme values of the supply voltage in a bounded time.

Figure 2.2 shows a 6T CMOS SRAM cell during read operation. The bit line voltages V_{BL} remains at the pre-charge level, practically this bit line may discharge due to leakage current flowing through NMOS and the complementary bit line voltage V_{BLB} is discharged through transistors M5 and M1. Total leakage current can be as large as Ioff times the number of cells in a column. Transistor M5 and M1 forms a voltage divider which develops a ΔV potential at node Q0. This node potential should be less than the trip point of the inverter pair M2-M4, if it is greater than the trip point, it will result in destructive read operation. In order to keep ΔV below trip point of M2-M4, M1 should be stronger than M5. Strong M1 draws more current and thus ΔV can kept low below trip point. How this may

contradict in the write operation which will be covered in next section. ΔV Depends on the cell ratio which is written as below $\beta = CR = (\frac{W1}{L1})/(\frac{W5}{L5})$.

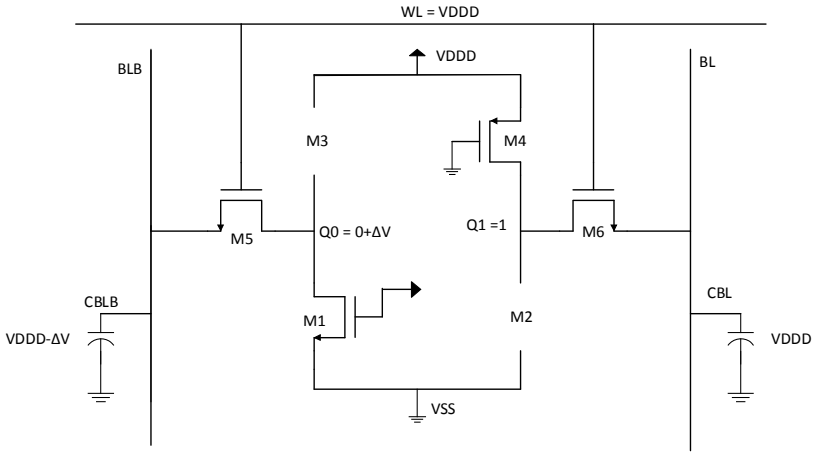


Fig.1.4 6T CMOS SRAM Read Cell

Since the cell is symmetrical, the CR is same for M2 – M6. Larger ratios provides higher read current I*read* (with minimum length geometry) which provide high speed memory and better stability but at the expense of the large area. An optimized cell ratio can provides low cell area, adequate stability and optimal operating speeds to achieve yield and position in the business lead.

1.5.2 Write Operation

In write operation the data to be written is loaded on BL with its compliment loaded on BLB. When data is loaded, WL is activated and now the cell is accessed to write the data. Suppose we have to write '0' on a node storing '1'. BLB is loaded to '1' and compliment '0' on BL. When WL is activated, there will be ΔV generated at Q0 node. If this node potential is above the trip point of M2-M4 then Q1 will become '0' on conduction on M2.However, ΔV is set by read margin and this contradicts in SRAM cell design. The

write structure is share by all column cells in addition to driving BL and BLB and as a result will be both large and accurate.

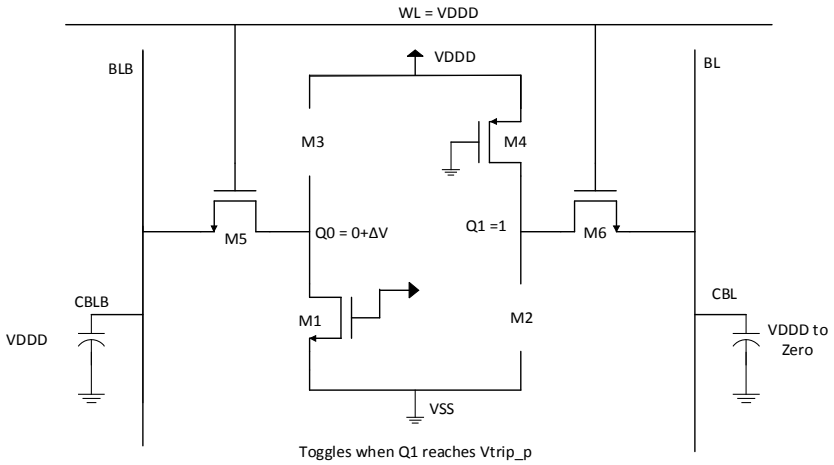


Fig.1.5 6T CMOS SRAM Write Cell

For a reliable write operation pass transistor M5 and M6 should be strong than M3 and M4 respectively. Once Q1 is pulled down below the trip point of M1-M3, M3 turns on and pull node Q0 to logic '1' which then pulls down Q1 to logic '0'. A positive feedback is used in the write operation. Pull up ratio $PR = \left(\frac{W4}{L4}\right) / \left(\frac{W6}{l6}\right)$ define the write margin. Since the cell is symmetrical, PR is same for M3 and M5. Thus a better cell design can be done by keeping Strength (PMOS PU) < Strength (NMOS Access/Pass) < Strength (NMOS PD) [2].

1.6 CELL DESIGN FAILUERS

There are two types of failures, catastrophic and parametric failure [7]. This report sticks to only parametric failures in memory cell. A failure in SRAM means unreliable write or incorrect read operation. If a single memory cell fail's then a whole row and column fails impacting the yield [8]. To improve yield, additional columns are typically added for error correction. Here, yield is calculated with probability of word failure. Below are five different cell failures which should be considered by memory designer. Cell failure in a word line

should be less than redundant columns present in a memory cell for error correction [8]. Following are few important cell failures.

i) Read Failure: A read failure takes place when WL is activated and the node storing logic '0' rises to ΔV due to a voltage divider formed due to access and PD transistor. If this ΔV is sufficient enough to flip the data stored, then there is a read failure. This can be a significant error, but should be addressed to maintain minimum noise margin.

ii) Write Failure: Writing depends on how fast the node having logic '1' discharges below the trip point (only when opposite data is written on the node). If this discharging is not fast enough and the node potential does not reach to the trip point before the WL is deactivated, then there is a write failure. Due to the size of the write amplifier this is a low probability error.

iii) Access Failure: Both the bit lines BL and BLB are connected to sense amplifier. Sense amplifier is used to sense ΔV between the bit lines and amplify the output to extreme supply rails using regenerative cross coupled pairs. Every sense amplifier has inherent input offset voltage to which ΔV has to overcome. If the timing analysis to fire sense amplifier is not done correctly then there is a chance we amplify the inherent input offset voltage which leads to incorrect data read. This is called access failure. Access failure is the major failure among other failures which affects the silicon yield [9], as a part of thesis we will be discussing access failure in detail in further chapters. This is a significant error.

iv) Data dependent bit line leakage: A row and column is selected to access the memory cell. The worst case scenario for data dependency can be every other cell in the column which is not read had opposite data saved on the node. The bit line which has bit '1' should not droop, but due to leakage this voltage bit line may also droop. An early decision would result in access failure. There is a limit on number of cells connected

to the bit line which depends on Ion/Ioff ratio, this is called data dependency error. This is a significant error.

v) Hold Failure: A low VDD is applied when memory is in sleep mode. Due to leakage of pulldown NMOS transistor, stored logic '1' could drop below trip point of an inverter pair to flip the cell data. This causes hold failure. This can be a significant error.

vi) Half Select: When a cell data is written, WL is set to logic high '1', this exposes the node data off all unselected column cell present in the row which is accessed by the bit lines. If the cell is not properly designed, then it can flip the data resulting into change in data. This is called as half select. It is very much similar to read failure but occurs in the write operation. This can be a significant error.

1.7 DESIGN CHART

There are many different parameter's a memory designer must consider. Most of the parameters ultimately concentrate on balancing Power, Performance and Area (PPA) of a memory cell or sense amplifier. This will eventually set PPA for whole memory bank. Before proceeding further the reader is assumed to have knowledge of noise margins of SRAM and parameters calculated using butterfly and N- Curve simulations. A detailed explanation of static, R/W (read and write) margin for memory cell is given in [10-15]. A new approach using N-curve simulation is given in [16, 17] which shows the importance of current details in calculating R/W noise margins. Both butterfly and N-Curve simulations were performed for various width and length transistor to show the effect of noise margins. Monte-Carlo simulations were performed to check the performance of individual cell across the PVT corners. Since, the model we have designed is generic and can be used for any cell design, simulations were completed for academic understanding as well as proof of concept which will be discussed in section 4.2. A robust cell and proposed model which uses PDK, Monte-Carlo simulations, statistical

modeling and error correction leads to an accurate maximum yield estimation. Table 1.1 below gives a brief overview of important parameters and trade off to achieve them.

Table 1.1 Parametric design chart for SRAM cell topology

PARAMETERS	EFFECT	HOW TO ACHIEVE	ADVERSE EFFECT OF OTHER PARAMETERS
• Cell Density	• Less area, more capacity	• Minimum geometry devices • Move to new process node for high capacity	• Mismatch Increases • Reduces yield • Difficult design validation.
• SNM at low voltage	• Less read failure • Low Power • High yield	• High Cell Ratio • More area per cell	• Reduced Density
• Cell stability (PVT)	• High Yield	• More area per cell	• Less Density
• High Cell Current	• Fast memory and high performance	• Increase CR • Low vt (threshold)	• Less Density, high leakage • Increased mismatch error with low vt • Reduces yield
• Low leakage Current	• Less Power consumption • Reduced read failure	• High vt (threshold) transistors	• Slower memory

It is observed that there is always a tradeoff between Power, Performance and Area. A smart SRAM cell design can lead to a low power, high speed and less area memory.

1.8 PROMISING ALTERNATE SRAM CELLS

Due to multiple issues discussed in section 1.6, there are some limitations on using a 6T SRAM cell. As seen in design chart, using high threshold devices does improves the cell stability of the SRAM cell, but at the same time due to low cell current read performance is impacted. However, there are few promising alternate SRAM cell topologies which might be used according to the application. Although alternate SRAM cell topologies might not solve all the issues as there is always a trade off, but can be surely used to achieve target specific parameters. Most of the topologies discussed below have similar read and write function as discussed in section 1.5, there are minor difference in the working operation. Different topologies were introduced to overcome few of the greater error sources like Icell variation, access failure, read failure and greater Ion/Ioff ratio. This section is restricted to the extend explaining different topologies driving down one or more than one error sources and not the read/write operations of individual SRAM topology.

1.8.1 7T SRAM Cell

A 7T SRAM cell is shown in fig.18.1. 7T SRAM cell is similar to the 6T cell where M7 is added to break the loop during read and write operation. M7 is connected to word line bar, which keeps M7 off when memory cell is accessed [18].

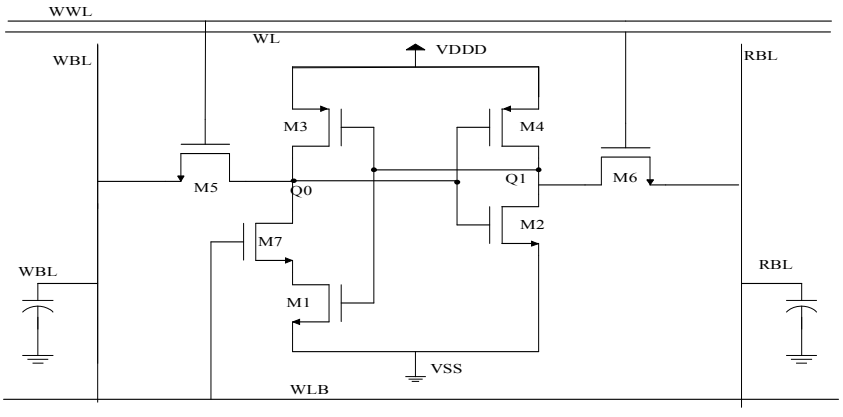


Figure 1.6 7T SRAM Cell [14]

7T SRAM cell was introduced to overcome extra area spent in the 8T SRAM cell, but at the same time provide similar SNM as of 8T. This topology improves the write margin asymmetrically. This topology allows memory cells to work on lower supply voltages due to improved SNM. In a nutshell, 7T was introduces to overcome the SNM issue of 6T and to avoid area constraint of conventional 8T [18]. 7T SRAM cell drives down the read failure error source and can be used to design independent read time to improve the performance. There are two different types of memory read techniques, one is differential ended sensing technique explained in conventional 6T and the other is single ended sensing technique which uses only one bit line to read the data used in 7T. In figure 1.6 RBL is the read bit line used to sense the data. This bit line is then connected to a process tolerant comparator which takes the decision by comparing to reference voltage. This topology addresses read failure and data dependency.

1.8.2 8T Half select condition cross point(CR8T) SRAM cell

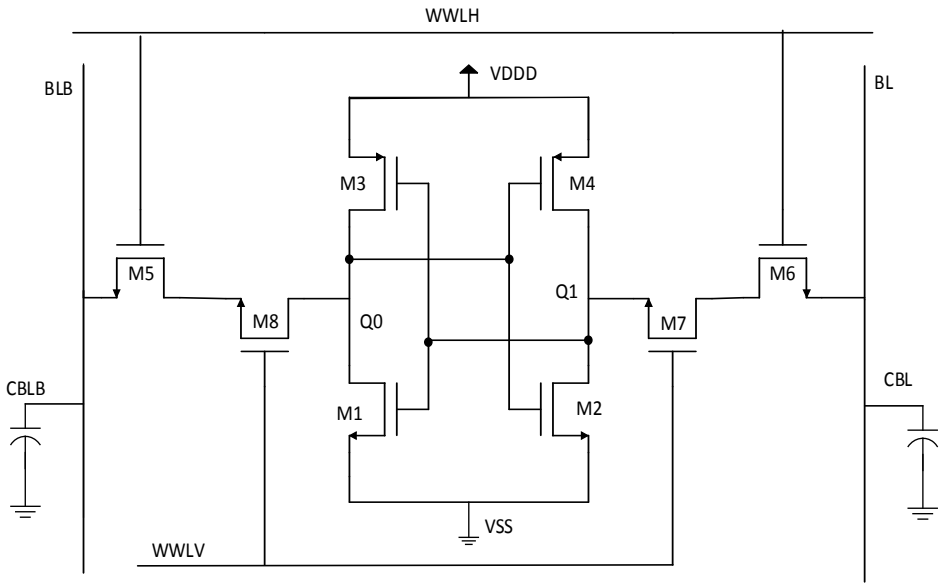


Fig. 1.7 Half select condition Free Cross point 8T [19]

Figure 1.7 shows a Half select condition free cross point 8T SRAM cell. Unlike 6T it has two different word lines. This is similar to 6T with two pass transistors controlled with two different control signals, Horizontal word line (WLH) and Vertical word line (WLV). These control signals are generated using address decoders. Node voltages are accessed only when WLH and WLV are high. This improves the read margin of cell but slows down the memory. This topology was proposed to overcome the half select failure. Since the unselected cells in the row will have low WLV which keeps bit lines to be isolated from the node voltages. To operate at low voltage, negative BL biasing can be used to improve the access speed. This topology specifically addresses half select issue while also improving read failure by adding one more access transistor in series.

1.8.3 9T Decoupled read SRAM cell

A 9T Decoupled read SRAM cell is shown fig.1.8 which uses three additional transistors for read operation. There is no direct access to the node voltages during read operation. Read margin in 9T is equal to the hold/static noise margin of the cross coupled inverter pair used in 6T. Thus read margin is improved when compared to 6T.

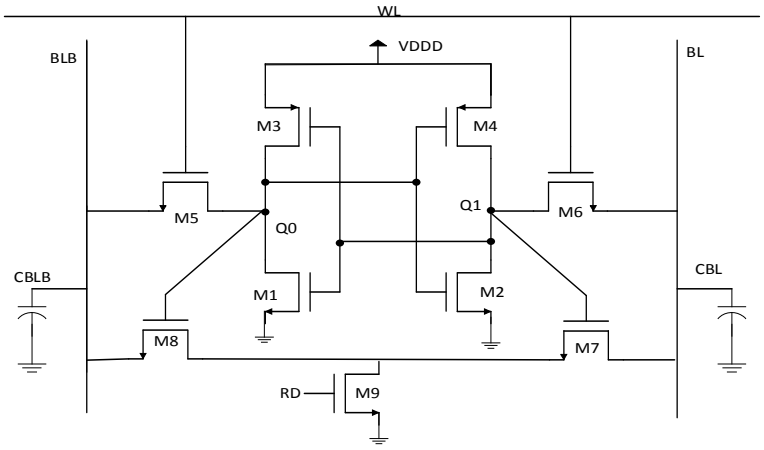


Fig. 1.8 9T Read Decoupled SRAM cell [20]

The read noise margin is equivalent to static noise margin of M1-M3 and M2-M4, 9Tmemory cell can be operated at low voltage. To increase the memory performance M9 can be increased. As the area per cell is increased the memory density goes down. In 6T, read and write margin conflict exist in designing, 9T was designed to have individual design approach for read and write margin. M3 and M4 are designed to achieve desired write margin whereas M8 and M9 are used to design read margin. The topologies addresses more than one error. It eliminates the read failure, achieves access time without affecting write margin, and shows a tighter leakage power distribution as compared to 6T leakage power distribution [20], which reduces data dependency bit line leakage error. 9T topology looks promising for future advanced process nodes.

1.8.4 10T SRAM cell with high cell per bit line

A 10T SRAM cell is shown in fig.1.9 which eliminates the data dependency bit line leakage error. The topology is designed to have low leakage current to have reliable bit line differential voltage to have less access failure errors in memory. Transistors M7 – M10 are used to provide decoupled read operation. When memory cell is not accessed RWL = 0, M10 turns ON, where as M7 and M9 are turned off. Drain of M8 is held high enough, thus there is less leakage current. This allows to have high cell per bit line using 10T SRAM topology as shown in fig. 1.9.[21]

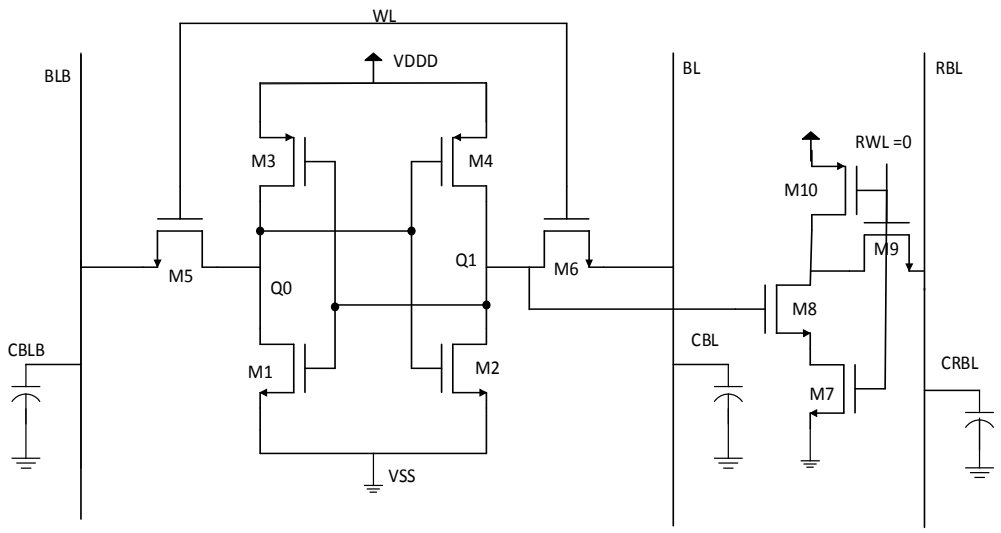


Fig. 1.9 10T SRAM cell with high cell per bit line [21]

Industry are introducing 12nm-10nm devices in applications, the projection is till 7nm-5nm. 3D transistors and finFET's have remarkable less variation in threshold [8, 22, 23]. Having good analysis of variation and a correlation of modern process node variation and planar process node variation could help choosing the right topology for given specification. There can be further development in architecture of SRAM to support the threshold variation and designing access time using statistic approach. To understand this approach we select 6T SRAM cell. It is important to understand that this approach can be extended to any cell topology as it only depends on the memory cell current variation.

1.9 Thesis Organization

This thesis presents a statistical approach on SRAM architecture design to have maximum yield. All the simulations are done using Cadence 6.1.4 IBM 7RF process. CHAPTER II presents variation and stochastics in CMOS devices, it gives the basic understanding of statistics, source and effects of variation and process corner design consideration. CHAPTER III presents literature review of different memory read techniques and proposed high level architecture to improve the accuracy of read timing. CHAPTER IV presents the implementation of proposed idea and simulation results. CHAPTER V presents the comparison results between conventional and proposed idea. Future improvements are discussed to improve the accuracy of the architecture.

CHAPTER II

VARIATION AND STOCHASTICS IN CMOS

2.1 STOCHASTIC APPLICATION

Stochastics is defined as a process which considers variables having random probability distribution; which are analyzed, interpreted and presented using statistics [24]. Statistics is used to take precise predicted decisions with error margins. Statistics is a process of collecting data, analyzing and interpreting using different statistical models. There are different parameters in a semiconductor device which are affected throughout the process of fabrication. A device would not work "exactly" the similar way it is simulated. A predicted variation in the operating parameters is observed in fabricated device. Statistic application is used for better understanding of pre-fabricated design and post fabricated working of device. Following section explains basic concept of statistics and its application for a better design. This chapter introduces basic concepts of statistics and covers the source and effects of variation. Below are few basic and important definitions in statistics.

2.1.1 Population: A population is all possible data entries of interest represented [24]. Recording threshold Vth, of every single device fabricated on all the wafer's is a population data set, or recording the cell current value of every single memory cell of the memory bank fabricated on all the wafer is also a population data set.

2.1.2 Sample: It is both time and resource consuming to analyze the effect of all data samples present in population data set. A sample collects random data points (sample size) from population to analyze and predict population behavior. A random sample of N devices is taken to plot Vth distribution. Sample distribution is then used to understand the population Vth distribution.

2.1.3 Central Limit Theorem: The sample mean approaches the population mean by increasing the number of random samples. A sample size of more than 30 is adequate to have a sample mean approximately equal or closer to the population mean with considerate sampling error. It is given in the [25] that sample size between 100 - 1000 is sufficient enough to understand the threshold variation of a devices.

2.1.4 Sampling error: Too big sample size is an investment of both time and resources, while too small sampling size can lead to inefficient designing. Sampling error margin helps deciding the number of samples for efficient designing. Sampling mean gets closer to the population mean as we increase the sampling size. An error between sample mean and population mean is given by $MOE(Margin\ of\ Error) = (\bar{X} \pm \mu) = Z_{\frac{\alpha}{2}} * \left(\frac{\sigma}{\sqrt{n}}\right)$; where $\bar{X}$ is the sampling mean, μ is the population mean, Z is the Z score value calculated form the normal distribution table, α is the significance level, σ is standard deviation and n is the number of samples taken from the population data [24].

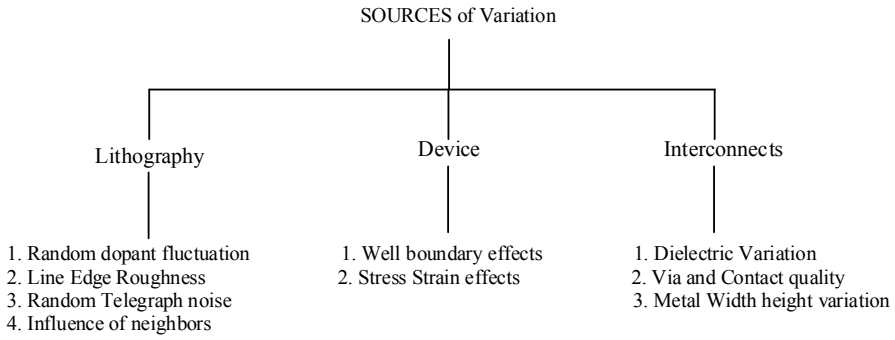
2.1.5 Confidence Interval: A confidence interval is the range defining how confident the predicted population mean lies between defined ranges. Confidence Interval is given in percentage. Population mean is calculated using confidence interval estimation given by equation (2.1.5) [24].

$$\text{C.I} = \bar{X} \pm Z_{\frac{\alpha}{2}} * \left(\frac{\sigma}{\sqrt{n}}\right) \quad (2.1.5)$$

All the statistical terms mentioned above are summarized for their design use in the proposed memory architecture. Statistical analysis is used in analog, mixed signal, digital and memory designs to improve the yield and performance. This introduction to statistic supports to the further discussion of thesis.

2.2 SOURCE OF VARIATION

The statistical term variation is the standard deviation of data set from mean value, whereas variation in semiconductor process is related to standard deviation of electrical parameters with designed electrical parameters of device. There are different sources which drive parametric variation in devices. Following are few important sources which causes variations.



Among above mentioned sources of variation there are few significant sources discussed further in this section. One of the major source to contribute in parametric variation is Lithography. RDF (Random Dopant Fluctuations), LER (Line Edge Roughness), RTN (Random Telegraph Noise) and Influence of neighbors are different types of source variation under lithography. There are other variation contributors such as dielectric variation, via and contact quality, gate oxide thickness variation, channel width variation, Stress strain effects on mobility and also metal width height variation. Let us discuss major sources of variation in semiconductor devices.

2.2.1 Random dopant fluctuation

To continue Moore's law scaling, industry is advancing to new and smaller process nodes. As a result the number of dopants in the depletion region are decreasing in newly scaled process nodes. For an example at W, L = 0.1μ with $Na = 10^{18}\ cm^{-3}$, depletion width= $350\dot{A}$, where number of atoms in the depletion region are given by $N = Na \bullet L \bullet Wdmo =$ $350\ atoms$ [8]. The count of number of dopants in depletion region for the latest technology beyond 32nm node has reached to less than 100's of dopants. A small variation in doping results in significant performance error. Fig. 2.2.1 shows simulated RDF [8]. The green dots shown are the dopant atoms. The dependence of threshold voltage on number of dopant atoms when source-body voltage is zero is shown in equation (2.2.1.a).

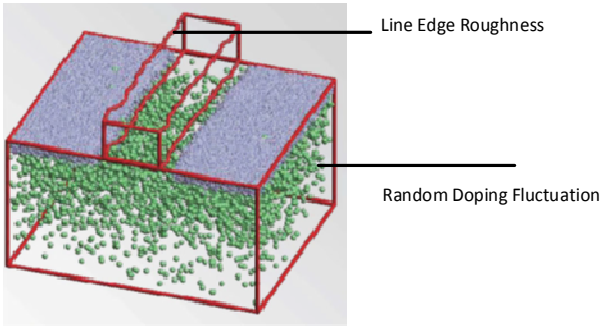


Fig. 2.1 Atomistic process simulation incorporating RDF and LEF as the source of intrinsic fluctuation [8]

$$Vt = \frac{\sqrt{2qN_a \in (2\emptyset f)}}{C_{ox}} + 2\emptyset f + \emptyset_{ms} - \frac{Q_{ss}}{C_{ox}} \qquad (2.2.1.a)$$

The standard deviation of number of dopants from device to device follows a Poisson's distribution due to its discrete statistical nature [8, 26]. Therefore $\sigma N = \sqrt{N}$. The overall threshold variation due to RDF is given by equation 2.2.1.b.

$$\sigma_{\Delta Vth} = \left(\sqrt[4]{(2q^3 \in_{Si} Na \emptyset_B)}\right) * \left(\frac{T_{ox}}{\in_{ox}}\right) * \frac{1}{\sqrt{3WL}} \cong \frac{Avt}{\sqrt{WL}} \quad (2.2.1.b)$$

Since, threshold is a continuous function, the standard deviation is statistically modelled using Gaussian distribution (Normally distributed). Avt is a pelgrom coefficient of the process and technology dependent [4].

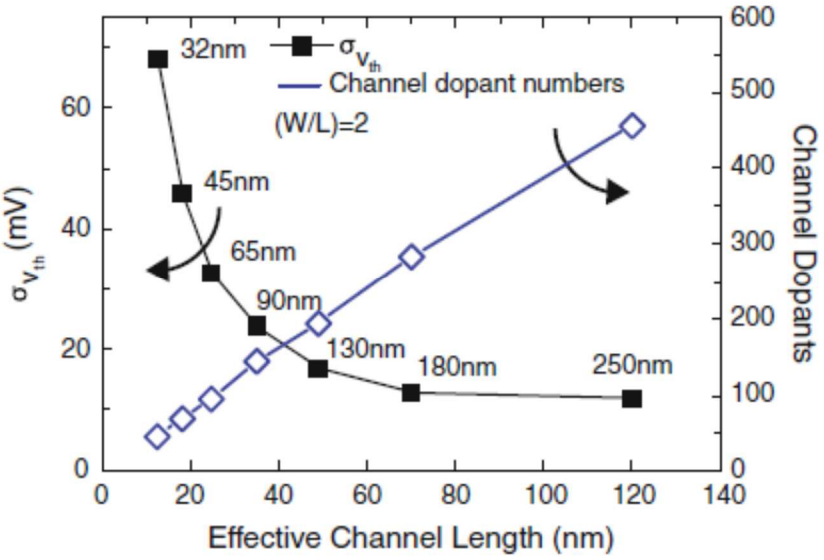


Fig. 2.2 Impact of RDF on σ *V*th and number of dopant atoms in the depletion layer of a MOSFET for different technology nodes [4, 8, 26, 27]

Fig. 2.2 shows that Vth variation increases with advancement in process node technology. Blue line in fig. 2.2 shows a decrease in number of dopants as process node technology enhances towards smaller node lengths. Variation is inversely proportional to the square root of the area and to reduce threshold variation area of the device has to be increased to in effect a greater sample size. In order to reduce variation in modern node technology, it is not possible to granularly increment area of devices and therefore area is increase by adding number of fingers. To understand this concepts, let's take an example where an

input offset voltage is calculated for a minimum geometry NMOS device in a 0.18um technology.

$$Vosn = \frac{Avtn}{\sqrt{((W-Kvtw)*(L-Kvt))}} = \frac{12mV.\mu m}{\sqrt{(0.22+0.58)*(0.18-0.058)}.\mu m} = 40mV$$

Now, if the required input offset is $Vosn \cong 35\text{mV}$ then, width of NMOS transistor can be granularly increased to 0.38um which gives $Vosn = \frac{Avtn}{\sqrt{((W-Kvtw)*(L-Kvt))}} = \frac{12mV.\mu m}{\sqrt{(0.38+0.58)*(0.18-0.058)}.\mu m} = 35mV$. If we did not have granularly increment of width/length but only the number of fingers then from one finger (minimum geometry) to two finger geometry the input offset will change to $Vosn = \frac{Avtn}{\sqrt{(2((W-Kvtw)*(L-Kvt)))}} = \frac{12mV.\mu m}{\sqrt{2(0.22+0.58)(.18-.058)}.\mu m} = 28\ mV\ ; Nf = Number\ of\ fingers.$ Having Nf over granular increment to increase the area of the devices helps improving input offset voltage, but there are few circuits where extra area is spend due to Nf option than rather incrementing it to the calculated width. Observer that a 22nm node with Avt approaching 2 mV-um Vos per finger equals 48mV. These results demonstrate that memory design becomes a greater yield issue primary due the Poisson nature of the doping as it relates to Vth.

2.2.2 Line Edge Roughness

Error related to the inaccurate gate patterning is referred as line edge roughness. Lithography wavelength for modern nodes has reached from 500nm to 193nm for Gate patterning. Fig. 2.3 shows lithography wavelength scaling for different technology nodes. Beyond 180nm device fabrication optical lithography with enhancement techniques are used. These techniques are aperture improvement using OPE's (Optical proximity effects) and immersion technology [8, 25, 28, 29].OPE's are the major contributors to variations and also decides the smallest feature size fabricated in a node generation. Beyond 50nm process node, LER effect is a significant contributor to threshold variation [8, 25, 28, 29].

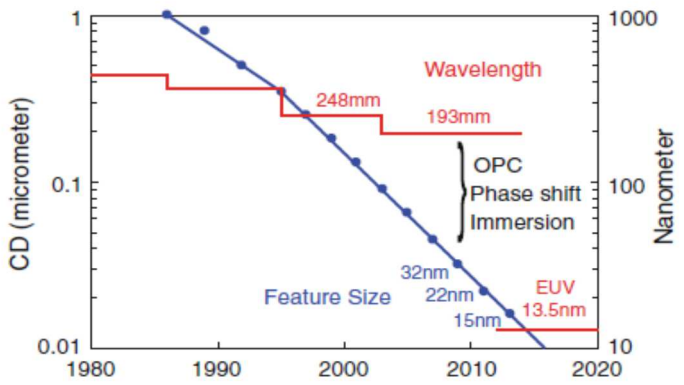


Fig. 2.3 Lithography wavelength scaling for different technology nodes [25].

Inaccurate gate patterning results in variation of length and width. This effect is known as NWE (Narrow Width Effect) and SCE (Short Channel Effect) which results in threshold variation [30-32].DIBL (Drain Induced barrier lowering) also contributes to the threshold variation [30, 31]. Vth variation modeling due to SCE and DIBL is shown in equation (2.2.2.a)

$$Vt = Vt0 - (\zeta + \eta * VDS)e^{-\frac{L}{\lambda}} \quad (2.2.2.a)$$

Where ζ is short channel effect coefficient and η is DIBL coefficient [8, 30, 31].

A Velocity saturation drain current is shown below in Equation (2.2.2.b)

$$Id = Vsat * WCox(Vgs - Vt) \quad (2.2.2.b)$$

In Equation (2.2.2.b), $Vsat$ is inversely proportional to length of the device. Drain current is directly proportional to threshold voltage Vt and inversely proportional to length. From Equation (2.2.2.a) and (2.2.2.b), threshold variation due to RDF and LER contribute to the variation in drain current. Variation in current affects performance of the system.

Total threshold variation can be given by equation 2.2.2.c

$$\sigma_{Vth} \approx \sqrt{\sigma^2 Vth, RDF + \sigma^2 Vth, LER + \sigma^2 Vth, other} \quad [8] \quad (2.2.2.c)$$

2.2.3 Random Telegraphic Noise

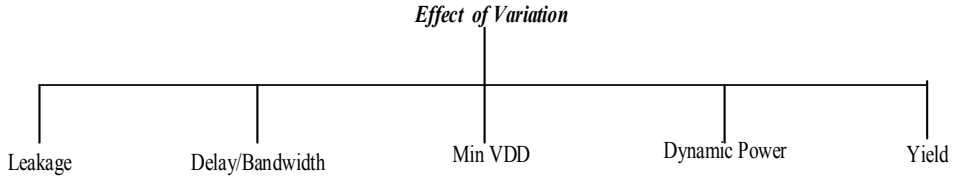
Drain current in transistor flows due to both horizontal and vertical electric field contributions. This results in a scattering term in electron flow from source to drain. Since the velocity of these electrons can be high, a few of the electrons hits atom and knocks off free electrons and holes, some electrons get trapped in oxide and some electrons travels to drain while the holes goes to the body substrate and eventually ground. This trapping continues for a while until an electron hits and de-traps all the electrons. There is sudden current increase due to trapped and de-trapped electrons. Trapping and de-trapping of electrons also changes the threshold of the device. This is also referred as $\left(\frac{1}{f}\right)$ noise. RTN follows a discrete statistical model. There is significant difference in RTN and RDF. Equation (2.2.3) shows threshold variation due to RTN.

$$\Delta V_{th,RTN} = \frac{q}{W_{eff} L_{eff} C_{ox}} \tag{2.2.3}$$

From Equation (2.2.3) it is observed that RTN is indirectly proportional to effective width W_{eff} effective length L_{eff}, and C_{ox} gate capacitance per unit area [33, 34]. RDF is inversely proportional to the square root of the area shown in equation (2.2.1.c) whereas RTN is inversely proportional to effective device area shown in equation (2.2.3). At new process nodes RTN is expected to mask/show combine effect with RDF [8]. RTN has non-Gaussian long tail behavior which makes design critical beyond $\pm 3\sigma$ [8] so at new process node a design engineer should be aware of skewed probability function after $\pm 3\sigma$, otherwise the assumptions results will be misleading.

2.3 EFFECTS OF VARIATION

Following flow chart shows different effects of variation. Few significant effects of variation are leakage in device, delay/bandwidth, minimum VDD requirement, and increased dynamic power.



In most of the today's applications, power consumption has become a critical issue. With variation in device parameters, power and performance are significantly impacted. The issue of variation can be referenced in a nutshell by knowing total power consumption equation (2.3.a)

$$P = n\left[C\ (V_{DD})^2 f\ +\ V_{DD} I_{off}\right] \tag{2.3.a}$$

Where C is parasitic or load capacitance, n is the number of devices, V_{DD} is the DC supply voltage of an application, f is the operating frequency and I_{off} is leakage current given by Equation (2.3.b)

$$I_{off} = Is * \exp\left[-\frac{V_{th}}{nU_T}\right] \tag{2.3.b}$$

$$Is = 2n * \mu n * Cox * \left(\frac{W}{L}\right) * U_T \tag{2.3.c}$$

Where I_{off} is leakage current of a MOS device when $V_{GS} = 0$, n is subthreshold slope, Cox is the gate oxide capacitance, U_T is the thermal voltage, V_G is gate voltage, and V_{th} is threshold voltage. The expectation here is to reduce V_{dd}, increase operating frequency f and reducing I_{off}. Variation in different parameters of devices affects ideal operation of the

application. Let us discuss which application parameters are affected due to variation in the device, ultimately affecting total power consumption.

Fig. 2.4 shows change in current at different threshold. Here, it is observed that if the threshold voltage randomly takes lower value than the designed (ideal/typical) value then leakage current I_{off} increases exponentially. This in turn increases the power consumption. NWE, SCE, DIBL discussed earlier and oxide thickness tox together contributes to the threshold variation resulting variation in leakage current. Delay/Bandwidth is majorly affected due to threshold variation. It is explained using Equation (2.3.d), where Δt is the delay/time defined to charge C (load or parasitic capacitance) to ΔV voltage with current I. Therefore a threshold variation affects the time delay ultimately effecting bandwidth.

$$I = C * \left(\frac{d_v}{d_t}\right) \cong C * \frac{\Delta V}{\Delta t};\ \Delta t = \frac{C}{I} * \Delta V \qquad (2.3.d)$$

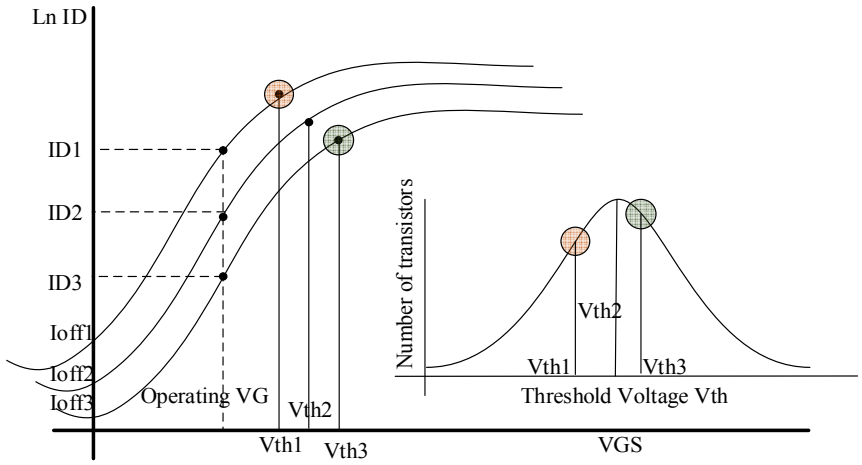


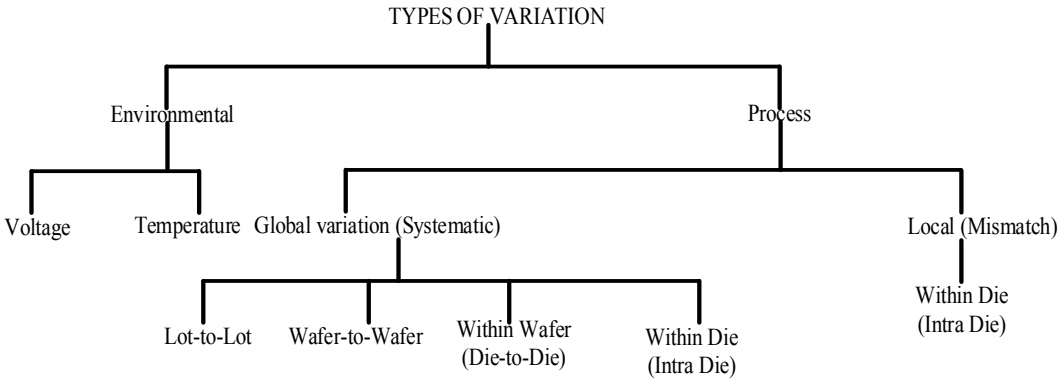
Fig. 2.4 ln $ID\ Vs\ V_{GS}$ curve for different threshold voltage

Variation in C also effects dynamic power in addition to changing performance and leakage Equation (2.3.a) and (2.3.d).

In [3, 10] it is shown that a 90% for a 1MB SRAM yield can be achieved by requiring -6σ SNM greater than 0.04xV_{dd} or restated $SNM_{min} \geq 20\% SNM_{typ}$ for a 90% yield. The same holds true for the inverter noise margin or any logical device noise margin. Required minimum noise margin sets limitation on minimum operating voltage. Therefore a larger variation would violate a SNM_{min} requirement limiting minimum V_{DD}. This implies a supply budget comprised as follows; V_{DD}minimum is set as follows 2xnoise margin + 2xnsigma threshold variation + 2xoverdrive margin [35]. $V_{DD} = 0.4V_{DD} + 2nAvt + 2\Delta V$; Where Avt is threshold variability, $0.4V_{DD}$ $is\ the\ noise\ margin\ approximately$ 20% $of\ Vdd$ and ΔV is the over drive voltage of the gate. In this section it was shown that variation (primarily threshold) effects leakage current, delay/bandwidth, minimum V_{DD}, dynamic power and yield with a concise explanation.

2.4 DIFFERENT TYPES OF VARIATION

A brief overview of stochastic application, source and effects of variation has led to the final discussion about different types of variation. Following classification gives basic idea of different types of variation.



There are two types of variation, Environmental and Process followed with the sub classification. Process variation is covered in this section and Environmental variation is discussed in Section 2.5.

2.4.1 Global variation (Systematic/Process)

Global variation is the gradient variation across the wafer caused due to physical errors during manufacturing a device. It is caused due to misalignment in the lenses and change in properties of elements used in the lithographic process. It is statically modelled as random probability function which follows spatial correlation referred to as gradient/process variation. Devices fabricated at the center of the wafer will have different properties when compared to the devices fabricated at the edge of the wafer.

2.4.2 Local variation (Random mismatch)

Local variation is predominantly observed due to sources of variation discussed in section 2.2.1 RDF. Local mismatch is between the devices placed in close vicinity of each other. Local mismatch variation is mostly observed in within die. For a certain area local mismatch dominates over systematic variation. If the design, in this case memory bank, is within the process defined area then it is dominated by local mismatch. If the design (memory bank) is large enough then systematic variation will dominate over random variation [25]. Since local variation is due to RDF it is well defined using statistics to design parameters with predictive models.

2.4.3 Systematic and Random variation

- Stochastic perspective

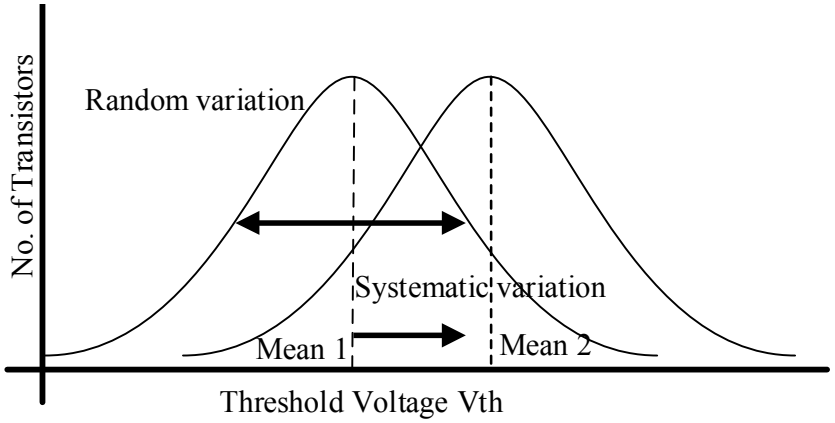


Fig. 2.5 Stochastic perspective for random and systematic variation [25].

Stochastic perspective of random and systematic variation is shown in fig. 2.5 [25]. The mean value in systematic variation of threshold changes when the device is spatially displaced and still follows the random variation effect at that point. Random variation remains constant in all the die's on all wafers, only its mean value differs depending on the gradient of wafer.

- Device engineer perspective

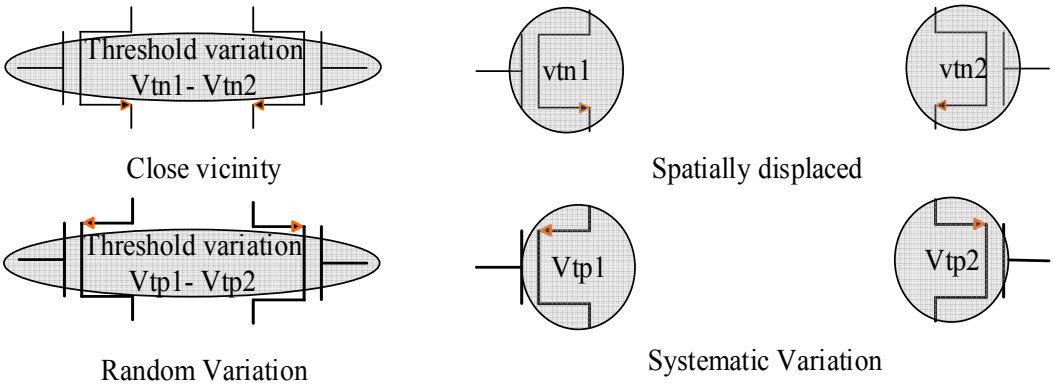


Fig. 2.6 Device engineer perspective for random and systematic variation

Device engineer perspective for random and systematic variation is shown in fig. 2.6. Random mismatch is considered between closely placed devices. It is a dominant error in circuits like differential amplifier, sense amplifier and comparators etc. Systematic variation is considered between spatially placed devices. Systematic variation should be considered when the memory bank is huge enough that extreme columns will have effective different mean current.

It is difficult to find the numbers for an area where systematic variation starts dominating over local mismatch variation. Numbers can be determined using statistically significant test data of the wafer or may be found in the PDK when provided by the process manufacturer. IBM 7RF process provides this data in PDK. An IBM 7RF design under $200\mu m$ x $200\mu m$ is dominated by local mismatch over systematic variation [36]. In paper [37] test results shows that systematic variation is shared by Die-to-Die and Within Die, Systematic variation starts increasing with increases in die size.

- Radial Gradient

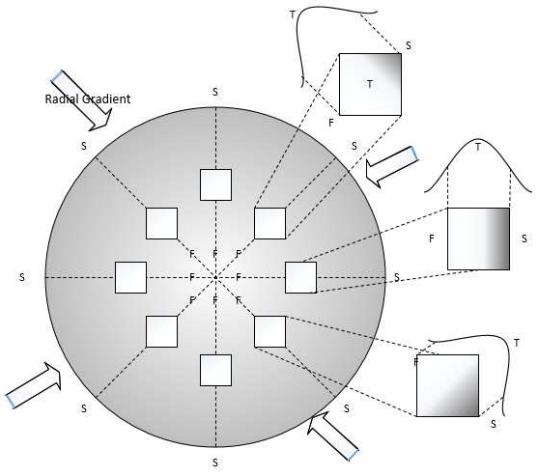


Fig. 2.7 Radial gradient on wafer

All wafer's have a process gradient due to which devices function either Fast (F), Slow (S) or Typical (T). Fig. 2.7 shows the radial gradient over wafer. A wafer gradient exists because of the doping profiles. The center of the wafer has *Fast* devices. The edges of the wafer have *Slow* devices. Most of the area on wafer has *Typical* devices. Due to the radial nature of the gradient it is very difficult to nullify the gradient effect as we are typically unaware of the die position. However, the manufacturer does have these test result data which are used for better designing. Gradient impact on each die is shown in fig. 2.7. The die position sets the gradient to be either diagonal, right to left, top to bottom and vice versa. As discussed in section 2.4.1 systematic variation is statistically modelled as random probability distribution. This distribution is shown around the small die in fig 2.4.3 with mean value as typical and slow-fast being at the tail. The gradient effect is considered in the proposed architecture which will be discussed in the chapter III.

- Lot-to-Lot and Wafer to Wafer variation

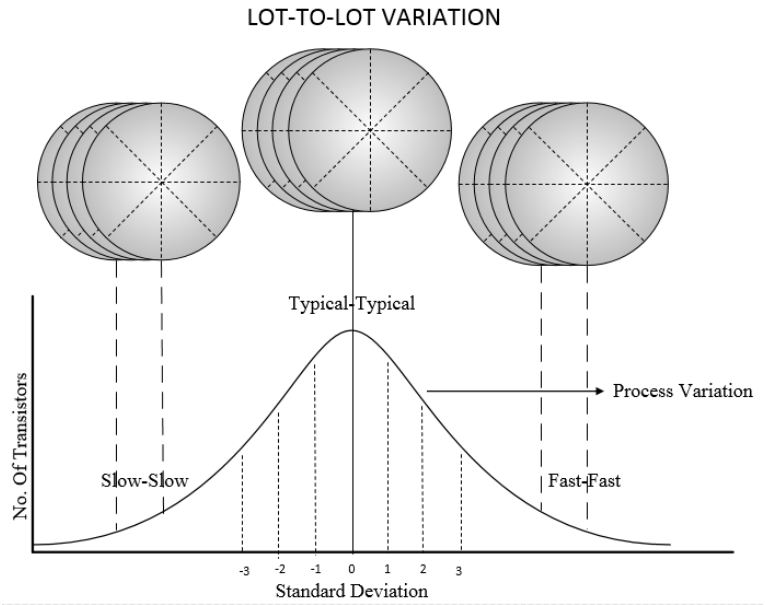


Fig. 2.8 Lot-to-Lot variation

In a fabrication process, multiple wafers in a lot are fabricated. During lot fabrication, some physical changes in mechanical and fabrication process takes place. These changes results in Lot-to-Lot variation. There are three sets formed in the lot. Slow process lot, Typical process lot and Fast process lot. Fig. 2.8 shows lot to lot variation and fig. 2.9 shows Wafer to Wafer variation. A Slow process lot will have a slow-typical-fast corners as discussed in radial gradient section and a Fast process lot will also have a slow-typical-fast corner. In fig. 2.9 Slow-Slow represents the slow corner in slow process lot whereas Fast-Fast represents fast corner in fast process lot. These points can be seen in fig. 2.9 which shows Wafer to Wafer or Within Lot variation.

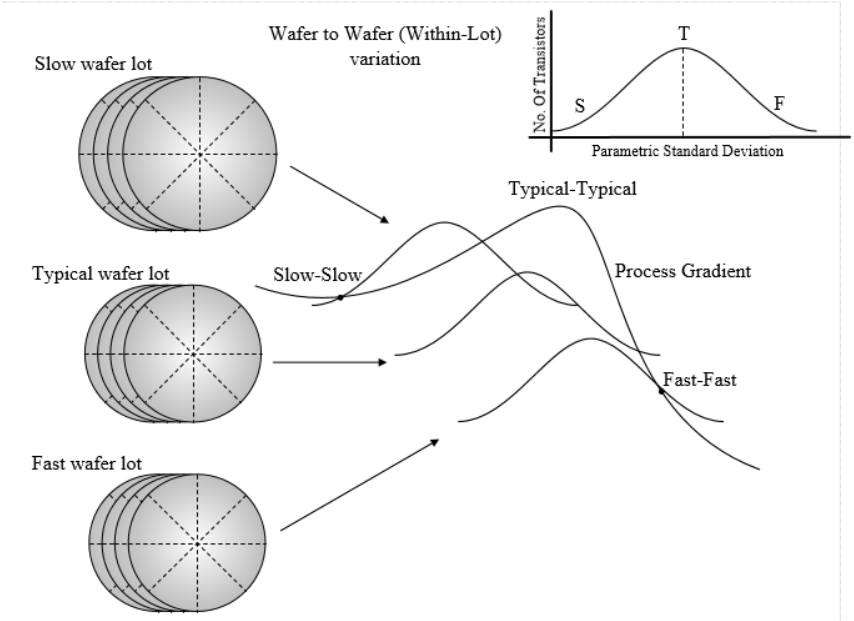


Fig. 2.9 Wafer to Wafer variation

- Die-to-Die and Within Die variation

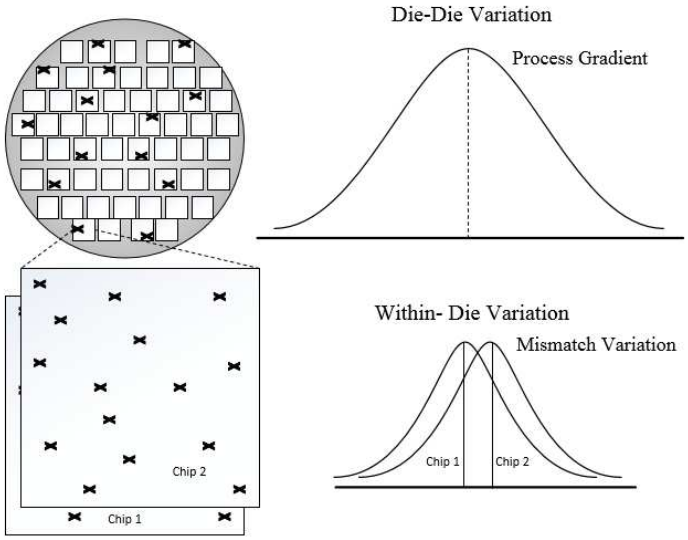


Fig. 2.10 Die-to-Die and Within Die variation

Die-to-Die and Within Die variation is shown in fig. 2.10. Fig. 2.10 shows chip (die) 1 and chip (die) 2 from the same wafer has different mean but same variation. In fig 2.10 it is assumed that die size is within 200 μ m x200 μ m for an IBM 7RF process.

2.5 Process Corners PVT (Process, Voltage and Temperature)

Process corners include statistical analysis of both process and environmental variation. Design parameters vary from their typical behavior when encountered with the environmental changes. There are two major factors which causes environmental variations and they are voltage and temperature. Let us discuss how voltage and temperature change affects designed parameters.

2.5.1 Voltage and temperature effects

Supply voltage is not constant through all devices. There is variation in supply voltage due to power supply noise, IR drop, capacitance tolerance and etc. Speed of the device is proportional to the supply voltage. Historically supply voltage variation was modelled as $\pm 10\%$ of the typical value. New design techniques and process nodes can achieve a tight supply budget of $\pm 5\%$ [38]. Voltage variation in today's date seems to have low priority among PVT. Threshold voltage of a device is inversely proportional to the temperature with a negative coefficient of $-0.83\mathrm{mV}/\mathrm{C}^0$ also mobility of the device is inversely proportional to temperature [39]. Therefore a change in supply voltage and temperature would result in change in drain current Id of the device which makes a device fast or slow.

2.5.2 Monte Carlo analysis

Monte Carlo is a statistical tool used to analyze total variation. It randomly allots the parameter values (Vth, Leff, Weff, temp, process, supply and etc.) to observe the behavior of the design. These values are drawn randomly such that all possible corners are considered. Cadence virtuoso provides Monte Carlo analysis tool to statistically observe the design results and support improved design yield.

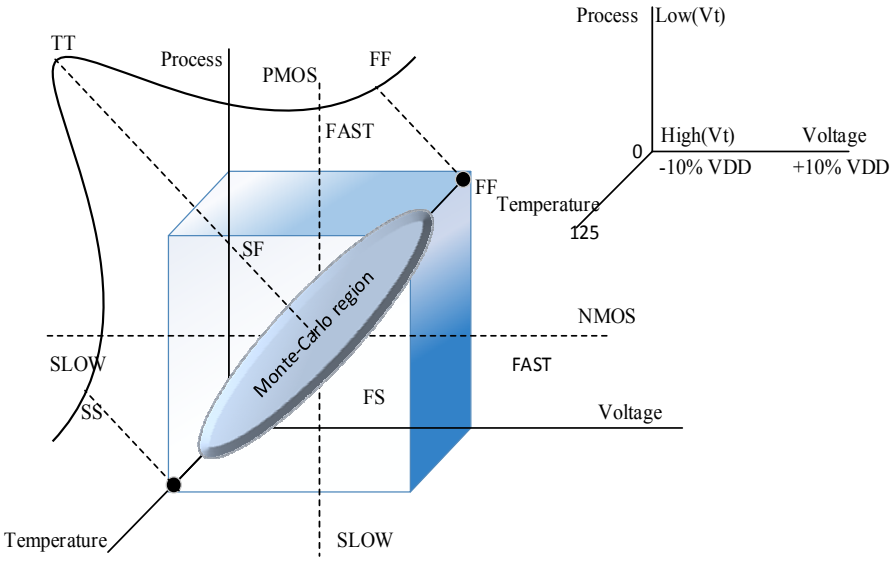


Fig. 2.11 PVT corners for Velocity Saturation device

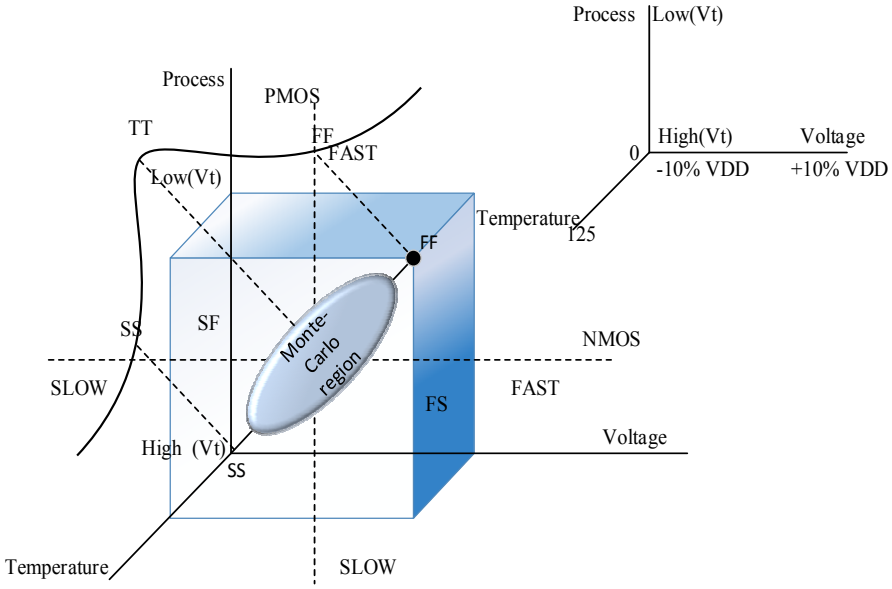


Fig. 2.12 PVT corners for subthreshold Saturation device

Fig. 2.11 and fig. 2.12 shows process corners and Monte Carlo region for velocity saturation and subthreshold saturation devices respectively. There are four main corners where design should pass. SS (Slow-Slow), FF (Fast-Fast), FS (Fast-Slow) and SF (Slow-Fast) where the SS, FF, SF and FS notation stands for (NMOS, PMOS). There are three variations axis considered, process variation which has (Systematic, Mismatch) and environmental variation which has (Voltage and Temperature).

Process and Voltage are directly proportional to the operating speed of the device. Therefore SS corner resides at the bottom right of the cube and FF corner resides at top right of the cube. SS and FF corner changes when the device goes from subthreshold to velocity saturation due to temperature effect. Fig. 2.13 shows change in drain current for different V_{GS} at different temperature.

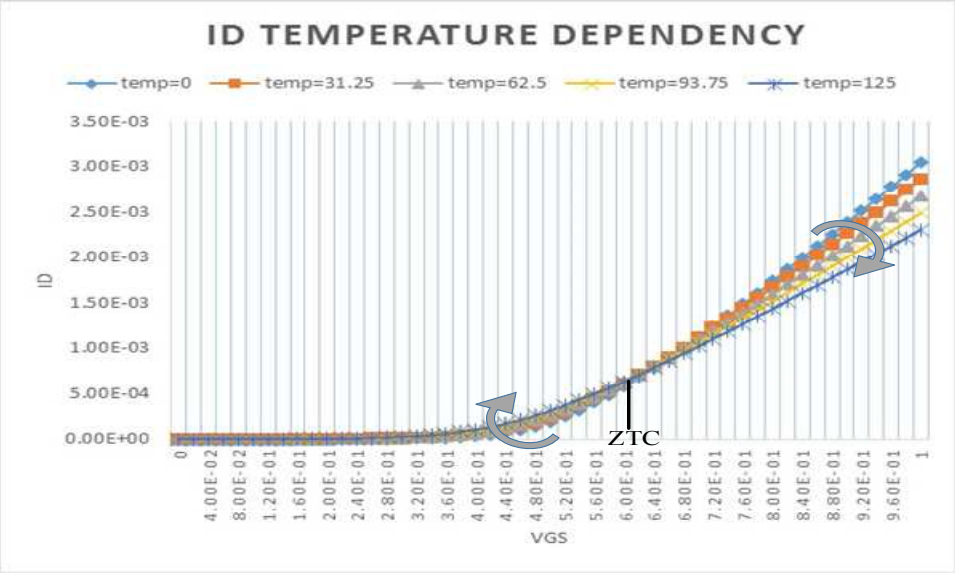


Fig. 2.13 Id dependency on temperature for subthreshold and velocity saturation

For subthreshold Id increases with increase in temperature whereas in velocity saturation Id decreases with increase in temperature. In subthreshold saturation threshold effect is exponential and it dominates mobility. In velocity saturation mobility dominates over

threshold effect. Therefore the process corner changes when device goes from subthreshold to velocity saturation. Observe the ZTC (Zero Temp Coefficient) point where mobility and threshold effect due to temperature is compensated. This gives constant current across the temperature

CHAPTER III

MEMORY READ TECHNIQUES

3.1 CELL DESIGN FAILURE PRIORITY

Understanding and knowing cell design failure priority helps designing high yield memory. The overview of different types of cell design failures is discussed in section 1.6. In this chapter we will present priorities for cell design failure, different memory architectures followed by a discussion of the new proposed architecture. The table below provides a brief overview of typical memory cell design failures, variation source(s) and a priority based on the likelihood of occurrence.

Table 3.1 Cell design failure variation sources and priorities

Error	Systematic Var.	Mismatch Var.	Priority	Comment
Write Failure	Yes	No	Very Low	Large drive requirement of "write" buffer minimizes effect of variability of V_{TN}
Read upset Failure	Yes	No	Low	Dependence on a "small" transistor pair
Hold Failure	Yes	No	Very Low excluding sleep state upset	Low VDD Fix use On chip LDO
Half Select	Yes	No	Low	Dependence on a "small" transistor pair
Bit line leakage	Yes	No	High	Mean column Ioff, V_{TN}
Access Failure	Yes	Yes	High	Sense Amp Vos (mismatch), Icell, V_{th}

To ensure that every memory cell on each bank of each wafer satisfy noise margin requirements, low priority errors should be fixed. Low priority errors can be fixed by selecting proper memory cell size to meet the requirement, adding redundant columns for error correction and by using advanced read-write assist techniques. Studying high priority errors drives improvement achieving more efficient memory.

For a high yield, the total failure probability should be low, beyond $5.5\sigma\ to\ 6.5\sigma$ [8, 40]. With modern process nodes, memory are getting denser and simulations to achieve statistical behaviors for every single failure are getting more intense scrutiny [8]. Among all the errors discussed earlier, bit line leakage and access failure have the greatest impact on yield ultimately determining performance and memory power consumption [9]. Read current variation, bit line leakage current, sense amplifier offset variation and sensing window variation contributes to access failure.

3.1.1 Read current variation and bit line leakage current

Threshold variation results in drain current (I_{read}) variation which follows the same statistical model as of threshold variation. ΔV is expected to be constant for every read memory cell, but due to current variation, ΔV is not constant. Statistical simulation of ΔV helps enhancing yield by selecting a minimum ΔV such that memory bank will have low access failure probability.

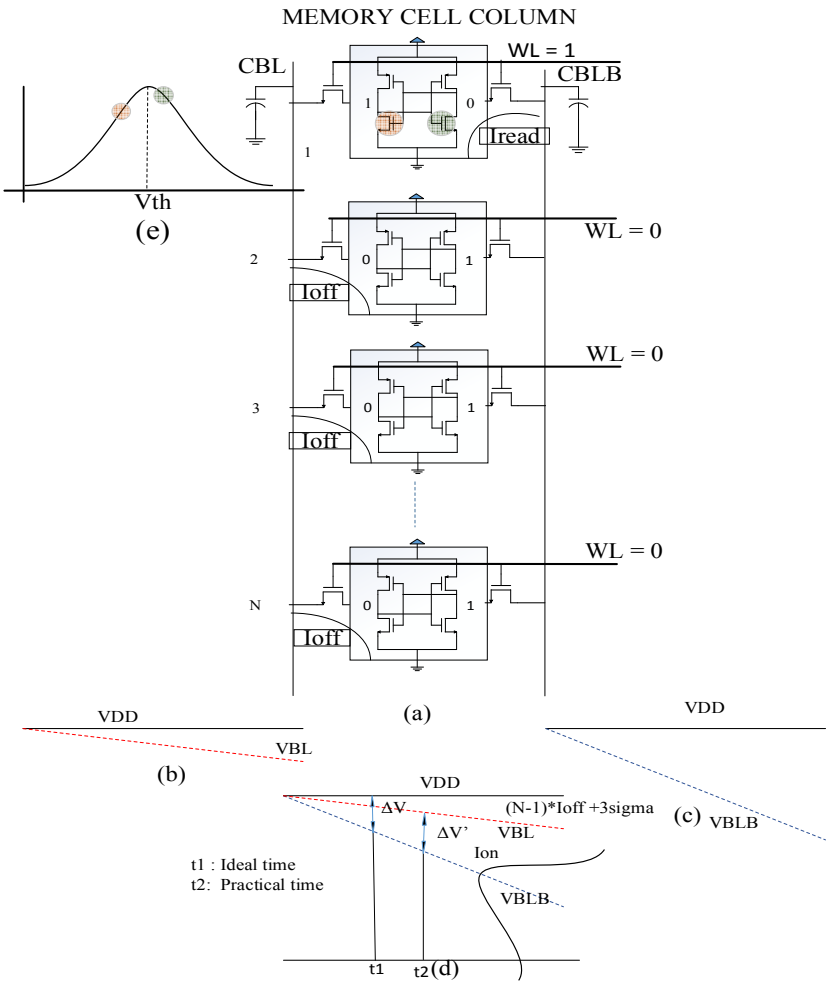


Figure 3.1 Read memory cell in a column

A read memory cell in a column is shown in fig. 3.1. There are N numbers of cells shown in a column. Memory cell 1 is accessed for reading cell data. In read operation, both bit lines are pre-charged to V_{DD}. When memory cell is accessed, I_{read} current flows throw access and pull down transistor. I_{read} discharges the bit line capacitance. Discharging of C_{BLB} is shown in fig. 3.1.1.c. If C_{BLB} droops more than desired, then more power is invested to pre-charge. In [41] it is stated that $\Delta V \cong 10\% V_{DD}$ is sufficient for a sense amplifier to take a valid decision to save pre-charging bit line power. However, the ΔV requirement depends upon the sense amplifier input offset voltage statistics and can be greater or less

than $10\%\ V_{DD}$. The statistical matching of I_{read} and sense amplifier input offset improves power consumption.

Ideally if C_{BLB} discharges then C_{BL} should be held charged to V_{DD} and vice-versa. Due to leakage current flowing throw access transistor, C_{BL} discharges as a function of I_{off}. The worst condition for leakage current occurs when all other cells in the column have the opposite data saved in the memory cell to the cell being read (reading a "1", "0" stored in all remaining cells in a column). Worst bit line leakage condition is shown in fig. 3.1.1.b. As shown in the fig. 3.1.1.d, ideally ΔV should be developed at $t1$ access time but due to leakage current, ΔV is developed after time t and can be read at $t2$. Performance of the memory is effected as $t2 > t1$. Fig. 3.1.1.d shows current distribution over the bit line implying ΔV varies statistically as current varies. Root cause of random variation in current is shown in Fig. 3.1.1.e where there is a threshold mismatch between two transistors in the memory cell taking random threshold values. Every cell in the bank experiences WID variation.

$$t_{read} \cong \frac{CBL}{I_{cell}} * \Delta V;\ I_{cell} = \left(I_{read\mu} - n * \sigma_{Iread}\right) - [(N-1) * Ioff + n * \sigma_{Ioff}] \quad (3.1.1)$$

Since the ΔV is directly proportional to I_{read}, If read access is designed for slow current which is $I_{read\mu} - n * \sigma_{Iread}$, then every cell having read current greater than $I_{read\mu} - n * \sigma_{Iread}$ will be read without failure. For an example, if a Monte-Carlo is run to statistically calculate a memory cell current which gives a cell current distribution with $(\mu, \sigma) = (33.33uA, 1.67uA)$ then $I_{read\mu} - 5\ \sigma_{Iread}$ will be $33.33uA - 5 * 1.67uA = 24.98uA$. Worst case I_{off} for memory design consideration should be $(N-1) * Ioff + n * \sigma_{Ioff}$ where N is number of cells in a column. Equation (3.1.1) shows the sense amplifier trigger time for differential voltage ΔV. It can be observed from equation 3.1.1 that if the read

access is designed for $24.98uA$, then any memory cell current greater than that will surely meet timing constraints and when it does not the failure rate is deemed to be acceptable.

3.1.2 Sense amplifier input offset variation

Sense amplifier observes the differential bit line input voltage and taking the decision when the read data is logic '1' or logic '0'. It is assumed that if the input is greater than 0, $V_{BL} > V_{BLB}$ then sense amplifier detects logic '1' and if the input is less than 0, $V_{BL} < V_{BLB}$ then sense amplifier detects logic '0'. Figure 3.1.2 shows a sense amplifier with input offset voltage having a statistical distribution with mean 0.

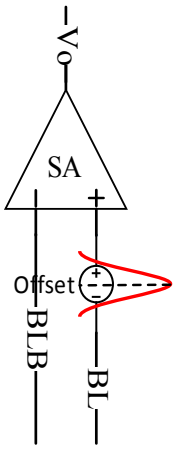


Fig. 3.2 Sense amplifier with offset distribution

Ideally, input offset voltage of the sense amplifier should be zero, but due to RDF in differential pair and PMOS pair of sense amplifier, the design experiences input offset voltage. A properly laid out sense amplifier input offset is dominated by mismatch variation and not by the systematic variation. However, due to the systematic process gradient some SA are slower or faster, but this timing is managed in the proposed

architecture by the timing compensation routine of the dummy cells. Every sense amplifier of the memory bank on all the wafers of all the lots will follow same input offset variation. ΔV must overcome sense amplifier offset voltage to make the correct read decision. A larger input offset voltage forces the designer to design memory with larger bit line ΔV.

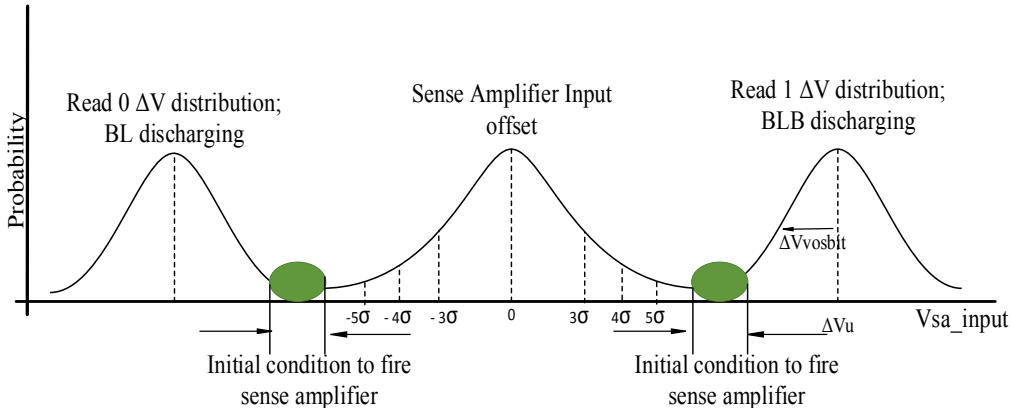


Fig. 3.3 Sense amplifier Initial condition to fire a sense amplifier

Fig. 3.3 shows sense amplifier initial condition to fire sense amplifier in both the read logic '0' and logic '1' condition. ΔV should be large enough than the worst case sense amplifier offset. $(\Delta V_{\mu} - \Delta V_{vosbit}) > \Delta V > n * \sigma_{input_{offset}} + OD$; Where n is number of standard deviation, σ_{input_offset} is sense amplifier input offset standard deviation, ΔV_{μ} is the bit line difference for mean cell current, ΔV_{vosbit} is the standard deviation of the bit line difference and OD is the over drive voltage assumed for better designing. If the ΔV is assumed to be $100mV$ for a 1V sense amplifier process and OD to be $50mV\ to\ 60mV$ which sets the initial settling voltage at cross coupled PMOS pair, then sense amplifier should be designed to have $n * \sigma_{input_{offset}} = 40mV\ to\ 50mV$. Note that the power and

area increases quadratic while the input offset voltage corresponding to the area decreases linearly.

The other important specification of a sense amplifier is settling time. Settling time is the time taken by the sense amplifier to settle the decision. Sense amplifier should settle before the next read/write cycle is initiated. Settling time of the sense amplifier is given by Equation (3.1.2.b).

$$tset = \frac{1}{\omega} ln\left[\frac{Voh-Vol}{V_{SAIN}}\right]; V_{SAIN} \propto \Delta V \pm n * \sigma_{input\_offset} \quad (3.1.2.b)$$

Where Voh is higher voltage drop given by $V_{DD} - V_{DS}$, where as Vol is the lower voltage drop towards V_{SS} , V_{SAIN} is sense amplifier input, ΔV is bit line differential voltage across the bit line and ω is the unity gain frequency of the device given by $\omega = \left(\frac{gm}{Cp} \approx \frac{gmp}{Cgsp}\right)$. Sense amplifier can sense the difference and pull the difference to Voh and Vol. According to Equation (3.1.2.b) too small ΔV will take longer time to settle and too large ΔV will consume more power. There is a tradeoff between power and performance and precise initial condition will utilize near exact power to achieve near exact performance improving the efficiency of the memory.

3.1.3 Sensing window variation

Sensing window is the time taken to fire the sense amplifier after activation of word line. Timing block shown in fig. 1.2 generates read, write, and sense amplifier trigger signals. These signals are generated with respect to clock. An inverter chain(s) is frequently used to trigger word line, sense amplifier enable/reset and etc. Inverters in the inverter chain experiences WID mismatch and process variation. There is variation associated with word line and sense amplifier enable (SAEN) at time t_μ with variation t_σ . Figure 3.4 shows sense window variation.

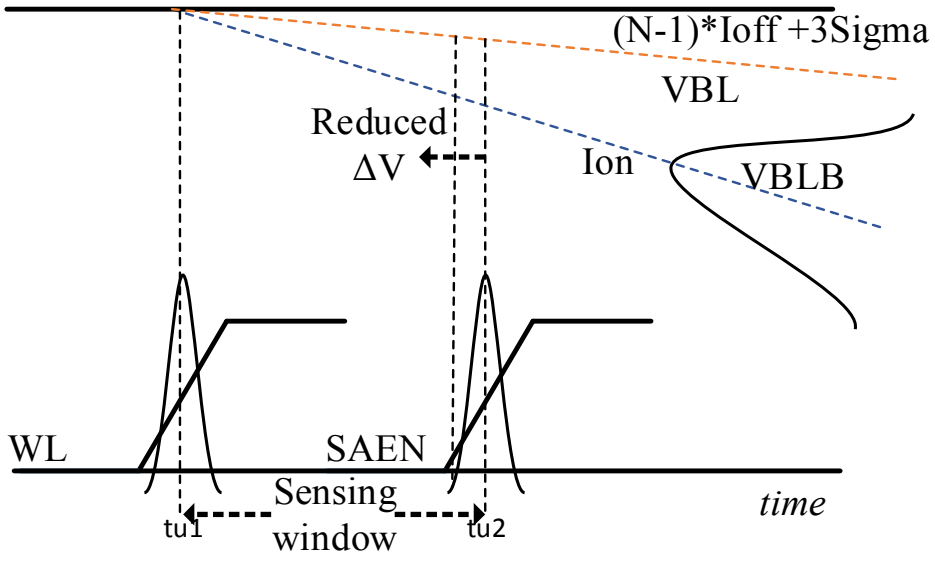


Fig. 3.4 Sensing Window Variation

BLB starts discharging on activation of word line. Sense amplifier timing is designed so that worst case current generates sufficient ΔV which is defined in section 3.1.2 to take a valid decision. Due to randomness of both word line trigger and sense amplifier enable trigger, sense amplifier can be fired before a proper ΔV is generated. A reduced ΔV across the bit line results in access read failure. Therefore designing a proper sensing window is important to improve memory yield.

3.1.4 Diffusive or lossy line

A transmission line can be either considered as lumped interconnect or distributed transmission line [42]. If the circuit/line is less than $\frac{\lambda}{6}$ then it is analyzed as lumped interconnect and line greater than $\frac{\lambda}{6}$ is analyzed as distributed transmission line [42]. When bit line is layout for the column and word line is layout for row/word, the total length of the line can act as a lumped element or distributed element. Line can be lossy or lossless, the bit line or word lines being too great in length increase the resistance which then become comparable to the characteristic impedance of the line. Bit line and word line are

laid out using metals Cu or Al. The total resistance offered by a metal line is given by $R = R_s\left(\frac{L}{W}\right)$; where R_s is unit sheet resistance of the metal, L is the total length of the metal and W is width of the metal. Characteristic impedance is given by $Z_0 = \sqrt{\frac{L}{C}}$; where L and C are the inductor and capacitance per unit length. For a line to be considered lossless $R \ll Z_0$, maintaining this result, there is restriction on maximum length of the bit line/word line to avoid any diffusive delay. Additionally, the bit line length also depends $\left(\frac{I_{read}}{I_{off}}\right)$ ratio.

3.1.5 Summary of errors

Table 3.2 Summary of SRAM errors

Errors to be corrected in SRAM	Consideration in proposed work	Method to reduce the errors	Priority
Write failure	Partially	1. Big driving buffer 2. Cell sizing designed with write assist	Low/Addressed to reduce power consumption
Read Upset Failure	Partially	1. Proper Cell Sizing 2. Read assist circuit	Low
Hold Failure	No	1. Using High threshold voltage devices at expense in read delay 2. On chip LDO	Usually low, High when used in Sleep state
Bit line leakage	Yes	1. Statistical simulation of populated bit line with worst case memory data	High
Sensing window variation	Yes	1. Row delay and sense amplifier enable 2. Statistical simulation to check worst case delay	High
SA input offset voltage	Yes	1. Mismatch variation for SA	High
Read Current variation	Yes	1. Inter Die(Mismatch current variation) 2. Worst case current calculation	High
Access failure	Yes	1. Includes Bit line leakage, SA offset, Sensing window, Read current variation	High
Tracking systematic variation	Yes	1. Dummy column routine using ANDing logic	Medium
Efficient timing control	Yes	1. Self-timed replica delay line	High
Word yield	No	1. Error encoding technique 2. Dual location read/write for critical words	Medium

Table 3.2 shows summary of SRAM errors in SRAM, while suggesting a fix for efficient SRAM design. Most of the errors have been discussed in earlier sections, let us now discuss the new but meaningful errors introduced in table 3.1.5. Tracking systematic variation and efficient timing control will be discussed in next section where conventional memory read techniques and proposed memory read techniques are introduced. Memory cells at the edge or corners of the bank have a high probable of failing no matter how well they are statistically designed, this is part of the reason that dummy row and columns are placed outside the bank core. An error correction code technique can be used in SRAM designs. A well know error correction code technique is Hamming code, where parity bits are used to correct one bit nibble in the word. 32 bit hamming can detect 3 errors, correcting 1 bit per 32 bit word increases area and power by 18.75%. One of the other popular error encoding technique is Golay encoding, where for a 24 bit word 12 bit parity is used for error detection and 3 bits can be corrected. Correcting 3 bits per 24 bit nibble increases area and power by 100%. Hamming error coding looks more efficient over Golay code techniques. However Golay detects and corrects 3 bits which improves the yield. Error encoding techniques can be used provided that power and area budget is flexible. An application of both the techniques combined together might give effective result, where Golay code will be applied on extremes and hamming at central bits of the word. One of the dual location read/write address technique can be used in SRAM's. Here one of the bank (Correction bank) will be designed with more than the usual dummy rows and columns. Middle locations can be mapped with few top and bottom words of all other (main) banks. Using parity check, valid words can be decided. Such similar techniques are used in RAID type server memories. It is expected that most of the SRAM architecture should address these errors. Now that it very clear what effects the performance and yield of the SRAM let us discuss few conventional memory read techniques in next section.

3.2 CONVENTIONAL MEMORY READING TECHNIQUES

This section reviews the different memory read techniques used to improve the yield. Each technique has attempted to design the memory at the golden spot where yield and PPA can be achieved. These techniques are classified by the type of timing methods used in timing control block. There are four basic types of timing methods used in read timing control block and they are as follows.

1. Direct clocking [43]
2. Inverter delay line [44]
3. Self-timed replica delay line using dummy cells [45]
4. Pipelined timing using registers between the sense amplifiers [46]

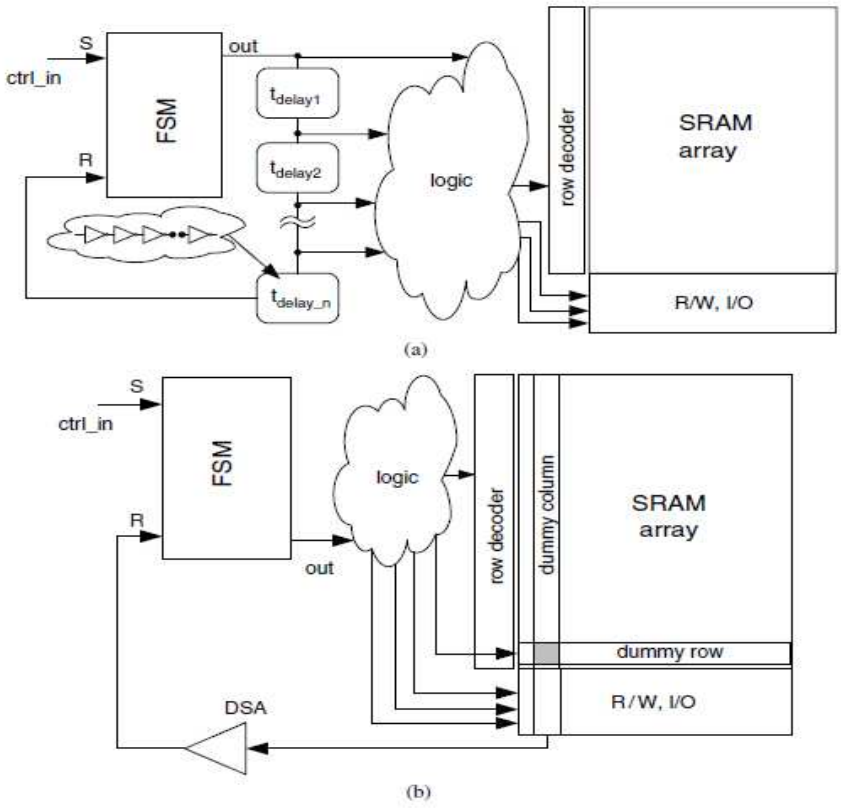


Fig. 3.5 (a) Inverter Delay Line (b) Self-timed replica delay line [3]

Inverter delay and self-timed replica delay lines are used often. Fig. 3.5.a and Fig. 3.5.b area a block level representation of inverter delay line timing and self-timed replica timing. In inverter delay technique, delay between different control signals are introduced using inverter chain. In fig. 3.5.a control signals are tapped after inverters to generate a delayed control signal. In self-timed replica delay timing, a localized reference signal is generated. A signal is then used to reset the FSM (Finite State Machine) and then fire all control signals using new time reference. Self- timed replica delay line uses local memory cells to generate local control signals (attempting to track local process variation) to fire the sense amplifiers and deactivates the word lines. [9, 41, 47] use inverter delay line and [39, 40, 45, 48, 49] self-timed delay timing techniques for memory read-write control signals.

- Inverter delay line [9, 41, 47]

In [41] a statistical device model is used to improve the yield, the paper concentrates on memory failures including read, write and hold failures. Monte Carlo simulations with process and mismatch variations together is carried for SNM, read SNM, write trip point, sense amplifier and read current. A timing analysis for 3σ worst conditions for SNM, read SNM, write trip point, sense amplifier and I_{read} are calculated. A 3σ yield is targeted here. With new or emerging process nodes and their increase in memory density a 3σ variation results lower and lower yields. This paper has failed to consider leakage current read considerations, sensing window variation, and sense amplifier input offset variation. These all increases the access failure rate affecting power consumption, performance and yield. The approach in [47] is similar to [41], however, a method to estimate cell design failure by calculating probability of failure of memory cell due to parametric variation is proposed. This paper fails to mention sense amplifier variation, bit line differential voltage variation and sensing window variation all of which resulting improper estimation compensation timing and reduced yield. In [9] an estimation yield loss due to read access failures is implemented, as this type of failure type has a strong impact on determining the performance and power consumption

of memory. This estimation methodology for access failures accounts for bit cell read current variations, sense amplifier input offset and sensing window variations as well as leakage currents from the remaining bit cells in the same column. This approach relies on a worst-case approach and analysis and well tend to result in an over design of the delay circuitry. Method proposed in [9] seems to be the most promising among inverter delay technique as the design addresses most of the high priority design failures discussed earlier. Process and mismatch variations together are used to design the timing control unit. However, there is still room to improve the timing by combining this technique with self-timed replica delay line.

- Self-timed replica delay line [39, 40, 45, 48, 49].

Self-timed replica delay line approach is briefly discussed in this section. Dummy columns are added as an assist in better cell timing. A dummy column has same number of memory cells as a single bit accessible memory column. Having similar memory cells tracks the parasitic bit line capacitance as of memory bit line capacitance. A replica timing circuit is designed for the time at which the worst case scenario memory cell in memory bank will generate sufficient ΔV to take a valid decision. Worst case timing t_{read} for ΔV is shown in Equation (3.1.1). Generating a differential bit line signal using dummy column gives a tighter ΔV_{dummy} distribution which provides precise timing to fire the sense amplifier and deactivate the word line. Dummy differential voltage ΔV_{dummy} is compared using a comparator with a reference $\left(\frac{V_{DD}}{2}\right)$ which sets the FSM to drive sense amplifier and deactivate word line signals. Equation (3.2.a) shows dummy trigger time which is comparable to t_{read} .

$$t_{dummy} = \left(\frac{C_{dummy}}{I_{reference}}\right) * \Delta V_{dummy} \tag{3.2.a}$$

Where, C_{dummy} is dummy bit line capacitance and is equivalent to the bit line capacitance of a memory column. $I_{reference}$ is replica current for reference and ΔV_{dummy} the dummy differential bit line voltage.

Comparing Equation (3.1.1) and (3.2.a) we get,

$$\left(\frac{C_{dummy}}{I_{reference}}\right) * \Delta V_{dummy} = \frac{CBL}{I_{cell}} * \Delta V \quad (3.2.b)$$

Since $C_{dummy} = C_{BL}, I_{reference} = \left(\frac{\Delta V_{dummy}}{\Delta V}\right) I_{cell}$ (3.2.c)

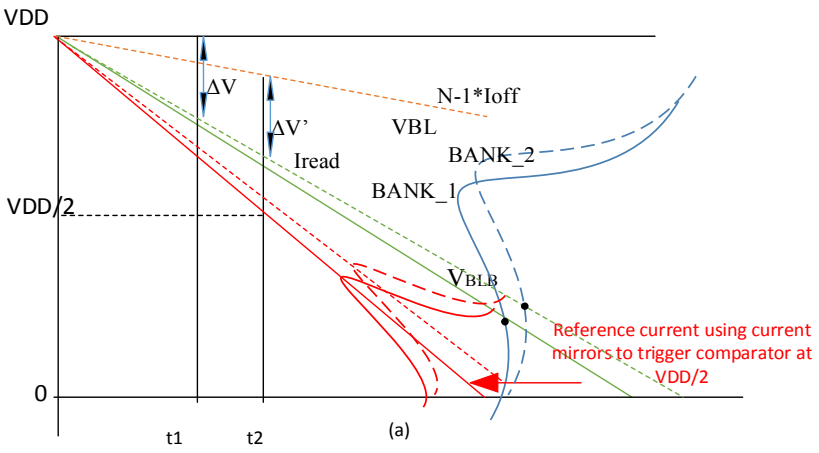
$$I_{reference} = \frac{\left(\frac{V_{DD}}{2}\right)}{\left(\frac{V_{DD}}{10}\right)} I_{cell} \; ; \quad I_{reference} = 5\, I_{cell} \quad (3.2.d)$$

Equation (3.2.d) can be achieved by designing 5 replica cells which are read on every read cycle. Therefore $I_{reference} = 5\, I_{cell}$ can be achieved by 5 replica cells in a dummy column. Another method to compare $t_{dummy} = t_{read}$ is by dividing the dummy bit line capacitance in required ratio.

The only difference between [39, 40, 45, 48, 49] is that each technique triggers different number of cells in a dummy column, also dummy columns are placed at different locations in memory bank. Self-timed replica bit line delay architectures discussed in [39, 40, 45, 48, 49] assumes that WID are only dominated by mismatch/local variation. If the memory is large in size then systematic variation will also contribute to WID variation. None of the architecture discussed the method to analyze and design for systematic variation. A new architecture is proposed in next section which addresses both process and mismatch variation issues. An attempt is made to improve the design for a process variant tolerant, low power and high performance memory.

3.3 PROPOSED STATISTICAL ARCHITECTURE

From [45] it is proven that the self-timed replica bit line delay techniques are more power and area efficient compared to inverter delay technique. Memory density is increasing with new process nodes and yield per die has become a concern. There will be tremendous power consumption with less performance if [9, 41, 47] memory architectures are used. The proposed idea uses a self-timed replica delay line technique to design a more accurate system compared to both conventional and new ideas discussed earlier. The proposed idea uses current mirrors to find mean read current for each bank in the memory. If the memory bank is large, then systematic variation becomes significant. Reference current generated from dummy columns can be less than the typical mean, large than the typical mean or equal/near to typical mean. The value of the reference current depends on the gradient of the wafer acting on the die. The Proposed architecture solves systematic and mismatch variation to design a precise sense amplifier timing to avoid access failure in memory bank.



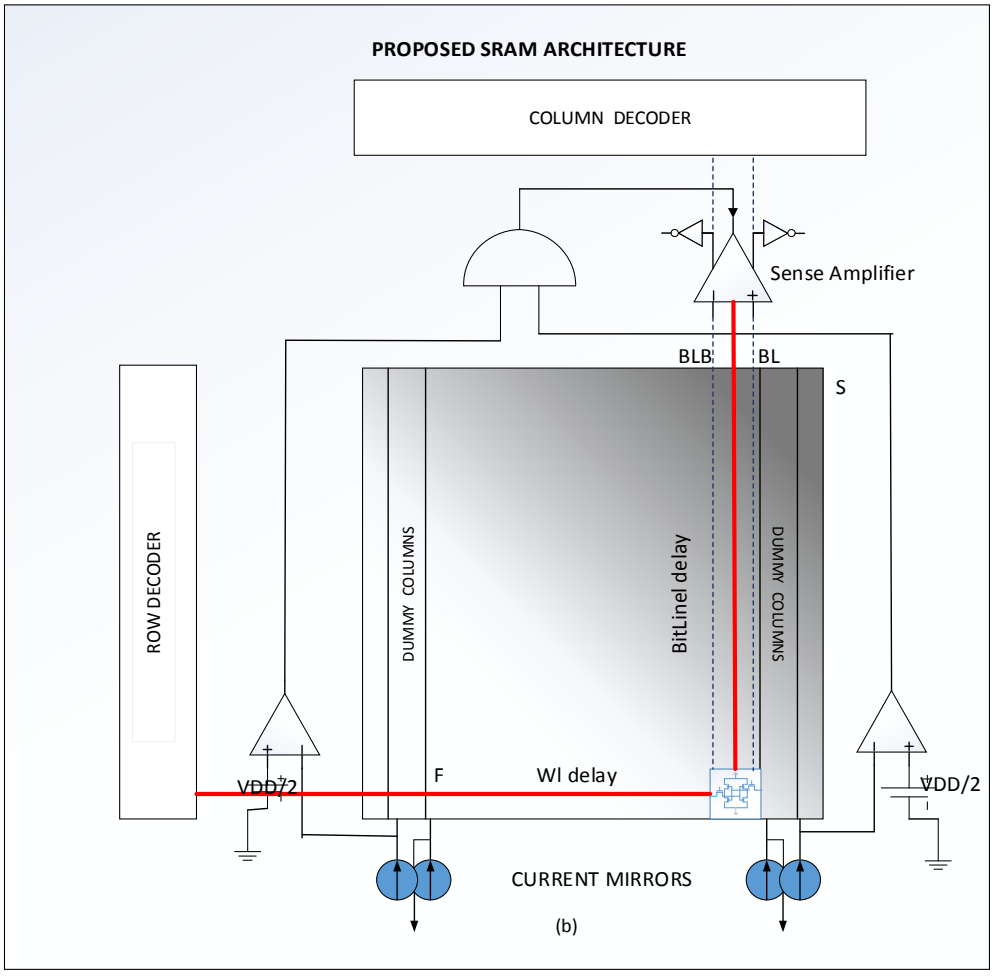


Fig. 3.6 (a) Reference and read current variation (b) Proposed SRAM Architecture

Fig. 3.6.a shows Reference and read current variation and fig. 3.6.b shows proposed SRAM architecture. In general, current mirrors are used to either scale current up or down by changing the (W/L) ratio of the devices [50]. As per the central limit theorem discussed in section 2.1.3, more than 30 sample size is adequate to have a sample mean approximately equal or closer to the population mean. Here, population would be parameters of all memory cell in a memory bank. Since WID variation is due to RDF, samples more than 30 will give better estimation of the population mean. More than 30 replica memory cell should be triggered which generates the $I_{reference} \cong 30\, I_{cell}$. $I_{reference}$ is divided using current mirror to a ratio where $I_{reference}$ can generate ΔV_{dummy} such that $t_{dummy} \geq t_{read}$.

Solid Red line shown in fig. 3.6.a is a reference current generated using current mirrors. Since the sample size is greater than equal to 30 replica cells the variation is tighter. Tighter variation makes design efficient. Memory is designed such that the worst case cell current achieves ΔV at t_{read}. The solid green line shows the bit line discharge due to $I_{cell} = \left(I_{read\mu} - n * \sigma_{Iread}\right) - (N-1) * Ioff + n * \sigma_{Ioff}$. To deal with systematic variation, dummy columns are placed at the extreme ends of the memory where worst case mean will be calculated locally. Slow reference dummy current among the two dummy columns triggers the comparator at $\frac{V_{DD}}{2}$ which then triggers the sense amplifier. Extreme dummy columns, AND gate logic, current mirror to find systematic and mismatch variation mean current, reduces the access failure rate. Once the current mirror ratio is fixed and the die moves on the wafer it will calculate its local reference current to trigger the sense amplifier as shown in red and green dotted line. The ratio between the worst cell current in the die and reference cell current remains approximately constant as die moves over the wafer.

This section shows a statistical proposed model which has considered most of high priority errors mentioned in table 3.2 leading to predictable yield specific read time for each bank in every memory on all wafers; Slow-Slow (SS), Typical-Typical (TT) and Fast-Fast (FF). This architecture is

capable of identifying the die position on the wafer and adjust the read access timing by generating the local control signals. This architecture can be more power consuming, but delivers high performance. This architecture certainly optimizes the power utilization in read operation but still lacks in putting efforts in improving write operation. In addition to this technique, all the design failures has been considered to design a smart stochastic/statistical approach which gives high performance, low power and maximum yield.

CHAPTER IV

IMPLEMENTATION

4.1 Memory Design Flow Chart

This chapter will discuss the simulation results and its interpretation. A memory design flow chart is shown in fig. 4.1. Sequential design steps are summarized to facilitate the steps in design of SRAM memory. SRAM architecture and cache requirements defines memory size and number of banks used to form the memory. However, further performance requirements addressed to verify a process can support the desired memory bank size. Once the number of banks are decided, the next step is to design memory cell geometry. Number of fingers in memory topology are swept or modified until the minimum noise margin requirements across process are satisfied. This ensures that the node geometry achieves a valid noise margin and that a design can be optimized to target both read/write performance and yield. At 0.18um process node, we can granularly increase the width and length of a transistor, but in this work we have strictly used the finger approach to understand what finger based design challenges other than variability will be offered by new process nodes.

Once memory geometries are fixed, statistical simulations are completed for read and leakage currents. 1) Selecting sense amplifier input offset voltage plus the minimum differential bit line voltage ΔV_{BL} approximately 0.1VDD and combing this with equation (3.1.2.b) arrives at a differential read voltage VDD/20 and a settling time of 2.5τ. Selection of a statistically significant ΔV_{BL} requirement was previously discussed in section [3.1.2]. 2) Calculate read timing for the worst case memory cell in the bank. Design dummy column and CTR (Current Transfer Ratio)

such that worst case memory cell has a reliable read. 3) Adjust CTR to achieve the read timing under TT wafer conditions. 4) Check if worst case memory cell on SS and FF process corners is read stable and as required adjust CTR such that all the worst case cell conditions of the memory bank on all the wafers; SS, TT and FF are read without failure.

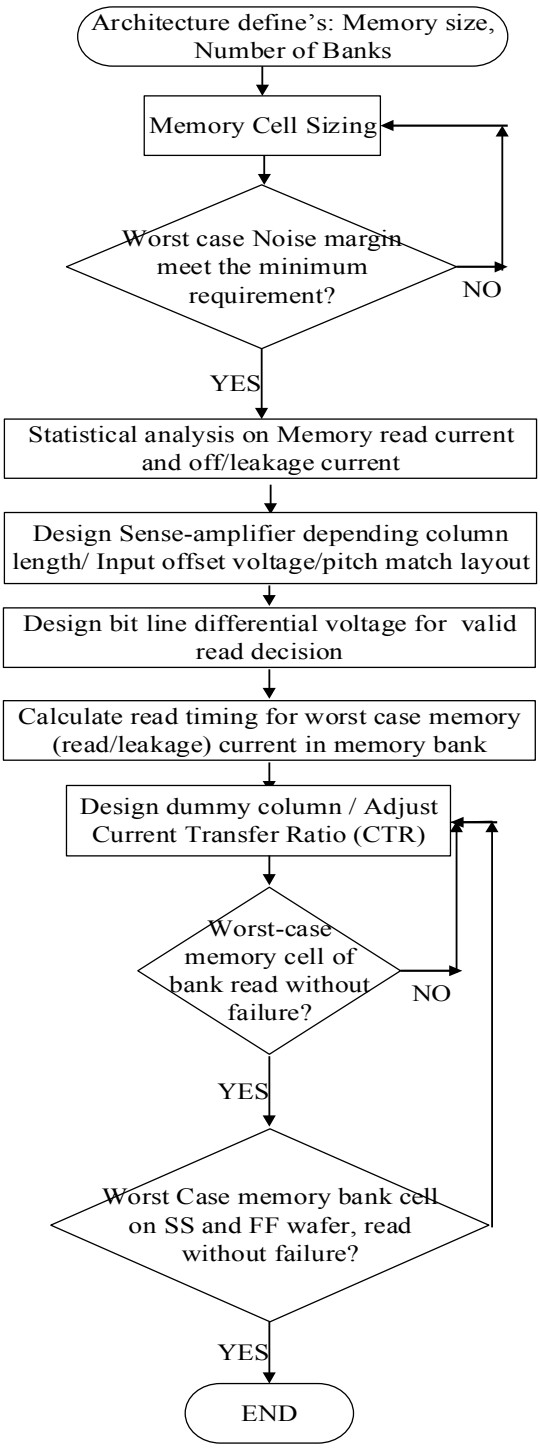


Fig 4.1 Memory Design Flow Chart

4.2 Memory cell design

Section 1.4, 1.5 and 1.6 discussed the basic workings of a memory cell and desired design requirements. Here we further discuss how to establish memory cell sizing. Before proceeding further the reader is suggested to revisit or recall [10-15] where N-curve and butterfly curve simulation techniques are discussed to define noise margin matrix. In this report both butterfly and N-Curve simulations are performed to design/conform memory cell sizing. There is a drawback in using only butterfly curve simulations to calculate noise margin of a memory cell. It is very difficult to perform an automated butterfly noise margin measures, also there is no information regarding the noise current sustained by the memory cell. An N-curve simulation allows for automated measures to perform statistical variation on read and write current and voltage noise margins in one single simulation run. A brief overview of noise margin matrix is reviewed below where SNM and RSNM are butterfly curve parameters and SINM, SVNM, WTI and WTV are N-curve parameters.

i) SNM (Static Noise Margin): Defined as static noise tolerated at the input of a device before switching its output state. Defined by the PU and PD transistors of a memory cell M1-M3 and M2-M4 shown in fig. 2.1.

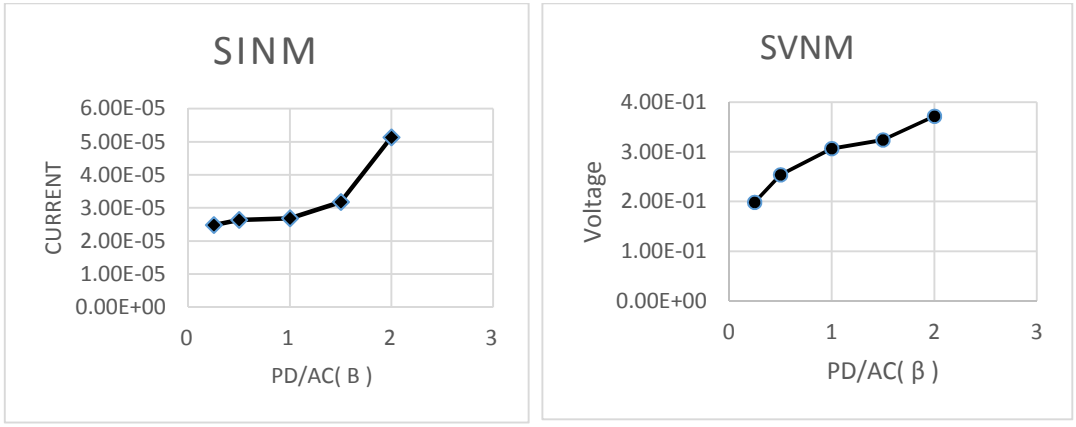
ii) RSNM (Read Noise Margin): Defined as the static noise tolerated by memory cell at the input before switching its output state during read operation. This is defined by the CR (cell ratio) of a memory cell.

iii) SINM (Static Current Noise Margin): The maximum current that can be injected at memory node before the memory cell switch's output state. SINM should be as high as possible.

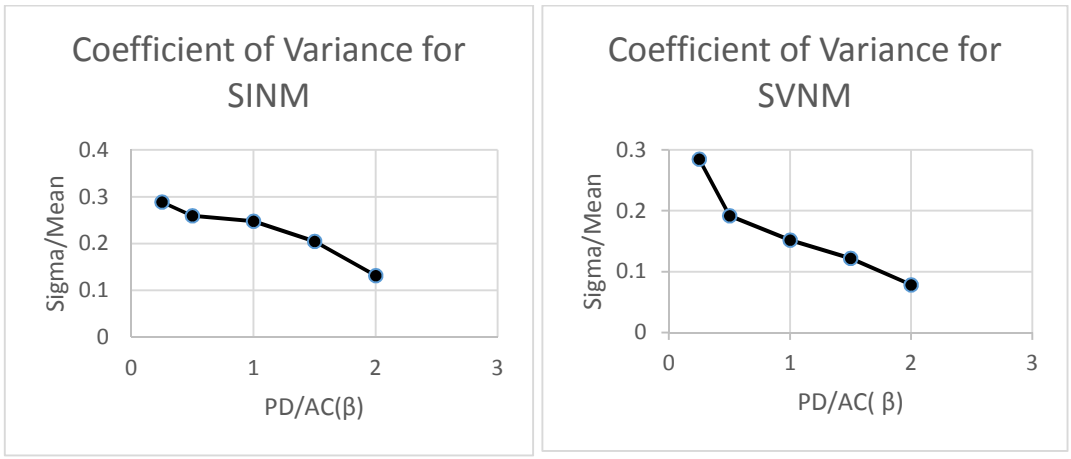
iv) SVNM (Static Voltage Noise Margin): The maximum DC noise voltage tolerable at the input of the inverter pair. SVNM is similar to SNM, but SNM is more conservative compared to SVNM.

v) WTI (Write Trip current): The amount of current required to write a cell when both the bit lines are held at high potential, this is similar to destructive read operation. A higher absolute value of WTI is sufficient enough to meet the read stability requirement.

vi) WTV (Write Trip Voltage): The bit line voltage drop required to write the opposite data on the node. A lower value of WTV results in less power dissipation during write operation.



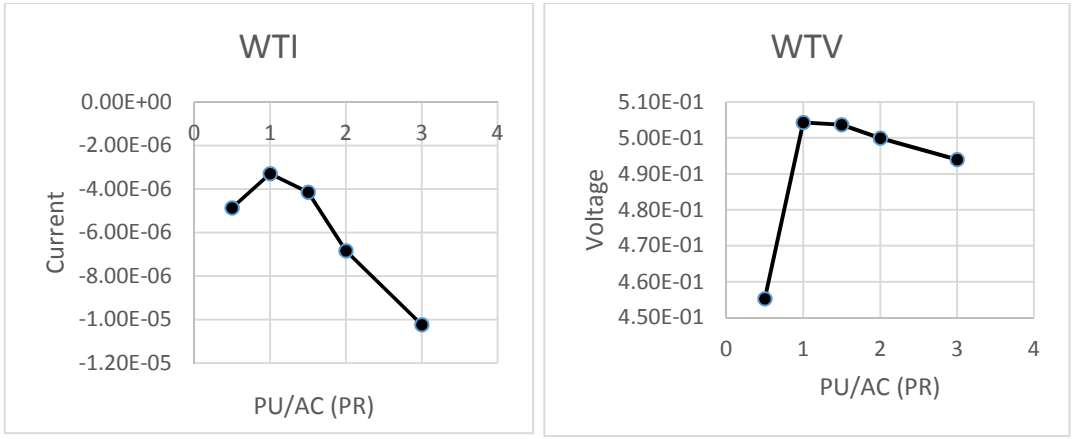
(a) (b)



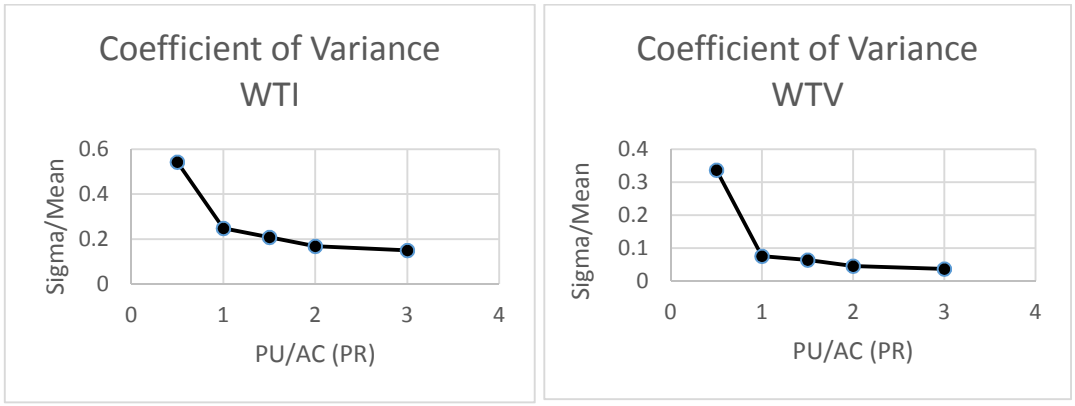
(c) (d)

Fig. 4.2 SINM, SVNM and Co-efficient of variance for different cell ratio

Fig. 4.2.1.1 (a) and (b) show SINM and SVNM for different cell ratio. It should be noted that current and voltage margin improves as we increase the cell ratio, also the coefficient of variance improves which indicates there is less variation in the noise margin at higher cell ratios.



(a) (b)



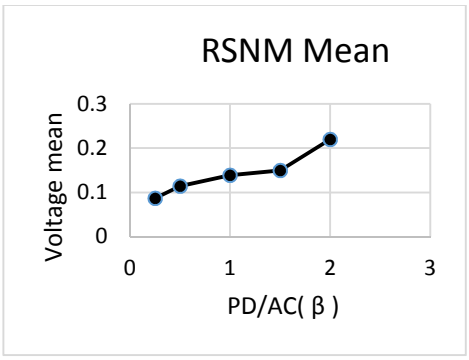
(c) (d)

Fig. 4.3 WTI, WTV and Co-efficient of variance for different cell ratio

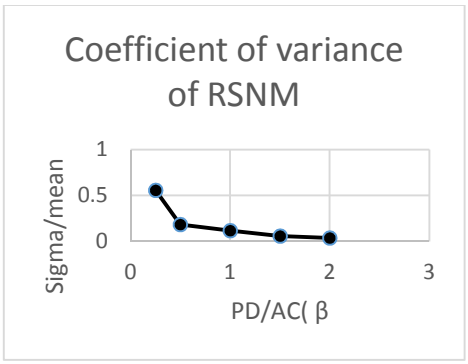
Fig. 4.3 (a) and (b) shows WTI and WTV for different cell ratio. In fig. 4.3 (b), when the access transistor is stronger than PU transistor, a much smaller bit line drop is required to write the opposite data save on the node. It is observed that access cell sizing contradicts with read and write margin.

Read margin desires the access transistor to be small compared to PD transistors whereas write margin desire the access transistor to be as large as possible. To solve this issue, the access transistor can be designed comparable to PU transistor but then requires a longer cell data write time. A second alternative is to decouple the read and write paths [49]. Dynamic write assist techniques are discussed in [49] which may be used to optimize write timings resulting in higher write yields. An Increase in PU also shows an improvement in write margin, however, this will increase the trip point of the inverter and degrade read margin. Since write operation is a positive feedback operation, it not the most significant issue. In fig. 4.3 an improvement in coefficient of variance is observed at higher PR, resulting from the increased area of the PU device. This again doesn't help in improving read margin but actually degrades in practice. AC transistors with high threshold drops less potential across itself which increases the node potential storing logic '0' and discharges '1' slowly. Using high threshold access transistors can improves the write margin but effectively degrades read margin. So a wise choice of PR to be selected is 1 and increase the word line signal length until the worst case memory is reliably written.

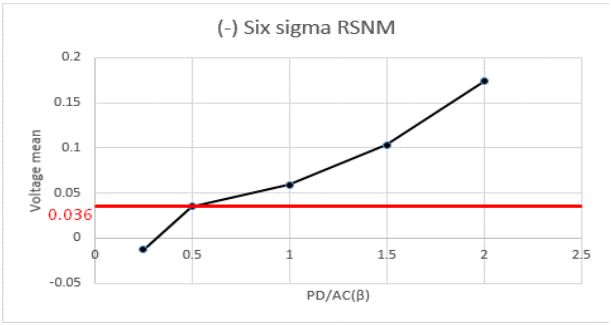
SNM is always greater than RSNM, hence if RSNM satisfies the worst case noise margin of $RSNM_{min} > 0.04 * V_{DD}$ for a 90% yield [3, 10] then SNM will surely satisfy the noise requirements. With a worst case $V_{DD} = 0.9V, RSNM_{min}\ should\ be > 0.036V$. A butterfly simulation is used to calculate RSNM. Fig. 4.11 shows (a) RSNM mean for different cell ratios, (b) Coefficient of variance for RSNM and (c) shows worst case read noise margin $RSNM_{\mu} - 6 * RSNM_{\sigma}$.



(a)



(b)



(c)

Fig. 4.11 RSNM mean, Co-efficient of variance and worst case margin RSNM

From fig. 4.11 (c) a cell ratio of 1 is sufficient to achieve 90% yield for worst case RNSM. But as the density increase with new process nodes, 90% yield is no longer sufficient. Therefore a design with cell ratio 1.5 and greater should be sufficient to avoid noise margin issues and this can be confirmed if memory test data is provided. Since with new process nodes width and length cannot be granularly increased, number of fingers must added in parallel and series to increase the width and length respectively. When not having the granular W and/or L increment option, the memory cell will be either under or over designed. The final memory cell size design is shown in fig. 4.3 where $\frac{PD}{AC}\ (\beta) = 2, \frac{PU}{AC}\ (PR) = 1$. Further simulations are carried on memory cell sizing shown in fig. 4.5.

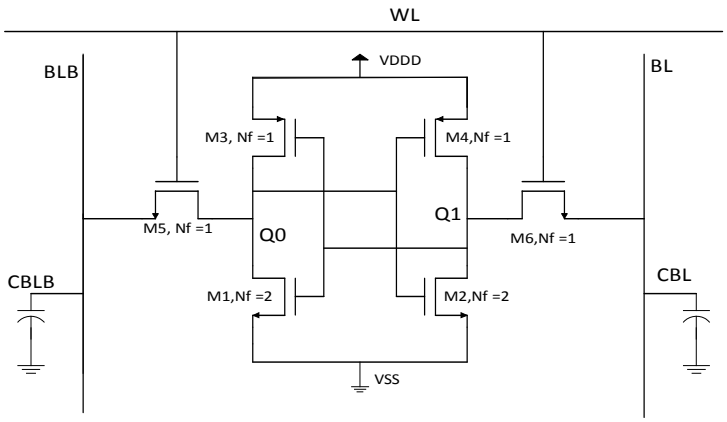


Fig. 4.5 Memory cell design

Table 4.1 and 4.2 shows statistical simulation results for read and leakage current across process and mismatch variation respectively. For every statistical simulation result, 99.9997% confidence interval is calculated with a 200 sample size. Once the confidence interval is achieved, a worst case value is selected for design. A small automated excel sheet is created to find the confidence interval.

Table 4.1 Monte-Carlo simulations on memory cell read current

Icell	(Process+mismatch)	Process	Mismatch T wafer	Mismatch S wafer	Mismatch F wafer
(μ,σ)	(30.078u, 5.5u)	(32.20u,4.646 u)	(32.02u,1.80 u)	(25.14u,1.49 u)	(42.90u,1.92 u)

Worst case $I_{cell} = I_{cell\mu} - n * \sigma_{Typical_{wafer_{Mismatch}}} = 32.180u - 3 * 1.80u = 26.78uA$

Table 4.2 Monte-Carlo simulations on off/leakage current

Ioff	N=128 (N-1)* Ioff_(Process+mismatch)
(μ,σ)	(5.229n,2.63292n)

Worst case $I_{off} = I_{off\mu} + n * \sigma_{total} = 5.229n + 3 * 2.6329n = 13.1277nA$;

$\frac{I_{cell}}{I_{off}} = \left(\frac{26.78uA}{13.1277nA}\right)$ = 2039 > 500, this is a sufficient Ion/Ioff ratio for memory design, but it is important to note that in modern process nodes this number will be difficult to achieve with high leakage current in devices. However one advantage of forth coming FinFETs and SOI devices have improvement of the subthreshold slope to '1' which alone will greatly improve Ion/Ioff ratios.

Layout of a memory cell is show in fig.4.3. A conventional memory layout technique is used here. M1 (Metal 1) is used for V_{DD} , M2 (Metal 2) is used for ground V_{SS}, M3 (Metal 3) is used for Word line WL and M4 (Metal 4) is used for bit lines BL and BLB. Higher metal provides less resistance and hence it is favorable using high level metal for bit lines and word lines to give high performance. It is observed that 18% area is increased in layout when we go from minimum geometry memory cell size to cell ratio of 2 [8]. A higher cell ratio reduces variation by 47% and increases performance by nearly 100% resulting in better cell stability, reduced variability, better performance and high yield.

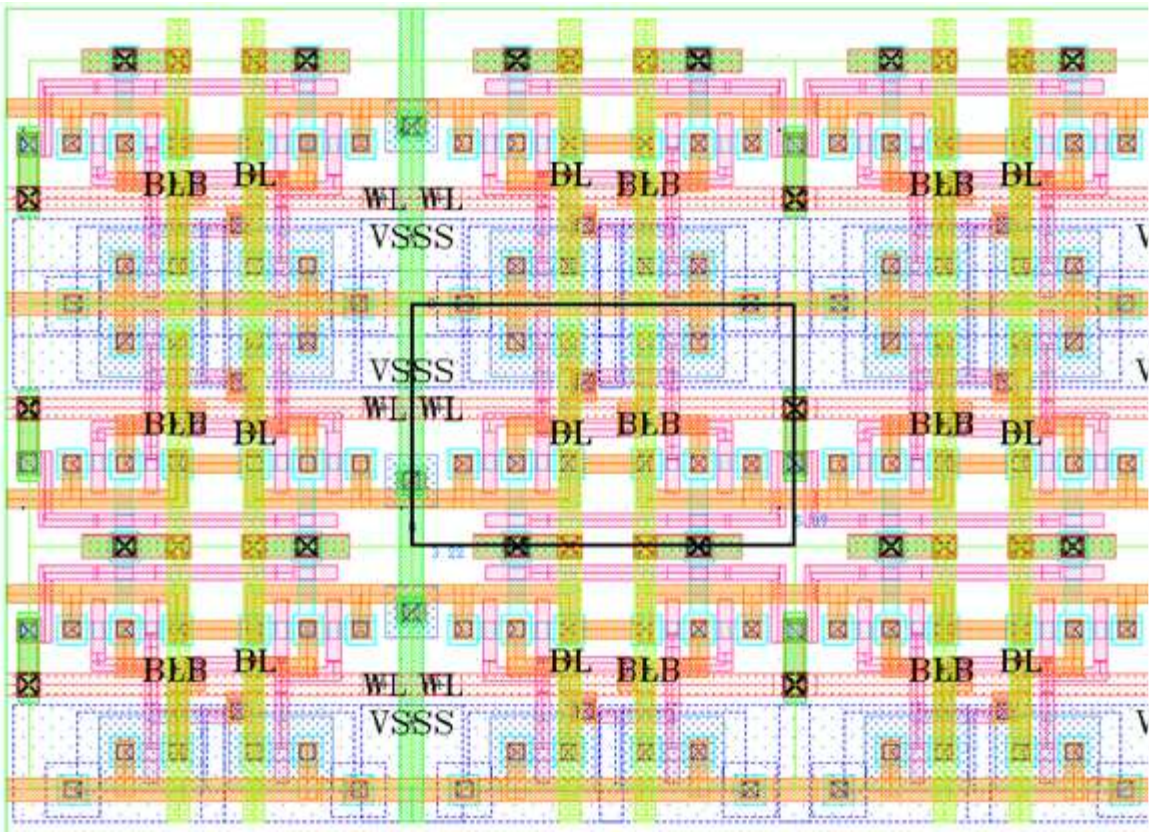


Fig. 4.6 3x3 Memory layout

A 3x3 memory cell layout is shown in fig. 4.6. Memory cells are abutted together to share V_{DD}, body bias and V_{SS} to reduce the effective area of a memory cell. Effective total area of memory cell achieved is 5.09umx3.19um = 16.24um^2. In [8] a road map of memory layout cell area for each process node is shown. A 'thin bitcell' layout approach can be used to reduce memory cell layout area [2], here sources and drains are shared while abutting cells. Lower metals in a single memory cell area offers higher capacitance compared to higher metals used in fig. 4.6 and there will be a tradeoff between area and performance. Using 'thin bitcell' layout topology, the 0.18um process node with 1:1:2 ratio design can achieve 6-8um^2 cell area which fit's on the memory cell area curve mentioned in [8].

4.3 Sense Amplifier Design

4.3.1 Design Method

A sense amplifier is shown in fig. 4.7. M1-M2 are PMOS cross couple pair used to pull up the output voltage to V_{DD}. M4-M5 are differential pair used to sense bit line differential signals and amplify bit line difference. M3 is a reset switch and M6 is a tail current transistor which maintains equal current through differential pair. A low clock signal is applied at M3 to reset the output voltage Vom and Vop. M1 and M2 gets shorted raising $and\ Vop = V_{DDD}$. Once the bit lines generates sufficient differential voltage ΔV_{BL}, SAEN (Sense Amplifier Enable) goes high which turns on M6 and after some 'ps' delay RESET switch is turned off. An initial voltage is generated at Vom and Vop and either of the sense amplifier legs starts pulling more current whereas the other leg reduces in current. A pull up and pull down cross couple operation identify logic '0' and logic '1' with sufficient ΔV_{BL}. Vom and Vop are further connected to minimum geometry inverter triggering or latch circuit creating master slave operation for clean logic '0' and '1'.

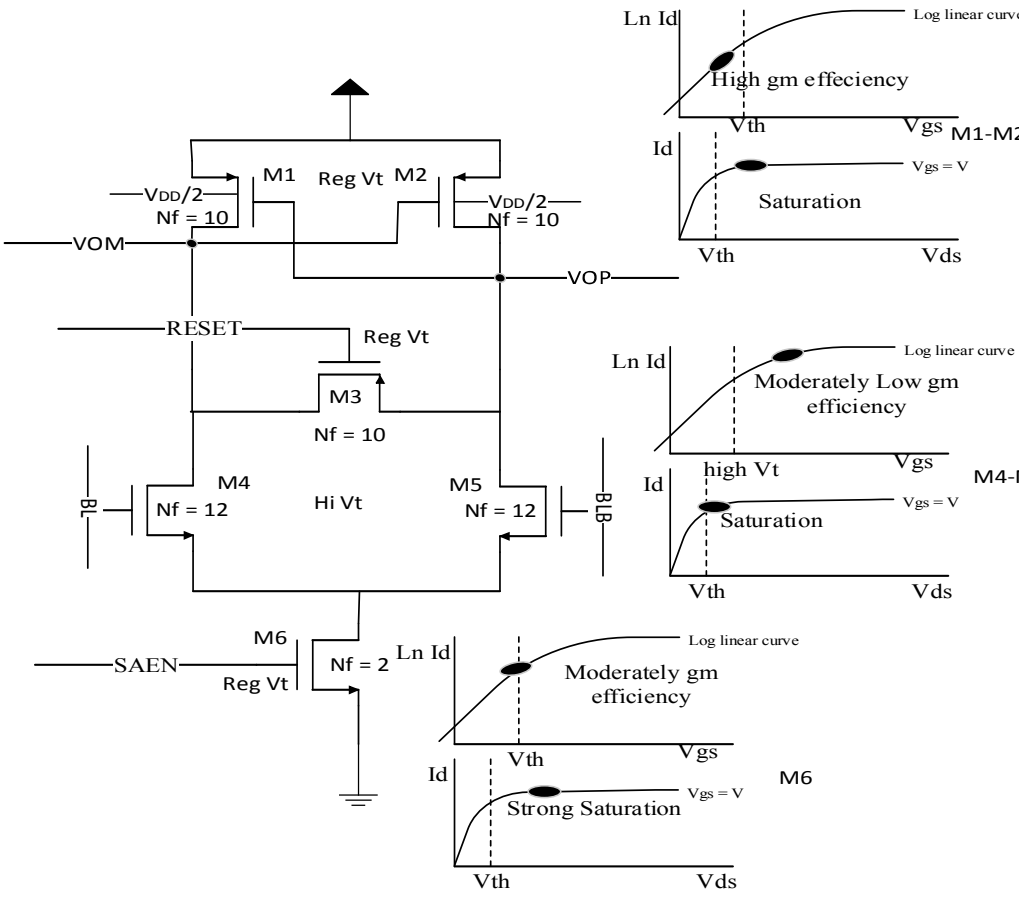


Fig. 4.7 Sense amplifier, operating region of transistors and geometries

The challenge in sense amplifier design is in proper layout and in estimating area to manage the input offset voltage. It is discussed in section 3.1.2, the ΔV_{BL} requirement is dependent on sense amplifier input offset voltage. Ideally, an infinite area can push input offset voltage to zero, but is quite impractical. There are practical limitations on sense amplifier area, usually (2 to 16) columns are multiplexed with a single sense amplifier. An area which in sense amplifier can be increased until it can pitch match the multiplexed column width. Settling time is one of the important design parameter in sense amplifier. Recalling equation (3.1.2.b)

$$tset = \tau \ln\left(\frac{VoH-VoL}{V_{BLmin}}\right) \approx \frac{Cgp}{gmp} \ln[\frac{Voh-Vol}{VBL_{min}}]$$

$\tau \ll t_{BL_{discharge}}$, settling time of the SA should be very small as compared to bit line discharge time. $\tau = Cgp/gmp$, gm of the cross couple device is set by the current flowing through each leg of the sense amplifier. M6 is the tail current device which set's the current flowing through each leg of the sense amplifier. Thus, settling time is controlled by the PMOS pair indirectly by the biasing of the tail transistor controlling the current. M6 should ensures that the differential current is balanced in both the legs, throughout the process to track sense amplifier performance. M1-M2 and M4-M5 are biased such that all the transistors are in saturation region. Once the DC biasing and settling 'gm' is achieved the area of the sense amplifier can be increased to achieve the required input offset voltage. In a nutshell to attain BW, current density at 0.9 VDD must be set by the correct combination of tail width and M1-M2 width and then increase all area as required to control Vos. It is derived in equation (4.3.2.d), the input offset voltage of a sense amplifier also depends on M1-M2 mismatch. In order to achieve the offset PMOS area should be increased and in this case BW is compromised as 'gmp' reduces. Using M4-M5 as high threshold transistors which has increased number of dopants in the channel reduces the input offset voltage.

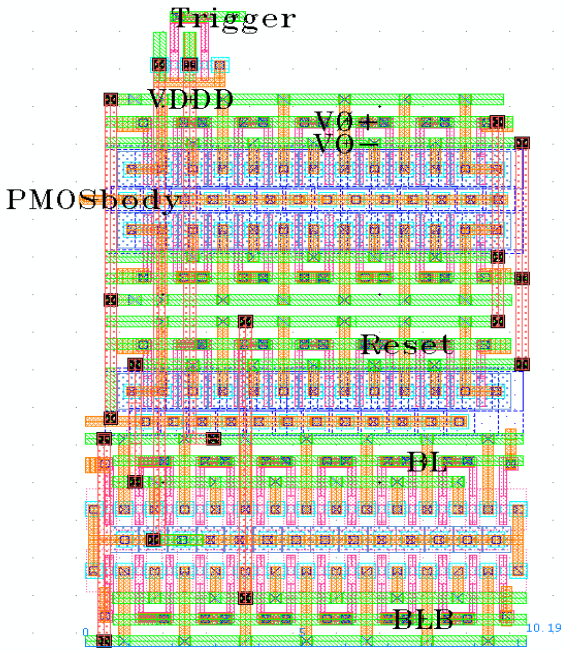


Fig. 4.8 Sense Amplifier Layout

A sense amplifier layout is shown in fig. 4.8. A sense amplifier layout is designed such that it is pitch match for two column widths. Sense amplifier can be stacked upon each other and with the upper sense amplifier connected to the second column. As another alternative, two columns can be multiplexed together to use one sense amplifier.

4.3.2 Input offset voltage

Input offset voltage is defined as the additional voltage required to balance the differential pair due to threshold mismatch in differential pair devices. Total output offset current ios of the sense amplifier can be written using equation (4.3.2.a).

$$ios = \sqrt{(ios_1)^2 + (ios_4)^2} \qquad (4.3.2.a)$$

From $ios = gm * vos$; offset current or offset voltage present is a reflection of transistor. The implication is that for small transistor mismatch either vos or ios can be represented through gm. Equation (4.3.2.a) can be modified to Equation (4.3.2.b).

$$ios = \sqrt{(gmp_1 * Vosp_1)^2 + (gmn_4 * Vosn_4)^2} \tag{4.3.2.b}$$

Substituting mismatch equation (2.2.1.c) in equation (4.3.2.b) we get equation (4.3.2.c)

$$ios = \sqrt{\left(gmp_1 * \frac{Avtp}{\sqrt{WLp1}}\right)^2 + \left(gmn_4 * \left(\frac{Ahvtn}{\sqrt{WLn4}}\right)\right)^2} \tag{4.3.2.c}$$

To find the input referred offset voltage, equation (4.3.2.c) is divided by differential pair gm.

$$\sigma_{input_{offset\_3sigma}} = \sqrt{\left(\left(\frac{gmp_1}{gmn_4}\right) * \frac{Avtp}{\sqrt{WLp1}}\right)^2 + \left(\frac{Ahvtn}{\sqrt{WLn4}}\right)^2\Bigg)} \tag{4.3.2.d}$$

Using M1-M2 as 10 fingers and incrementing differential pair M4-M5 number of finger, input offset voltage can be observed as decreasing. Theoretical input offset calculation using equation (4.3.2.d) is shown in fig. 4.3.2. A matlab code is used to calculate input offset voltage for different differential pair sizing. Pelgrom co-efficient values are provided by IBM in PDK. A 12 finger differential pair produces 8.3mV 1sigma input offset voltage. Historically, theoretical and simulation layout, mismatch match closely. While achieving the required input offset voltage. Beyond 4 to 6 fingers a point of diminishing returns appears.

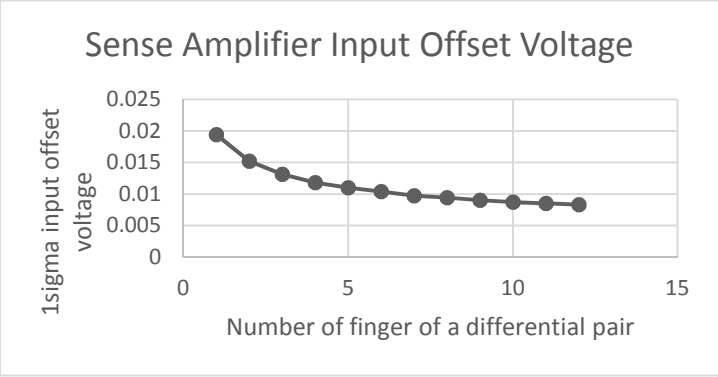


Fig. 4.9 Theoretical input offset voltage of a sense amplifier

Following are the simulation steps used to observe the input offset voltage.

1. Switch ON M3 and M6 simultaneously
2. Apply common mode input voltage to differential pair
3. Find offset current $Ios = id1 - id2$
4. Find average differential $gm_{diff} = \frac{gm_4 + gm_5}{2}$
5. Generate an equation for input offset voltage as $Vos = \frac{Ios}{gm_{diff}}$
6. Run mismatch Monte-Carlo for N points

Fig. 4.10 shows Monte-Carlo simulation output for an input offset voltage of a sense amplifier

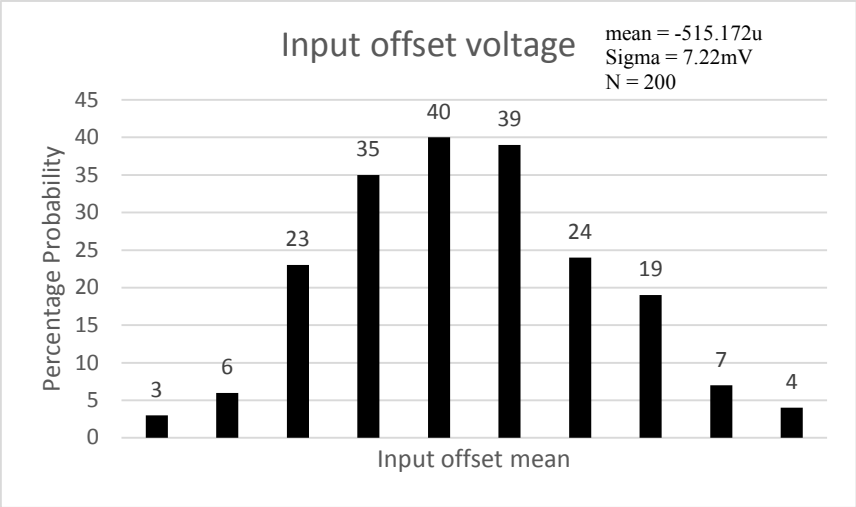


Fig. 4.10 Simulation input offset voltage of a sense amplifier

Input offset voltage from calculation and simulation via the theoretical PDK formula closely matches each other. A worst case input offset voltage is $\pm 3\ \sigma_{input_{offset}} = 21.6mV$. Observe that for 18nm process Vos would approach $\pm 3\ \sigma_{input_{offset}} = 36mV$. Even so read delay will be considerable faster as Icell/C_{BL} remains relatively constant.

4.3.2 Settling and Reset timing

Settling time and reset time variations are due to process variation in sense amplifier. Worst case design for settling and reset timings of a sense amplifier makes sure that Vom and Vop have reached their final values in order to take the decision. Settling and reset time are given in equations (4.3.2.a) and (4.3.2.b).

$$tset = \frac{C_{gsp1}+C_{dbn4}+C_{dbp1}}{gmp_1}\ln\left(\frac{VoH-VoL}{3\sigma Vos}\right) \quad (4.3.2.a)$$

$$= \frac{2.044f+674a+694.2a}{36u}\ln\left(\frac{900mV-100mV}{25mV}\right) \cong 350ps$$

$$treset = \frac{C_{gsp3}+C_{dbn4}+C_{dbp1}}{gmp_3}\ln\left(\frac{4*(VoH-VoL)}{3\sigma Vos}\right) \quad (4.3.2.b)$$

$$= \frac{1.504f+674a+694.2a}{32u}\ln\left(\frac{4*(900mV-100mV)}{25mV}\right) \cong 300ps$$

Table 4.4 shows settling of sense amplifier across the process. A minimum geometry inverter/latch circuit at the sense amplifier output should be triggered after >350ps ensuring the output is settled and the reset switch should be on at least for > 300ps to reset the output.

Table 4.4 settling and reset time across process

Timings	SS	TT	FF
tset	333.21ps	160ps	90ps
treset	467.75ps	240ps	172.42ps

4.4 Trip Comparator Design

This and following section discusses the comparator and dummy column design approach generate the replica delay timing. The replica delay line routine is designed such that a reference current is generated to charge the dummy bit line to Vref and trigger the sense amplifier to reliable read the worst case memory cell. A trip comparator is designed to trigger the sense amplifier when dummy bit line is charged to $\frac{V_{DD}}{2}$. A trip comparator mechanism and a schematic diagram of error amplifier is shown in fig 4.4.a. and 4.4.b. Vref is generated using two PMOS diode connects in their own well as shown in fig.4.4.c. A beta match inverter with unity feedback is shown in fig.4.4.a. Due to unity feedback and beta matched pair Vtrip is set at

$$V_{trip} = \frac{(V_{DD}.\beta p - |Vtp|.\beta p + Vtn.\beta n)}{(\beta n + \beta p)} = \frac{V_{DD} + Vtn - |Vtp|}{2} \quad (4.4.1)$$

V_{trip} is adjusted to $\frac{V_{DD}}{2}$ by controlling the body potential of a PMOS device only in this case since |VTP|> VTN. An error amplifier maintains the body potential to track Vref changes over the process and maintains $V_{trip} = \frac{V_{DD}}{2}$, but only to the extent that Vref is valid. If the trip point of the comparator is not "constant" and does not track the supply voltage and process, then access failure may occur. Width and the accompanying power to control Vos of the PMOS divider can be a concern and can be shared and averaged if we place one per each SA or one per 2 SA.

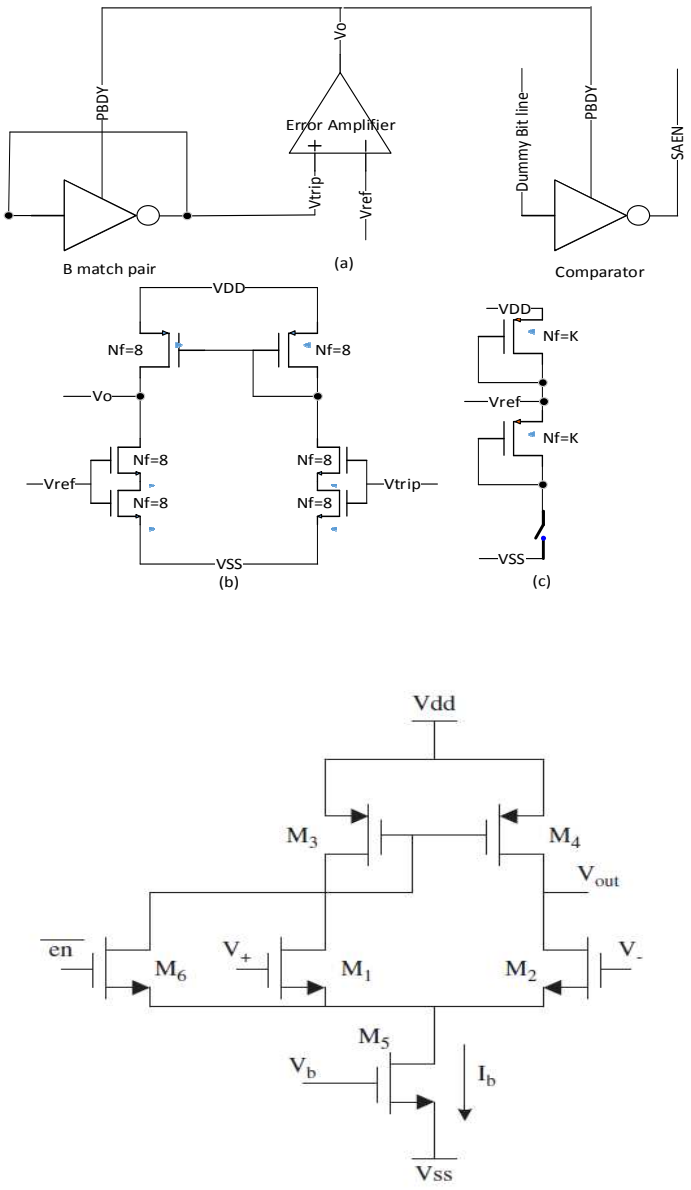


Fig 4.11 (a) Trip comparator (b) error amplifier (c) Vref (d) Enable trip comparator

Vref is expected to be constant throughout the process variation and track VDD, ensuring the worst case memory cell timing are tracked by dummy column replica timing as well. Due to RDF in PMOS devices Vref is not constant throughout the process and so the standard deviation of Vref can be given by equation (4.4.2).

$$\sigma_{Vref} = \sqrt{2} \cdot \frac{Avtp}{\sqrt{K.WL}} \quad (4.4.2)$$

Number of fingers (K) are added in PMOS until the worst case offset voltage at Vref is valid.

In fig 4.4.a, a comparator compares the input difference and amplifies the output until Vtrip equals Vref. A negative feedback circuit is used to balance the Vtrip at Vref. A negative feedback equation for the trip comparator is shown below.

$$Vo = Avol * (Vtrip - Vref) \quad (4.4.3)$$

Vtrip in terms of Vo can be mentioned as shown in equation (4.4.4)

$$Vtrip = -up.\chi.r_{on}.Vo \quad (4.4.4)$$

Where $\chi = \frac{\gamma}{(2\sqrt{(2\emptyset f + VSB)})} = \frac{gmb}{gm} = 0.1 - 0.3$ [50]

Substituting equation (4.4.4) in (4.4.3)

$$Vo = -Vref.\frac{Avol}{1+Avol.up.\chi.r_{on}} \quad (4.4.5)$$

A negative feedback is applied generated and control the Vtrip. DC gain $Avol$ should be sufficiently high enough to support the negative feedback accuracy. Avol is given by equation (4.4.6) where gmn and gmp are 'gm's'of error amplifier.

$$Avol = \left(\frac{gmn}{gmp}\right)_{err} \quad (4.4.6)$$

Comparator chain is designed to drive a number of (word length) sense amplifiers. A PMOS current mirror in the design uses 20 fingers and can drive 80x buffer which can drive 320 devices in a word length sense amplifier. A proper inverter chain is designed to achieve the desired bandwidth.

Body bias controls the threshold voltage of the PMOS and Vtrip is adjusted through negative feedback shown in equation (4.4.1) and (4.4.5). A robust trip comparator is designed to increases the system accuracy. Vtrip statistical distribution is shown in fig 4.12.

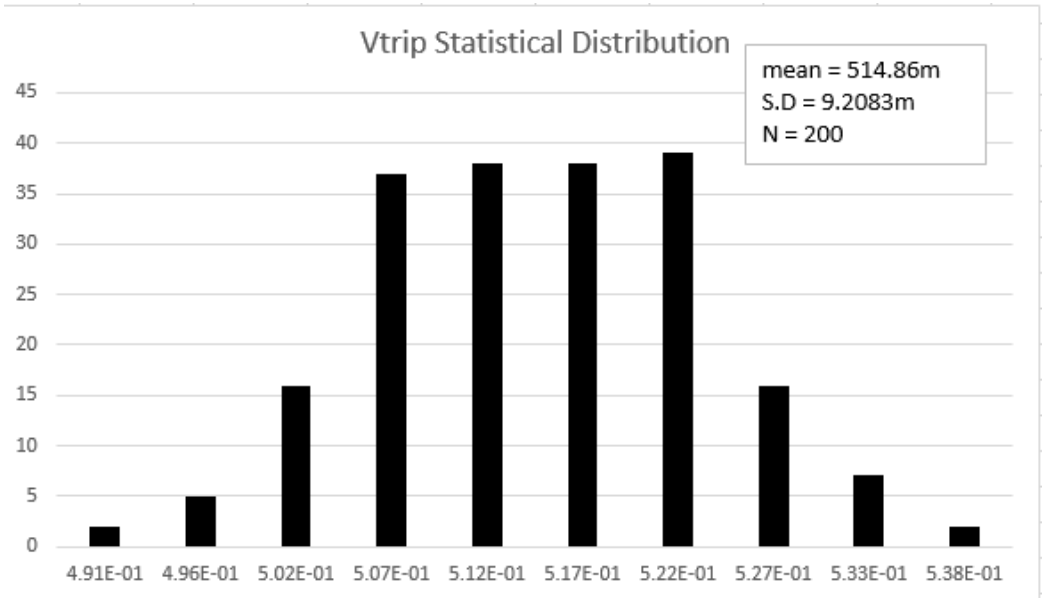


Fig. 4.12 Vtrip statistical distribution

Considering the worst case scenario, dummy column replica delay line should be designed for

$$Vtrip = Vtrip_{\mu} - 3 * \sigma_{trip} = 514.86mV - 3 * 9.2mV \cong 490mV$$

It should be noted that, a tailless trip comparator draws lot of current and an alternative trip comparator with enable signal can be used as shown in fig. 4.11.d. In enable trip comparator the tail current is switched with the bank select. Enable bar signal diode connects M6 and M3 and M4 forms a current mirror. A Vref is connected at one input terminal and dummy column capacitance to the other input terminal. One the dummy bit line charges above the Vref, enable trip comparator will trigger the sense amplifiers. The BW of the enable trip comparator is given in equation (4.4.7)

$$BW = \frac{gmn_{1-2}}{cpp} = \frac{gmn_{1-2}}{C_{dp}+C_{dn}+C_{bufferin}} \quad (4.4.7)$$

In equation (4.4.7) gmn_{1-2} is the differential pair 'gm' of M1 and M2 device (equal) shown in fig. 4.11.d. and $C_{bufferin}$ is the input buffer capacitance of the optimized comparator chain used to drive word length sense amplifiers. Differential pair 'gm' should be adequate to achieve the BW.

4.5 Dummy Column Design

A dummy column consists of identical or replica memory cells as in the memory bank columns. When the word line is activated in the memory bank, greater than 30 replica cells are triggered in each column. This current is summed by a PMOS current mirrors. Current is mirrored using the CTR (Current Transfer Ratio) to find a scaled mean current for the memory bank. The average current charges the dummy bit line to Vtrip tripping a comparator and enabling the sense amplifier. The reconfigured dummy column is shown in fig. 4.13. PMOS current mirrors are used to average and scale the bit lines current, but diode connected PMOS offers high resistance and the voltage drop is too high to support the dummy cell current. As a result less drain to source voltage is available and both PG and PD are in triode region. This issue can be resolve by modifying the memory cell in a dummy column.

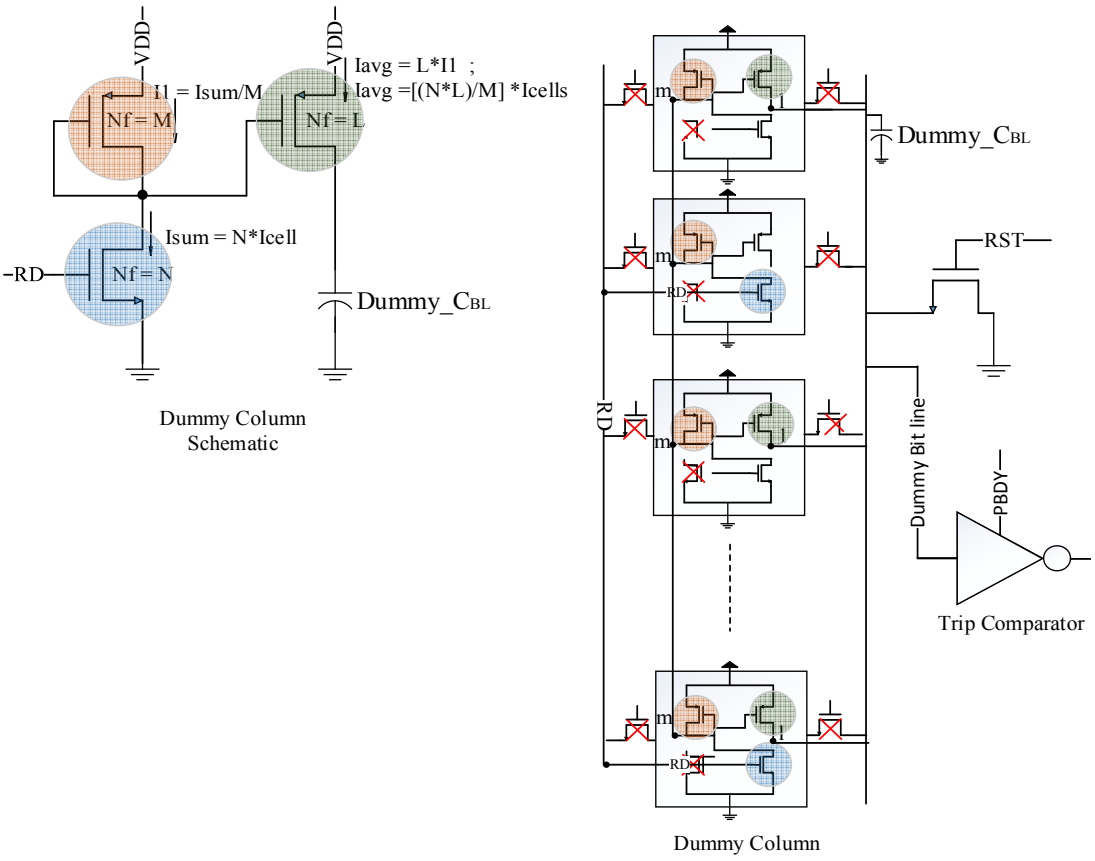


Fig. 4.13 Dummy column (red X indicates modifications to the dummy line)

Before discussing dummy column memory read current in greater detail. The memory read current topology shown in fig. 4.14 is reviewed.

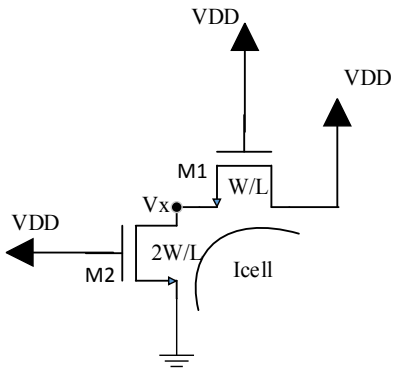


Fig. 4.14 Memory Cell Read Current

$$I_{cell} = I_{D1} = I_{D2}$$

(4.5.1)

$$I_{D1} = \text{Ksatn}(V_{DD} - V_x - V_{TN}) \quad (4.5.2)$$

$$I_{D2} = \frac{1}{2}\mu_n C_{ox}\frac{2W}{L}(2(V_G - V_{thn})V_x - V_X^2) = \text{Ksatn}(V_{DD} - V_x - V_{TN}) \quad (4.5.3)$$

Equation (4.5.1) holds true when M1 is in saturation and M2 is in triode. Since M2 is in triode region, the variation in M2 strongly dominates the cell performance.

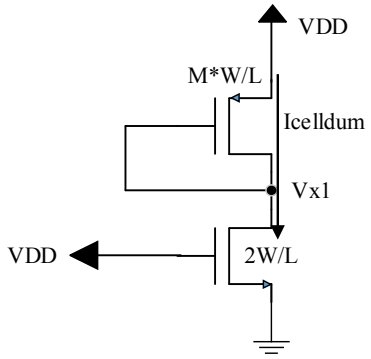


Fig. 4.15 Dummy Cell Read Current

A modified dummy memory cell is shown in fig. 4.15 where a PMOS is diode connected to measure the read current. As discussed earlier, PMOS current mirrors cannot be used to measure the current from bit line due to the resistive drop. This can be used to advantage with the topology in fig. 4.15.

Since the drop will be high enough, that NMOS in fig. 4.15 will always operate in triode region similar modestly matching memory cell current in equation (4.5.3)

$$I_{celldum} = \frac{1}{2}\,\mu_n C_{ox}\frac{2W}{L}(2(V_G - V_{thn})V_{x1}\ - V_{X1}^2) = \text{Ksatp}(V_{DD} - V_{x1} - |V_{TP}|)$$

(4.5.4)

To achieve $I_{cell} = I_{celldum}$ we compare equation (4.5.3) and equation (4.5.4).

$$\frac{Icelldum}{Icell} = \frac{\frac{1}{2}\mu_n C_{ox}\frac{2W}{L}\left(2(V_{DD}-V_{TN})V_{x1}\ - V_{X1}^2\right)}{\frac{1}{2}\mu_n C_{ox}\frac{2W}{L}\left(2(V_{DD}-V_{TN})V_x\ - V_X^2\right)} = \frac{\text{Ksatp}(V_{DD}-V_{x1}-|V_{TP}|)}{\text{Ksatn}(V_{DD}-V_x-V_{TN})} \tag{4.5.5}$$

$$\frac{Icelldum}{Icell} = \frac{\left(2(V_{DD}-V_{TN})V_{x1}\ - V_{X1}^2\right)}{\left(2(V_{DD}-V_{TN})V_x\ - V_X^2\right)} = \frac{\text{KR}(V_{DD}-V_{x1}-|V_{TP}|)}{(V_{DD}-V_x-V_{TN})} \tag{4.5.6}$$

$$\frac{Icelldum}{Icell} = \frac{\left(2V_{x1}\ - {V_{X1}^2}/{(V_{DD}-V_{TN})}\right)}{\left(2V_x\ - {V_X^2}/{(V_{DD}-V_{TN})}\right)} = \frac{\text{KR}(V_{DD}-|V_{TP}|-V_{x1})}{(V_{DD}-V_{TN}-V_x)} \tag{4.5.7}$$

$$\frac{Icelldum}{Icell} \approx \frac{V_{x1}}{V_{x2}} \approx \frac{\text{KR}(V_{DD}-|V_{TP}|-V_{x1})}{(V_{DD}-V_{TN}-V_x)}$$

$$Icelldum \approx Icell\ \frac{\text{KR}(V_{DD}-|V_{TP}|-V_{x1})}{(V_{DD}-V_{TN}-V_x)} = \text{KR}\frac{(V_{DD}-|V_{TP}|)}{(V_{DD}-V_{TN})}\left(1-\frac{V_{x1}}{(V_{DD}-|V_{TP}|)}\right)\left(1+\frac{V_x}{(V_{DD}-V_{TN})}\right) \tag{4.5.8}$$

$$Icelldum \approx Icell x\text{KR}\frac{(V_{DD}-|V_{TP}|)}{(V_{DD}-V_{TN})}\left(1-\frac{V_{x1}}{(V_{DD}-|V_{TP}|)}+\frac{V_x}{(V_{DD}-V_{TN})}\right) \tag{4.5.9}$$

$$Icelldum \approx Icell x\text{KR}\frac{(V_{DD}-|V_{TP}|)}{(V_{DD}-V_{TN})} \tag{4.5.10}$$

$V_x = V_{x1}$, which can be achieved by adjusting KR (Increasing number of fingers in PMOS). It is observed that when 3 fingers are used, $I_{celldum} \cong I_{cell}$ and follows same statistical variations. It obvious to the casual designer that the real difference is 1) difference between ($|V_{TP}|$ and V_{TN}) 2) KR both forcing dummy cell current deeper in to triode. The CTR (Current Transfer Ratio) ratio is derived using current mirror logic. [50] presents detailed understanding of current mirror and current mirror ratio.

$I_{avg} = \left(\frac{N*L}{M}\right) * I_{cell}$ Where, N are the number of memory cells trigger in a dummy column. M are the number of fingers in a PMOS diode connect, $M \geq 3 * N$ and L are the number of fingers used

to generate the average scaled current to meet the timing requirement. By the central limit theorem, $N \geq 30$ in this case M = 128 and L will be discussed in a later section.

Dummy cell layout is shown in fig. 4.16.a Observe that there are very minor changes made in the memory cell layout making a best effort to keep the differences in memory cell and dummy memory cell layout as few as possible. Specifically modifying the top layer first and attempting to avoid modifying the lower layers. Both PG devices are disconnected by removing the contact VIA placed in the bit line. The cross coupled inverter pair is decoupled, and one PMOS is diode connected to reroute to the drain of the PD transistor marked as N in fig. 4.13. Once the number of fingers in L is selected, one of the bit lines is used and VIA are placed to make the required number of (L) connections. The dummy cell layout is designed such that when possible the changes can take place at the highest layer (Metal 4). Another bit line is used to connect all the diode connections together. An extra top metal is run to trigger the gate of PD (N cells), coupled with read signal of the bank. The DCBL (dummy bit capacitance) closely matches the memory cell bit line capacitance, the parasitic extraction results will be discussed and compared in following section.

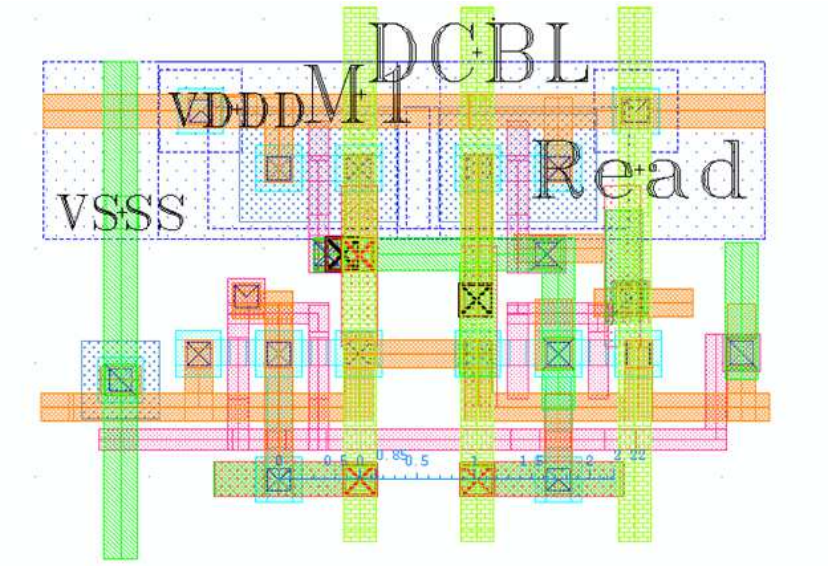


Fig. 4.16.a Dummy Cell Layout

The dummy cell pitch matches with the memory core cell layout. This allows abutment of dummy column and memory cell core. A brief example of memory cell and extreme column abutment with dummy column in a 3x3 memory array for theoretical understanding is shown in fig. 4.16.b.

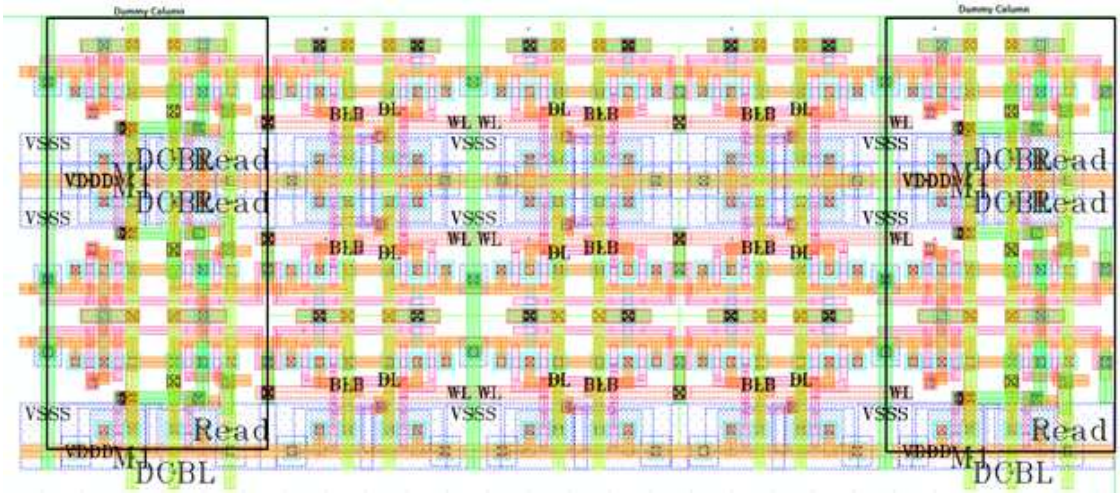


Fig. 4.16.b Dummy Cell abutment layout

4.6 Access Read Design

Now that we have reviewed memory read access time design let us summarize the design steps. The target here is to design targeted yield of 'n' sigma on process while spending exact power to achieve near exact high performance on each bank fabricated on each wafer.

Table 4.5 Memory read current distribution at different process corners at VDD 1V

Process	Icell Mismatch	Icell Sigma	Worst cell read current(3σ)	Coefficient of Variance in %
Slow	25.1 μA	1.49 μA	20.3 μA	5.92
Typical	32.0 μA	1.80 μA	26.6 μA	5.62
Fast	42.9 μA	1.92 μA	36.7 μA	4.47

The statistical distribution memory read current is shown in table 4.5. It can be noted that the coefficient of variance is almost equal throughout the process. Assuming that the pelgrom coefficient is constant throughout the process the access read time is designed for typical wafer and

then verified at process corners. However, there can be slight difference in the pelgrom coefficient across the process and a hypothesis can be made that this difference is due to line edge roughness not being constant throughout the process [51]. The worst case cell current calculation shown in table 4.6.1 is calculated as below.

$$I_{read} = I_{cell} - I_{off} = \left(I_{cell\mu} - n * \sigma_{mismatch}\right) - \left(I_{off\mu} + n * \sigma_{total}\right)$$

$$I_{read} = (32.02\mu - 3 * 1.80\mu) - (5.229n + 3 * 2.6329n) = 26.6\mu A \quad (4.6.1)$$

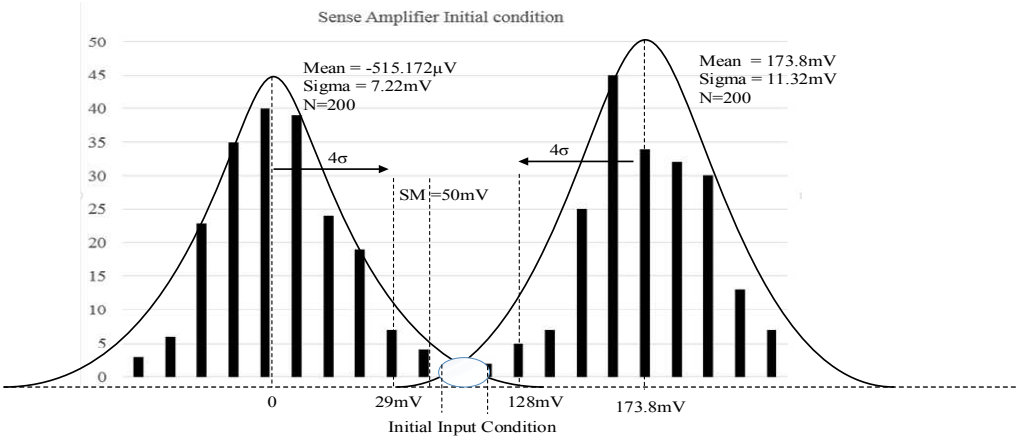


Fig. 4.17 Simulation result of sense amplifier initial condition

Simulation result for sense amplifier initial condition pre sense amplifier triggering is shown in fig. 4.17. As seen earlier in section 3.1.2, ΔV_{BL} is given by $(\Delta V_{BL\mu} - \Delta V_{vosbit}) > \Delta V_{BL} > n * \sigma_{input_{offset}} + OD$. Where OD is the minimum SA initial condition ensuring SA settling in $2.5\tau_{SA}$. From fig.4.17 the designed bit line ΔV_{BL} should be $(173.8mV - 3 * 11.32mV) > \Delta V_{BL} > 4 * 7.22mV + 50mV$.

$\Delta V_{BL} = 4 * 7.22mV + 50mV + 2 * 11.32mV \cong 100mV$. Worst case memory cell read current must develop a bit line voltage greater than $100mV$ prior to triggering the sense amplifier. Bit line discharge time is given by equation (3.1.1). Let's recall the equation and design the bit line discharge time. CBL is extracted from layout using parasitic extraction (QRC). The column layout extraction result is shown in fig. 4.18

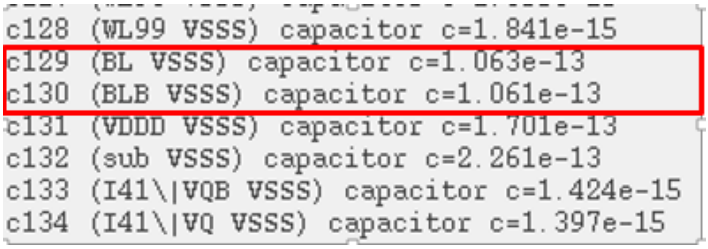


Fig. 4.18 Parasitic extraction of a memory column

where $CBL = CBLB = 106fF$.

Substituting ΔV_{BL} and equation (4.6.1) and (4.6.2) in equation (3.1.1) we get bit line discharge time as $t_{BL_{discharge}} \cong \frac{106f}{26.6uA} * 100mV \cong 398.50ps = 450ps$, Note that there are always gate delay associated with control signal path. Considering the gate delay in advance saves recalculation work. In applying the dummy circuit, we assume the dummy bit line capacitance to be equal to the bit line capacitance or 106fF in this design. Recalling equation (3.2.a)

$t_{dummy} = \left(\frac{C_{dummy}}{I_{reference}}\right) * \Delta V_{trip} = 450ps = \frac{106fF}{I_{reference}} * 490mV$; $I_{reference} = \frac{106fF}{450ps} *$

$490mV \cong 120\mu A$

All the equations in the following discussion are with referenced to fig. 4.13, dummy column. By the central limit theorem greater 30 memory cells are averaged in this replica design. Here 38 dummy memory cells are selected and average from a 128 word line bank.

$Isum = N * I_{cell} = 38 * 32uA = 1.2mA.$ (4.6.2)

Using simulations $Isum = 1.14mA$, a 5% current error is observed due to modified memory cell shown in fig. 4.13. See eq. (4.6.2) to (4.6.6) and fig 4.13.

$I1 = \frac{Isum}{M} = \frac{1.14mA}{128} = 8.9uA;$ (4.6.3)

$I_{reference} = N * I1;$ (4.6.4)

$N = \frac{I_{reference}}{I1} = \frac{120uA}{8.9uA} = 13.4 \cong 14$ (4.6.5)

Current transfer equation $I_{reference} = \left(\frac{N*L}{M}\right) * I_{cells}$, in future node technologies M (number of cells in a column) will increase and to maintain the $I_{reference}$ N and L should be increased proportionally as dictated by bank design and the process. This can further reduces the error increasing accuracy. The number of fingers in the current mirror (L) are adjusted in order to get the desired yield across the process. In this design M=128, N=38 and L=20. Using L=20, dummy column layout is designed by just adding specific node contacts. It always desirable to keep M, N, and L as large a practical while keeping power consumption in mind. Once that is done QRC extractions are run on the dummy column. Parasitic extraction of dummy column is shown in fig. 4.19. Observe only 3.7% error is observed between dummy bit line and memory column bit line capacitance post modification. A high correlated process tracking can be expected between the dummy column design and each memory column.

```
// Library name: new_test
// Cell name: Dummy_Column
// View name: av_extracted
c1 (CBL VSSS) capacitor c=7.649e-14
c2 (L VSSS) capacitor c=1.027e-13
c3 (M1 VSSS) capacitor c=2.298e-13
```

Fig. 4.19 Parasitic extraction of dummy column

A transient memory read access simulation at TT corner is shown in fig. 4.20. A logic ‘1’ is stored at the Q0 node of a memory cell. BLB is populated with the worst case data stored in a column.

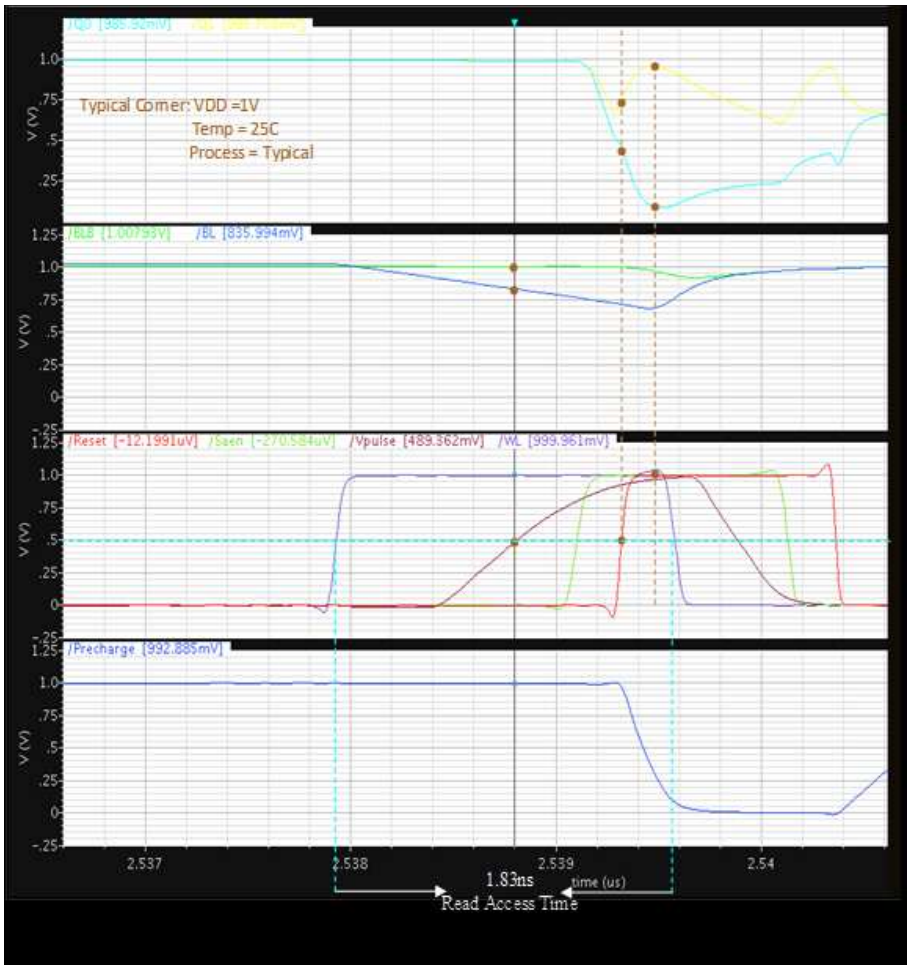


Fig. 4.20 Transient memory read access at TT Corner

A global word line and read signal is initiated to read the memory word. The local word line is then activated which starts discharging the bit line. The read signal and the word line signal are coupled together with logic to trigger the dummy column logic at the same time the local word line is triggered. The dummy bit line starts charging, shown in fig. 4.20 as 'Vpulse'. The comparator is triggered when Cdummy rises to (Vpulse = VDD/2). This in turn triggers all the sense amplifiers to read the word. In the meantime the sense amplifier has sensed the bit line difference and generates the initial voltage at output of the sense amplifier, the initial voltage is then amplified when the reset switch is turned off. It is observed that memory cell is triggered when$\Delta V_{BL} = 165mV$. Memory is designed to handle 3σ mismatch variations and $\Delta V_{vosbit} =$

$11.32mV$. A $\Delta V_{BL} - 3 * \Delta V_{vosbit} = 130mV$. The design Target is 100mV which is nearly achieved accurately. Once the comparator triggers the sense amplifier, it also sends a signal to logic turning off the word line and pre-charging the bit lines to prepare for the next read cycle. This is done to reduce power consumption as well as speed up the read cycle.

A transient memory read access simulation at SS corner and FF corner is shown in fig. 4.20and fig. 4.21 respectively.

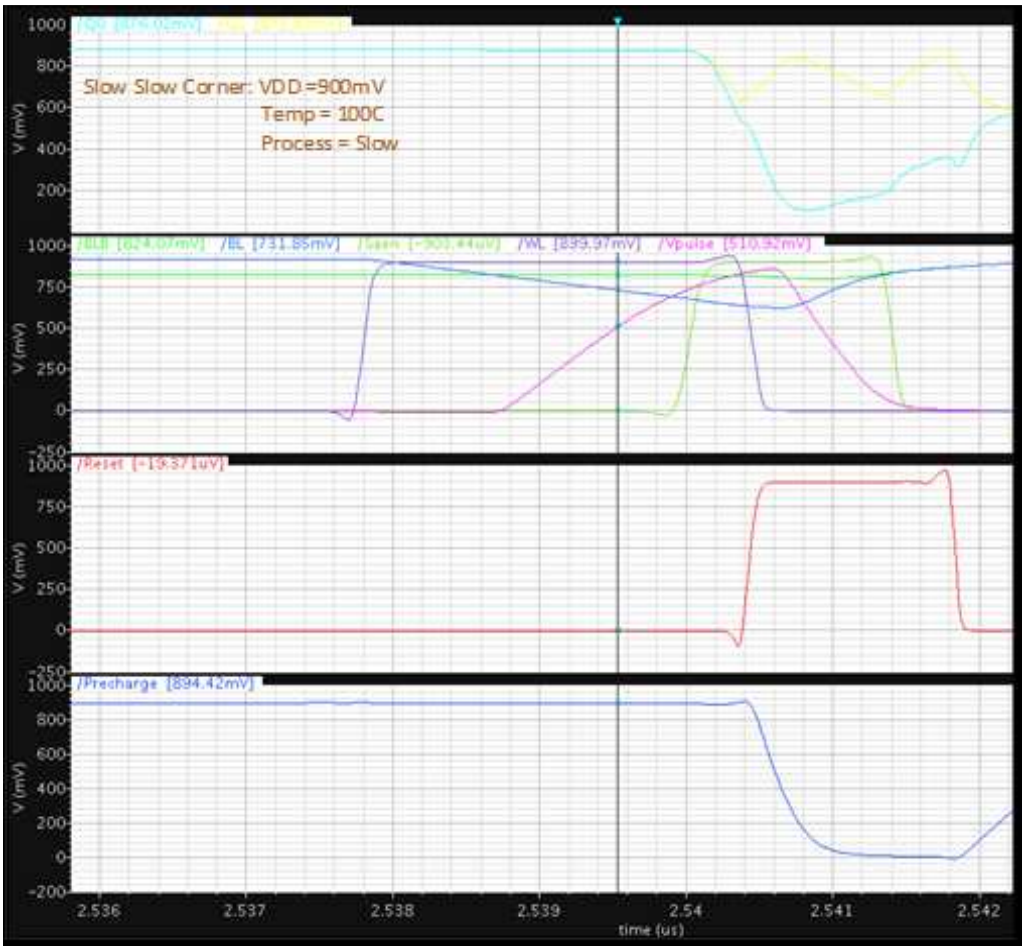


Fig. 4.21 Transient memory read access at SS corner

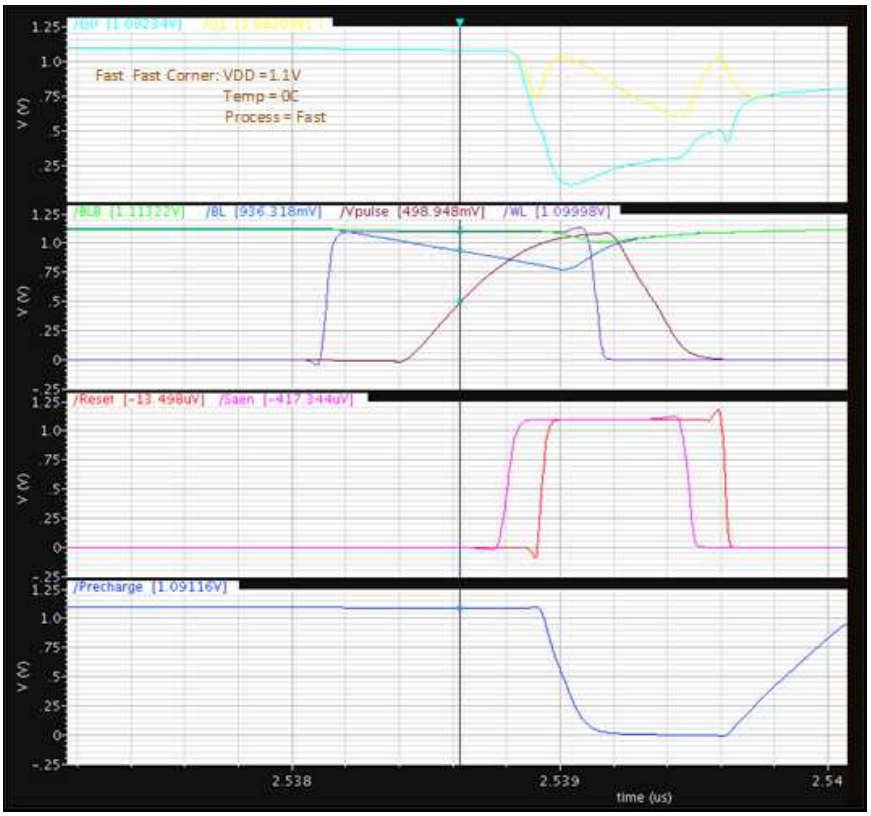


Fig. 4.22 Transient memory read access at FF corner

Table 4.6 Transient memory read summary

Process	Access Time	CFV%	Bit line Difference ΔV	CFV%	Max Bit line Difference ΔV	CFV%
Slow (μ,σ)	(2.463n,39.35p)	1.59	(187.2m,13.17m)	7.035	(304.2m,20.25m)	6.656
Typical (μ,σ)	(1.831n,27.34p)	1.329	(173.8m,11.32m)	6.51	(283.2m,17.2m)	6.07
Fast (μ,σ)	(1.348n,18.7p)	1.38	(156.1m,9.365m)	5.99	(261.1m,13.96m)	5.34

Table 4.6 provides a short summary on transient memory read access simulation at different process. CFV is Coefficient of variance. Max bit line difference is the maximum difference bit line achieves before the word line is switched off. All the bank read operation signals except the bank select are locally generated using dummy circuit. CFV are equal at different process, this indicates that dummy column current tracks memory cell current across the process.

Table 4.7 provides a PVT corner simulation summary. Recalling fig. 2.13, Ioff current increases with increasing temperature, and is observed in table 4.7 at the slow corner BLB droops by 824mV. Observe that if bit line leakage is ignored in memory design memory read will fail at SS corner. This makes SS the worst case corner.

Table 4.7 PVT Corner Simulation

P-V-T Corners	SS - 900mV - 100C	TT - 1V - 25C	FF - 1.1V - 0C
VBL	733.8mV	834.2m	936.1m
VBLB	824mV	1	1.1
ΔV_{BL}	90.23m	173.8m	177.1m
Max ΔV	186.1mV	283.5m	292.5m
Access Time	2.626ns	1.646ns	1ns

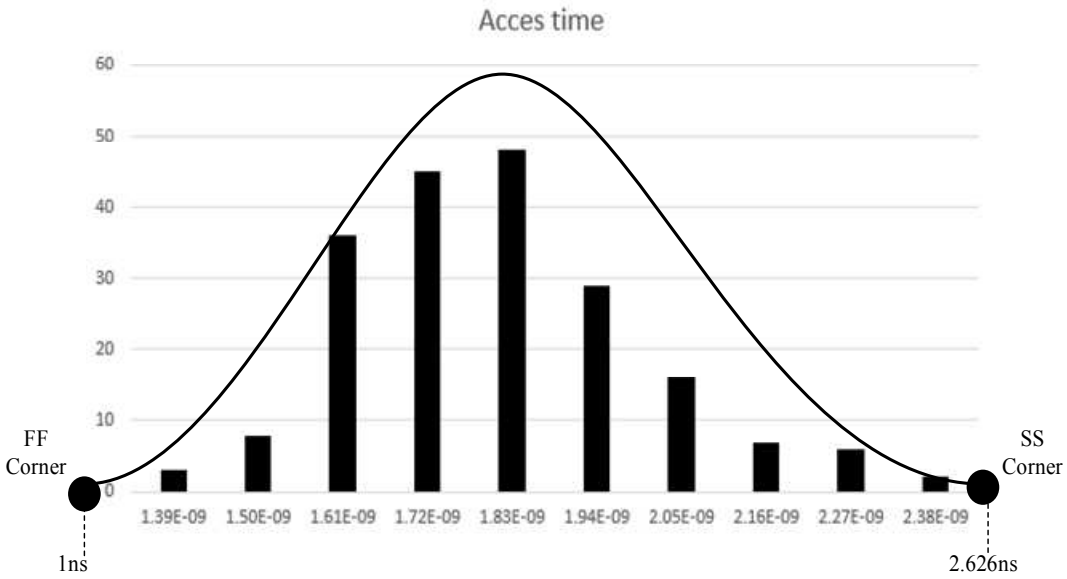


Fig. 4.23 Monte - Carlo analysis on memory access time

A Monte-Carlo analysis simulation on memory access time is shown in fig. 4.23. Fig 4.23 and table (4.5.2) content proves the theory introduced in section 2.5.2 of Monte-Carlo analysis. Monte-Carlo analysis shows the results within the corner simulations. A faster way to analyze the circuit is to do the Monte-Carlo analysis and then confirm the results using corner

simulations. This shows that the design will work across all the process and environmental variations.

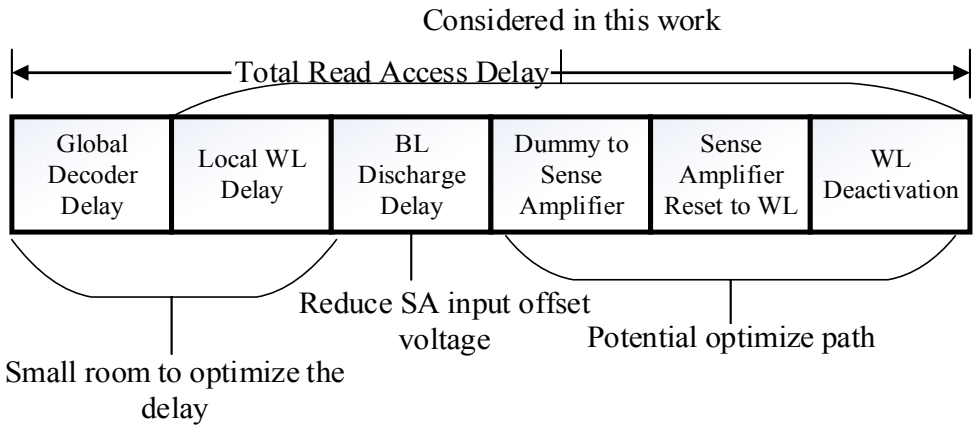


Fig. 4.24 Total read access delay

A total read access time delay is shown in fig. 4.24. Clock initiates global word line and read signal with memory address to perform read operation. There is delay associated with decoder while decoding the address, further the local word line is triggered through a buffer-inverter chain. Once the word line is activated, bit line discharge delay is accounted, the comparator then makes the decision, and dummy circuit triggers the sense amplifier through a buffer-inverter chain. The sense amplifier releases the reset signal and also initiates a control signal to turn off WL. Access time simulation in fig. 4.23 consists of delay from Local WL delay – WL deactivation delay.

In this chapter, a detailed methodology was discussed to achieve a fixed yield by locally adjusting each bank's performance such that a constant yield across the process is achieved. The design is sigma based and by changing only sigma and current transfer ratio the desired results can be achieved. A conceptual output representation of yield, read access time and process variation tolerant architecture is shown in fig. 4.25.

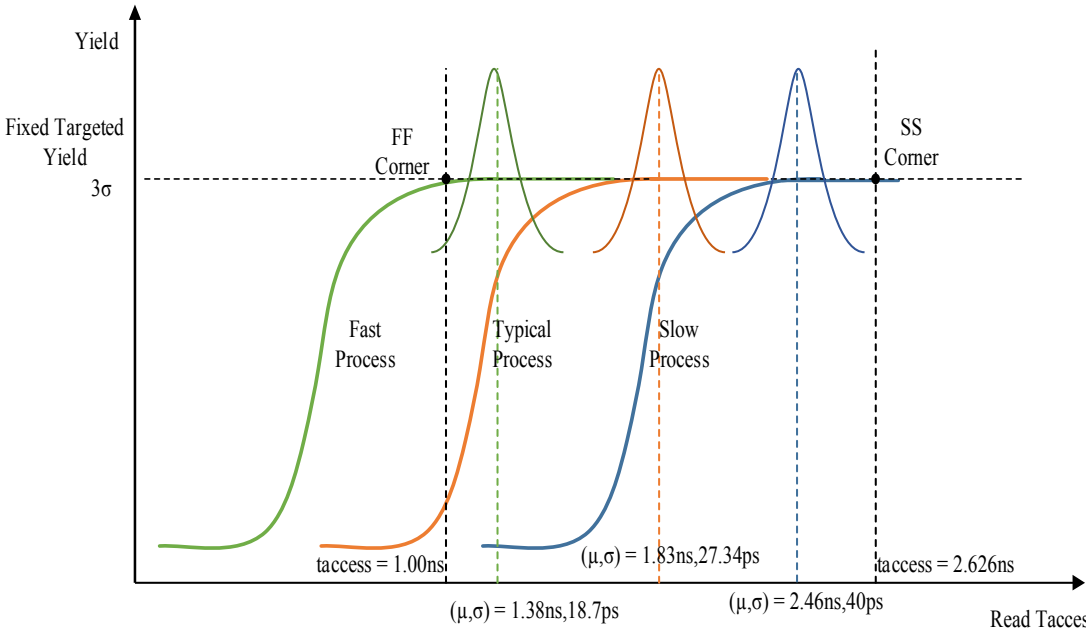


Fig. 4.25 Conceptual output representation of yield, access time and process variation

This shows that every memory bank system gives high performance and maximum yield across all die and across all the wafers; slow-slow (SS), typical-typical (TT) and fast-fast (FF) to be ahead in the memory business.

CHAPTER V

CONCLUSION

5.1 Result Comparison

Proposed work architecture is compared with conventional self-time replica model [45]. In this work, a dummy design area provides a dual use matching performance while with minor modifications generate reference cell current with significantly reduced variations. The locally developed “reference” cell current is used to generate local signals to enhance memory read performance. In memory read techniques [39, 40, 45, 48, 49] 4-16 memory cells are triggered in a single column or multiple dummy columns placed randomly in the bank. The low number of memory cells does not provide a proper estimation of the bank or cell current. Additionally none of [39, 40, 45, 48, 49] track or compensate the process gradient. The accuracy improvement can be calculated by comparing previous efforts with the dummy cell current statistics presented in this work. A coefficient of variance is calculated for a conventional architecture and proposed work in table 5.1

Table 5.1 Comparison between conventional and proposed architecture

Conventional Architecture	Mismatch Typical	Coefficient of Variance (CFV) in %
Icell (μ, σ)	(32.0 μA, 1.80 μA)	5.62129346
Proposed Architecture	Mismatch T	Coefficient of Variance in %
Iavg (μ, σ)	(170 μA, 3.37 μA)	1.987145468
Isum (μ, σ)	(1.14mA, 7.21 μA)	0.635242291

(a)

Conventional Architecture	Mismatch Slow	Coefficient of Variance (CFV) in %
Icell (μ, σ)	(25.1 μA, 1.49 μA)	5.920917617
Proposed Architecture	Mismatch S	Coefficient of Variance (CFV) in %
Iavg (μ, σ)	(119 μA, 2.62 μA)	2.201680672
Isum (μ, σ)	(860 μA, 5.71 μA)	0.664109716

(b)

Conventional Architecture	Mismatch Fast	Coefficient of Variance (CFV) in %
Icell (μ, σ)	(42.9 μA, 1.92 μA)	4.475524476
Proposed Architecture	Mismatch F	Coefficient of Variance (CFV) in %
Iavg (μ, σ)	(229 μA, 4.42 μA)	1.930131004
Isum (μ, σ)	(1.57mA, 9.17 μA)	0.585322272

(c)

Table 5.2 Accuracy improvement in proposed architecture

Wafer	Slow	Typical	Fast
Expected Improvement	88.78%	88.69%	86.92%
Actual Improvement	62.81%	64.64%	56.87%

Using CFV, accuracy improvement is calculated and summarized in Table 5.2. It can be observed that accuracy is improved almost 62% throughout the process. The dummy area sum current has significantly reduced variation compared to the average current and it is expected that the average current follows same variation. However, there are some errors due to an imperfect current mirror

[50]. Due to the early voltage effect, the mirrored drain current is not constant / accurate in saturation region. This results in reduced accuracy in actual read compensation improvement.

5.2 Implementation Cost

Other than the main memory core there are two dummy columns used in the memory architecture and two comparators of the same size and footprint as each sense amplifier. For a 128 bit word memory a 2% area increment is observed. This architecture should trigger more than 30 memory cells in each dummy column to generate an accurate mean measurement and so the power consumption should be closely monitored. A Power budget for the proposed memory architecture is shown in table.5.3.

Table 5.3 Power Budget

Memory	SA Current	Memory Cell Current	Leakage current	Memory Power	Dummy Circuit Power	Power Consumption Increase
128x128	$22uA$	32 uA	13nA	6.913mW	2.62mW	37.90%
256x128	22 uA	32 uA	13nA	13.82mW	2.62mW	18.95%
512x128	22 uA	32 uA	13nA	27.648mW	2.62mW	9.47%

From table 5.3 it is clear that the power consumption is significantly increased in proposed work, however read time performance and yield are improved by 62%. As discussed in earlier sections, at new process nodes memory cell current probability will be skewed at low cell current by the Poisson tail. In this case, more than 40 samples should be sampled to estimate approach the mean cell current value. If process gradient cell current test data is available then skewness of the cell current across the die can be predicted. Knowing skewness of cell current, extreme dummy columns can be designed precisely. The combined column logic shown in fig. 3.3 will improve the sense amplifier firing timing, access time and yield. In proposed work, bank understands the position of the die on a wafer, accommodates the process gradient, takes the decision based on slow gradient

mean cell current. Detailed insight about the gradient tracking ability in proposed work is shown in fig. 5.2.

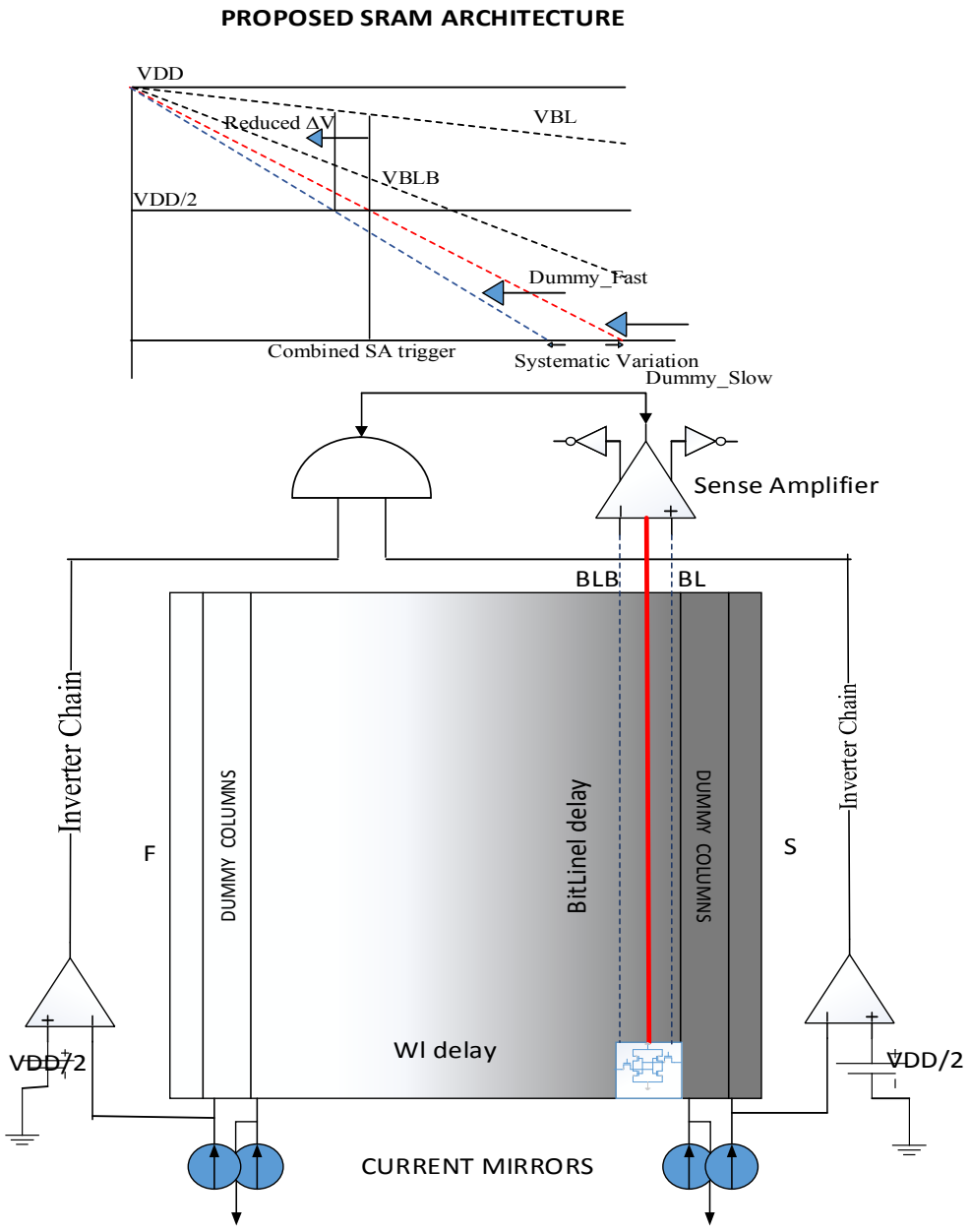


Fig. 5.1 Gradient Tracking Ability

At new process nodes the bank density increases. Systematic variation is a concern with increases density and bank size. Since the bank size is increased we can now have more than 40 samples to

have a better estimate of the bank current. With a skewed/ poisons tail current distribution small current samples won't give better estimation. The bank current achieved from the dummy column using current mirror will be skewed with the process. For an instance let us assume bank has left to right gradient as shown in fig. 5.1. The left column will generate a higher mean current compared to the right column mean current. Note that the magnitude of mean current is differed but the CFV remains nearly same throughout the process, in this case 62% shown in table 5.1. The proposed design forces to take the read decision depending on the slowest mean current as seen in fig. 5.1. Now, when the die moves on the wafer with a gradient right to left, still the decision will be made at the slowest mean current. This is not the case with conventional techniques discussed in [53-57], since there is no combine effect of dummy columns there is a possibility that die could generate the fast mean current and sense amplifier will be fired with reduced ΔVas shown in fig. 5.1. This either increases the failure rate or the timing is still designed for worst condition. In a nutshell, with increased bank size and higher current sampling requirement at new process nodes, proposed architecture seems to be solid.

5.2 Future Improvements

This architecture has concentrated on improving read access failure, yield prediction and performance. This work does not include any improvement in write access time. A dynamic read-write assist approach is presented in [40, 49]. In read assist, a voltage regulator is used to charge the bit line. Bit lines are charged 68-78% of V_{DD} [49]. This decreases V_{DS} across the pass transistor and PD. Decreased V_{DS} reduces Vbump which can trip inverter to flip the data content. The read assist technique allows pass transistor to be big enough in order to support write margin. In write assist technique negative bit-line boosting is used [40, 49]. This increases the Vgs across the pass transistor which discharges the node storing '1' fast to write the opposite data. In this way a decoupled read-write assist technique can be used to achieve read-write margins across the process. Such advanced techniques make 6T design still popular at new process nodes. In the future a

detailed focus on current transfer ratio design, reduction in current mirror error, on chip LDO to reduce hold failure and techniques discussed in [40, 49] together used can lead to more accurate system.

A sincere attempt is made for any initiate who wants to start the research in memory (SRAM) design. This work has addressed all possible challenges an SRAM design offers such as write failure, read failure, hold failure, bit line leakage, sensing window, SA input offset, read current variation, access failure, tracking systematic variation, efficient timing control signals and word yield. This report has considered reducing top priority errors such as access failure, read upset failure, robust new systematic tracking and efficient timing control signals using "A dual column replica bit line delay technique using stochastic current processing for a PVT, low power SRAM". The report has also discussed different ways which could lead to a robust memory design for future process nodes. The overall accuracy is improved by 62% compared to conventional idea and optimal word line activation to SA set time delay is optimized. The proposed architecture should show promising results at future process nodes and big memory banks in terms of power, performance, area and yield.

REFERENCES

1. Available from: http://www.oecd.org/internet/ieconomy/technology-foresight-forum-2014.html.
2. Singh, J., S.P. Mohanty, and D. Pradhan, *Robust SRAM Designs and Analysis*. 2012: Springer Science & Business Media.
3. Pavlov, A. and M. Sachdev, *CMOS SRAM circuit design and parametric test in nano-scaled technologies: process-aware SRAM design and test*. Vol. 40. 2008: Springer Science & Business Media.
4. Pelgrom, M.J.M., H.P. Tuinhout, and M. Vertregt. *Transistor matching in analog CMOS applications*. in *Electron Devices Meeting, 1998. IEDM '98. Technical Digest., International*. 1998.
5. Puttaswamy, K., et al. *System level power-performance trade-offs in embedded systems using voltage and frequency scaling of off-chip buses and memory*. in *System Synthesis, 2002. 15th International Symposium on*. 2002.
6. Fern Nee, T., et al. *SRAM Core Modeling Methodology for Efficient Power Delivery Analysis*. in *SoC Design Conference (ISOCC), 2009 International*. 2009.
7. Navabi, Z., *Digital System Test and Testable Design*. 2011: Springer.
8. Rahma, M.A. and M. Anis, *Nanometer Variation-Tolerant SRAM: Circuits and Statistical Design for Yield*. 2012: Springer Science & Business Media.
9. Abu-Rahma, M.H., et al. *A methodology for statistical estimation of read access yield in SRAMs*. in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*. 2008.
10. Seevinck, E., F.J. List, and J. Lohstroh, *Static-noise margin analysis of MOS SRAM cells.* Solid-State Circuits, IEEE Journal of, 1987. **22**(5): p. 748-754.
11. Jiajing, W., et al. *Statistical modeling for the minimum standby supply voltage of a full SRAM array*. in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*. 2007.
12. Hanwool, J., et al., *One-Sided Static Noise Margin and Gaussian-Tail-Fitting Method for SRAM.* Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 2014. **22**(6): p. 1262-1269.
13. Calhoun, B.H. and A.P. Chandrakasan, *A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation.* Solid-State Circuits, IEEE Journal of, 2007. **42**(3): p. 680-688.
14. Calhoun, B.H. and A.P. Chandrakasan, *Static noise margin variation for sub-threshold SRAM in 65-nm CMOS.* Solid-State Circuits, IEEE Journal of, 2006. **41**(7): p. 1673-1679.
15. Nakata, S., et al. *Increasing static noise margin of single-bit-line SRAM by lowering bit-line voltage during reading*. in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*. 2011.
16. Grossar, E., et al., *Read Stability and Write-Ability Analysis of SRAM Cells for Nanometer Technologies.* Solid-State Circuits, IEEE Journal of, 2006. **41**(11): p. 2577-2588.
17. Arandilla, C.D.C., A.B. Alvarez, and C.R.K. Roque. *Static Noise Margin of 6T SRAM Cell in 90-nm CMOS*. in *Computer Modelling and Simulation (UKSim), 2011 UkSim 13th International Conference on*. 2011.

18. Takeda, K., et al., *A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications.* Solid-State Circuits, IEEE Journal of, 2006. **41**(1): p. 113-121.
19. Yabuuchi, M., et al. *A 45nm 0.6V cross-point 8T SRAM with negative biased read/write assist*. in *VLSI Circuits, 2009 Symposium on*. 2009.
20. Zhiyu, L. and V. Kursun, *Characterization of a Novel Nine-Transistor SRAM Cell.* Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 2008. **16**(4): p. 488-492.
21. Ik Joon, C., et al., *A 32 kb 10T Sub-Threshold SRAM Array With Bit-Interleaving and Differential Read Scheme in 90 nm CMOS.* Solid-State Circuits, IEEE Journal of, 2009. **44**(2): p. 650-658.
22. Karl, E., et al., *A 4.6 GHz 162 Mb SRAM Design in 22 nm Tri-Gate CMOS Technology With Integrated Read and Write Assist Circuitry*. Solid-State Circuits, IEEE Journal of, 2013. **48**(1): p. 150-158.
23. Satish, D., et al. *A 22nm IA multi-CPU and GPU System-on-Chip*. in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*. 2012.
24. Freund, R.J., W.J. Wilson, and D. Mohr, *Statistical Methods, Students Solutions Manual (e-only)*. 2010: Academic Press.
25. Kuhn, K.J. *CMOS transistor scaling past 32nm and implications on variation*. in *Advanced Semiconductor Manufacturing Conference (ASMC), 2010 IEEE/SEMI*. 2010.
26. Mizuno, T., J. Okumtura, and A. Toriumi, *Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's.* Electron Devices, IEEE Transactions on, 1994. **41**(11): p. 2216-2221.
27. Takeuchi, K., T. Tatsumi, and A. Furukawa. *Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuation*. in *Electron Devices Meeting, 1997. IEDM'97. Technical Digest., International*. 1997. IEEE.
28. Croon, J.A., W.M. Sansen, and H.E. Maes, *Matching properties of deep sub-micron MOS transistors*. 2005: Springer.
29. Tze-chiang, C. *Where CMOS is going: trendy hype vs. real technology*. in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*. 2006.
30. Cheng, Y. and C. Hu, *MOSFET modeling & BSIM3 user's guide*. 1999: Springer Science & Business Media.
31. Taur, Y. and T.H. Ning, *Fundamentals of modern VLSI devices*. 2009: Cambridge university press.
32. Liao, R., *A low power digital baseband core for wireless micro-neural-interface using CMOS sub/near-threshold circuit*. 2013, Oklahoma State University.
33. Ghibaudo, G., *On the theory of carrier number fluctuations in MOS devices.* Solid-State Electronics, 1989. **32**(7): p. 563-565.
34. Sonoda, K., et al., *Discrete Dopant Effects on Statistical Variation of Random Telegraph Signal Magnitude.* Electron Devices, IEEE Transactions on, 2007. **54**(8): p. 1918-1925.
35. De, K., *Design and implementation of a low power T-Gate cell library and comparison with its CMOS equivalent*. 2014, OKLAHOMA STATE UNIVERSITY.
36. *IBM 7RF PDK*
37. Nakanishi, J., et al. *Analysis technique for systematic variation over whole shot and wafer at 45 nm process node*. in *ASIC, 2009. ASICON '09. IEEE 8th International Conference on*. 2009.
38. Kanj, R., R. Joshi, and S. Nassif. *SRAM Yield Sensitivity to Supply Voltage Fluctuations and Its Implications on Vmin*. in *Integrated Circuit Design and Technology, 2007. ICICDT '07. IEEE International Conference on*. 2007.
39. Wolpert, D. and P. Ampadu, *Managing temperature effects in nanoscale adaptive systems*. 2011: Springer Science & Business Media.

40. Pilo, H., et al., *A 64 Mb SRAM in 32 nm High-k Metal-Gate SOI Technology With 0.7 V Operation Enabled by Stability, Write-Ability and Read-Ability Enhancements.* Solid-State Circuits, IEEE Journal of, 2012. **47**(1): p. 97-106.
41. Heald, R. and P. Wang. *Variability in sub-100nm SRAM designs*. in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*. 2004.
42. Lee, T.H., *The design of CMOS radio-frequency integrated circuits*. 2004: Cambridge university press.
43. Tachibana, S., et al., *A 2.6-ns wave-pipelined CMOS SRAM with dual-sensing-latch circuits.* Solid-State Circuits, IEEE Journal of, 1995. **30**(4): p. 487-490.
44. Schuster, S.E., et al., *A 15-ns CMOS 64K RAM.* Solid-State Circuits, IEEE Journal of, 1986. **21**(5): p. 704-712.
45. Amrutur, B.S. and M.A. Horowitz, *A replica technique for wordline and sense control in low-power SRAM's.* Solid-State Circuits, IEEE Journal of, 1998. **33**(8): p. 1208-1219.
46. Sharma, A., *Introduction to Advanced Semiconductor Memories*, in *Advanced Semiconductor Memories:Architectures, Designs, and Applications*. 2003, Wiley-IEEE Press. p. 1-18.
47. Mukhopadhyay, S., H. Mahmoodi, and K. Roy. *Statistical design and optimization of SRAM cell for yield enhancement*. in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*. 2004.
48. Li, Y., et al., *An area-efficient dual replica-bitline delay technique for process-variation-tolerant low voltage SRAM sense amplifier timing.* IEICE Electronics Express, 2014. **11**(3): p. 20130992-20130992.
49. Pilo, H., et al., *An SRAM Design in 65-nm Technology Node Featuring Read and Write-Assist Circuits to Expand Operating Voltage.* Solid-State Circuits, IEEE Journal of, 2007. **42**(4): p. 813-819.
50. Gray, P.R., et al., *Analysis and design of analog integrated circuits*. 2001: Wiley.
51. *https://www.youtube.com/watch?v=6LcTrp6SB3o*.

VITA

Aniket Suhas Kulkarni

Candidate for the Degree of

Master of Science

Thesis: A DUAL COLUMN, REPLICA BITLINE DELAY TECHNIQUE USING STOCHASTIC CURRENT PROCESSING FOR A PROCESS VARIATION TOLERANT, LOW POWER SRAM

Major Field: Electrical and Computer Engineering

Biographical:

Education:

Completed the requirements for the Master of Science in Electrical and Computer Engineering at Oklahoma State University, Stillwater, Oklahoma in December, 2015.

Completed the requirements for the Bachelor of Engineering in Electronics and Telecommunications at University of Mumbai, Navi Mumbai, Maharashtra /INDIA in May 2011.

Experience:
Assistant Manager at Reliance Communication/Alcatel lucent July 2011 – July 2013

Professional Memberships: