A Novel 10T, 256 Cells Per Bitline 1KB Stacked SRAM Design

A DISSERTATION

Submitted in partial fulfillment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

(With Specialization in Semiconductor Devices & VLSI Technology)

GAURAV GONTIYA

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE -247 667 (INDIA)

JUNE, 2008

PUBLICATION

[1] Gaurav Gontiya, S. Dasgupta, "1.25GHz, 10T Stacked SRAM Cell Design at 1.1V and 90nm Technology," *TENCON International IEEE Conference 2008*, August 2008, (Under review).

ABSTRACT

SRAM performance and stability play important role in designing high performance microprocessors where speed is the main criteria. On the other hand wireless sensors and bio-medical equipment have power saving as top priority. Along with these there are applications where memory access is required at fast rate for certain interval of time and it remains idle for rest of the period like laptops. In today's battery operated electronic devices saving power through voltage scaling of integrated circuits is best solution so increase battery life. But soon problems of reliability and failure of memory came. Hence main emphasis of all designers is to reduce power and optimize performance and reliability. Sub-threshold region of operation provides new opportunity so save power but at the cost of reduced speeds.

During memory design memory cell forms the core of complete architecture. Optimizing the power in one cell ensures large saving of power as these units may be repeated for 2^{10} to 2^{20} in just one processor. Peripherals and routing of word lines play important role in deciding the access time of memory. Hence it's important to design highly efficient peripherals. But it should be kept in mind that these peripheral circuits are continuously used, so there power consumed should not go beyond the limits.

This thesis is aimed at development of memory-speed optimized cells and decoders, SRAM cell where leakage prevention and robust design under process variation were top priority. Moreover it was found that 10T Stacked cell can perform successfully in voltage range of 400mV to 1.1V making it the choice for application where dynamic voltage scaling DVS is possible. This was made possible at the cost of increased cell area. In nanometer regime this increase in area is a appreciable tradeoff with process tolerance and less power consuming cell design. Complete layout of prototype 1KB SRAM design was made was tested with design rule check. Parasitic were extracted through layout versus schematics. Through circuit level techniques it was ensured that as high as 256 bits can be connected at voltage as low as 400mV.Write back circuitry was implemented to ensure proper writability. Simulations were performed at 90nm technology node.

LIST OF FIGURES

CHAPTER 2

2.1 A 2^{m+n} bit memory chip organized as an array of 2^n rows X 2^m columns 5

2.2 Two-level decoder for 6-bit address 6

2.3 Sense Amplifier 7

2.4 Read cycle using a precharge circuit 8

2.5 SRAM architecture for 2^{m+n} bit organized as an array of 2^n rows X 2^m columns 9

2.6 Stacked nMOS (lower three) and single nMOS(upper three) leakage at proc. corners 12

CHAPTER 3

3.1 Standard 6T SRAM cell structure 13

3.2 An illustration of SRAM inverter VTC deterioration under low-V_{DD} 14

3.3 Flip-flop representation of SRAM cell with inserted static noise, V_n 16

CHAPTER 4

4.1 10T SRAM cell design with M9 and M10 MOSFET used as buffer 18

4.2 (a) Read Bit line reading 1 (b) Read Bit line reading 19

4.3 Open Bit line architecture for single ended sensing 20

4.4 Six-transistor SRAM cell at the onset of write operation (writing '0'→'1') 21

4.5 Shows the Butterfly curve at supply voltage of 1.1V 22

4.6 Shows the SNM comparison of 6T and 10T cells by varying the supply voltage 23

4.7 (a) RNM comparison (b)Variation of normalized RNM with supply voltage 24

4.8 8 Maximum access frequency attained for different supply voltage 25

4.9 Shows WNM of 10T cell and (b) for 6T SRAM at 300mV 26

4.10 Layout of TG with metal 2 and Conventional 6T SRAM 28

4.11 Layout of 10T stacked SRAM 29

4.12 (a)Layout of Write Driver (b) Layout of Row and Column Decoder 30

4.13 Layout of 10T SRAM with Peripheral Circuits 32

4.14 Subthreshold Leakage paths in 6T cell = 3 paths and in 10T = 2 paths 33

4.15 Shows the leakage power during standby mode in single bit of SRAM 34

4.16 Variation of Idle read/write power with width of Driver 35

4.17 Variation of read/write power with access frequency 36
4.18 Variation of read/write power with supply voltag 36

CHAPTER 5

5.1 Write opteration in Sub-V_t region 37
5.2 Shows the Pseudo write in unselected column 38
5.3 Write back scheme for preserving row data during write operation 39
5.4 Normalized current under typical and slow case of nMOS transistor 40
5.5 I_{ON}/I_{OFF} ratio and its degradation with process variation 41
5.6 Worst case to determine the maximum number of transistor in the column 42
5.7 Shows controlled power supply, Foot Driver and Boosted Word line 43

CHAPTER 6

6.1 Schematic View for idle power, write operation and write power simulation ckt 45
6.2 Write operation at 1.1V with writing time of 0.2 ns 47
6.3 Write operating on at 300mV with writing time of 40 ns 48
6.4 Schematic for read time and read power simulation ckt 48
6.5 Read time and read power simulation waveforms V_{DD}=1.1 49
6.6 Read waveforms V_{DD}= 600 mV 49

CONTENTS

Topic Title	Page
CANDIDATE'S DECLARATION	i
CERTIFICATE	i
ACKNOWLEDGEMENT	ii
PUBLICATION	iii
CHAPTER 1: INTRODUCTION	1
1.1 Literature Review	2
1.1.1 The novel SRAM Cells	2
1.1.2 Dynamic biasing techniques	2
1.1.3 V_{DD}-gating Techniques	3
1.2 Motivation	4
1.3 Literature Survey	4
CHAPTER 2: SRAM ARCHITECTURE AND PERIPHERALS	5
2.1 Row/Column Decoders	6
2.2 Read Sense Amplifier	7
2.3 Precharge Circuit	8
2.4 SRAM Architecture	9
2.5 Power Dissipation Analysis Using Stack Effect	10
CHAPTER 3: DATA RETENTION VOLTAGE	13
3.1 DRV Theoretical Lower Bound	13
3.2 Low-Voltage SRAM Standby Stability Analysis	16

Topic Title		Page
CHAPTER 4: PROPOSED TECHNOLOGY:10T SRAM		17
4.1	Cell Structure	17
4.2	Read Operation	19
4.3	Writing Operation	21
4.4	Static Noise Margin	22
4.5	Read Noise Margin	24
4.6	Voltage Scaling	24
4.7	Sensitivity to Process Variations and Mismatch	25
4.8	Sizing and Layout	27
4.9	Current Leakage	33
CHAPTER 5: SUBTHRESHOLD SRAM DESIGN		37
5.1	Writeback Scheme for Row Data Preservation	38
5.2	Floating Bitline and Back Gate Biasing of pMOS	39
5.3	Data-Dependent Bitline leakage	40
5.4	Write margin Improvement	42
CHAPTER 6: RESULTS OF SIMULATIONS		45
6.1	Implementation Results and Schematics	45
6.2	Conclusion	51
6.3	Scope for Future Work	52
REFERENCES		53

CHAPTER 1

INTRODUCTION

Modern digital systems require the capability of storing and retrieving large amounts of information at high speeds [1] which is essential to all digital systems. According to ITRS [2] by the year 2014, 94% of chip area is going to be occupied by memory. Design of both embedded and stand-alone SRAMs requires the estimation and reduction of stand-by power consumption, design of high-speed peripheral circuits, and achieving reliability in deep submicron low-voltage operation. The ever-increasing demand for larger data storage capacity has driven the fabrication technology and memory development toward more compact design rules and, consequently, toward higher data storage densities. Thus, the maximum realizable data storage capacity of single-chip semiconductor memory arrays approximately doubles every two years. Appliances such as sensor nodes and mobile phones use memory in sub-threshold region of operation. This enables maximum power saving [3]. However design of cell [4] - [10] and its peripheral [11] becomes a challenging problem.

Techniques for Low Power Operation of Memory Structure are given by,

- Leakage current reduction (in active and standby mode) by utilizing various device/circuit/Architecture topologies.
- Operating voltage reduction.
- DC current reduction by using new pulse operation techniques for word-lines, periphery circuits and sense amplifiers.
- Read noise margin improvement at sub 1V range by using separate buffer line and Read bit line.
- Write noise margin improvement by using power line floating write operation.
- Eliminating data dependent bit line leakage through voltage boosting.
- Write scheme for row data prevention in unselected columns during write.

1.1 Literature Review

A large variety of circuit design techniques have been proposed to reduce the leakage power of SRAM cells and the memory peripheral circuits (decoding circuitry, I/O, etc). Previous work showed that leakage of the peripheral circuits can be effectively suppressed by turning off the leakage paths with switched source. The focus was on the leakage control of SRAM core cell. The existing SRAM cell leakage reduction techniques include novel SRAM cell design [9,12], dynamic-biasing [13], and V_{DD}-gating [14-16]. The following sections provide a detailed review of the existing techniques.

1.1.1 The novel SRAM Cells

As the supply voltage (V_{DD}) scales down in each new technology generation, in recent years several new SRAM cell designs were proposed with a reduced leakage power. A 10-T SRAM cell in CMOS technology improves the read margin by buffering the stored data during a read access, and enhances the write margin with a floating V_{DD} during write operation [9]. The improved operation margins allow this cell to operate at a V_{DD} lower than 400mV. Memory operations at such a low voltage effectively reduce both the active and standby power. In another work, a 4-T FinFET-based SRAM cell used back-gated feedback design to boost the static noise margin (SNM) and reduce cell leakage. [12]

1.1.2 Dynamic biasing techniques

The dynamic-biasing techniques use dynamic control on transistor gate-source and substrate-source bias to enhance the driving strength of active operations and create low leakage paths during standby period [17]. For example, the driving source-line (DSL) scheme connects the source line of the cross-coupled inverters in an SRAM cell to a negative voltage VBB during read cycle, and leaves the source line floating during write cycle. This bias configuration improves the speed of both read and writes operations in SRAM cell. Therefore, high threshold (V_{th}) transistors can be used to reduce leakage, without compromising the active performance [18]. Another similar technique is the negative word-line driving (NWD) scheme. NWD uses low V_{th} access transistors with negative cut-off gate voltage and high V_{th} cross-coupled inverter pair with boosted gate voltage. The result is an improved access time and a reduced standby leakage [13]. The dynamic leakage cut-off (DLC) scheme applies reverse-biased PMOS and NMOS substrate voltages on non-selected SRAM cells [13]. At the current technology nodes (130nm and 90nm), the above dynamic-biasing schemes typically achieve 5-7X leakage power

reduction. This power saving becomes less as the technology scales, because the worsening short-channel effects cause the reverse body bias effect on leakage suppression to diminish [19]. In order to design for a higher (>30X) and sustainable leakage power reduction, an SRAM designer needs to integrate multiple low-power design techniques, rather than using dynamic-biasing only.

1.1.3 V_{DD}-gating Techniques

The V_{DD}-gating techniques either gate-off the supply voltage of idle memory sections, or put less frequently used sections into a low-voltage standby mode. There are three types of leakage mechanisms in an SRAM cell: sub-threshold leakage, gate leakage and junction leakage. A lower V_{DD} reduces all of these leakages effectively. The reduction ratio in leakage power is even higher because both the supply voltage and leakage current are reduced.

An example of V_{DD}-gating is the Cache Decay technique, which gates off unused cache sections, and uses cache activity analysis to balance the leakage energy saving against the performance loss caused by extra cache misses. With adaptive timing policies in cache line gating, Cache Decay achieves 70% leakage power reduction at a performance penalty. To further reduce leakage power for caches with large utilization ratio, the Drowsy Caches approach was proposed to allocate inactive cache lines to a low-power mode, where a low standby V_{DD} is used to reduce leakage. The Drowsy Caches design assumes that the standby V_{DD} is higher than the voltage level required for SRAM data-retention. Therefore the cache data are preserved during the drowsy standby mode. A leakage power reduction is higher than 70% in a drowsy data cache [20].

In recent years as the need of leakage reduction in high-utilization memory structures increases, there have been many research activities on low-voltage SRAM standby techniques. Most of the reported circuit techniques in this field focus on the design of sleep control circuits. For example, an array of dynamically-controlled sleep transistors was used to provide a finely programmable standby V_{DD} [14]. In another design, a self decay circuit generates a periodical sleep pulse with an adaptive pulse period, which puts the SRAM array into a sleep mode more frequently at high leakage conditions (fast process, high temperature) and vice versa. The result is an optimized tradeoff between leakage power reduction and dynamic power overhead [15]. A recent work proposed an actively clamped sleep transistor design, which is capable of adaptively adjusting the level of standby V_{DD} based on the magnitude of leakage current. With this design

the cache standby power is minimized under all conditions during the lifetime of a processor [16].

1.2 Motivation

Leakage current under OFF state of transistor, is a concern for designers. It leads to power wastage during standby mode. It is also observed that RNM for conventional SRAM is very less for sub 1-V range. WNM also decreases under process variation. Reduced I_{ON}/I_{OFF} in subthreshold region causes number of cells connected per bitline to decrease. Poor writability and over-writing of unaccesed cells are also observed. This motivated us to stack driver transistors in SRAM. Stacking of transistor provides reduction in leakages. This approach does not need any architectural changes in SRAM, and there is no need to apply extra control signals to control the leakages. Further, stacking also results in reduction of gate leakages. Owing to stacking some delay penalty is introduced in read/write operations. However this problem can be solved by adding read buffer which effectively isolates data storing nodes during read operation. Also, WNM is more process tolerant with stacking. Use of Foot-Inverter decreases I_{OFF} current and I_{ON}/I_{OFF} ratio increases. Use of Write Back circuitry and floating supply voltage during writing improves writability.

1.3 Organization of Thesis

This thesis is divided into six chapters. Chapter-1 briefly describes the previously used techniques for leakage reduction and brings out the motivation behind the present work. Chapter-2 describes the SRAM architecture and circuitry associated with it. Functionality of Row/Column Decoder, Sense Amplifier, Precharge Circuit are also discussed here. In Chapter-3 minimum supply voltage for retaining the data in the cell and various factors that affect it are discussed. In Chapter-4, a 10T SRAM is proposed and its SNM, RNM, process tolerance are determined. Layout of the cell made on L-edit is presented here. Further the design rule check applied to layouts. Read/Write power and access frequency found for the proposed design are determined. In Chapter-5, further modifications on circuit level are introduced for sub-threshold operation. In Chapter-6 simulation results are presented, followed by conclusion and scope for future work.

SRAM ARCHITECTURE AND PERIPHERALS

The preferred organization for most large memories is shown in Figure 2.1. [21] This organization is random-access architecture. The storage array, or core, is made up of simple cell circuits arranged to share connections in horizontal rows and vertical columns. The horizontal lines, which are driven only from outside the storage array, are called wordlines, while the vertical lines, along which data flow into and out of cells, are called bitlines. A cell is accessed for reading or writing by selecting its row and column. Each cell can store 0 or 1. Memories may simultaneously select 4, 8, 16, 32, or 64 columns in one row depending on the application. The row and column (or groups of columns) to be selected are determined by decoding binary address information.

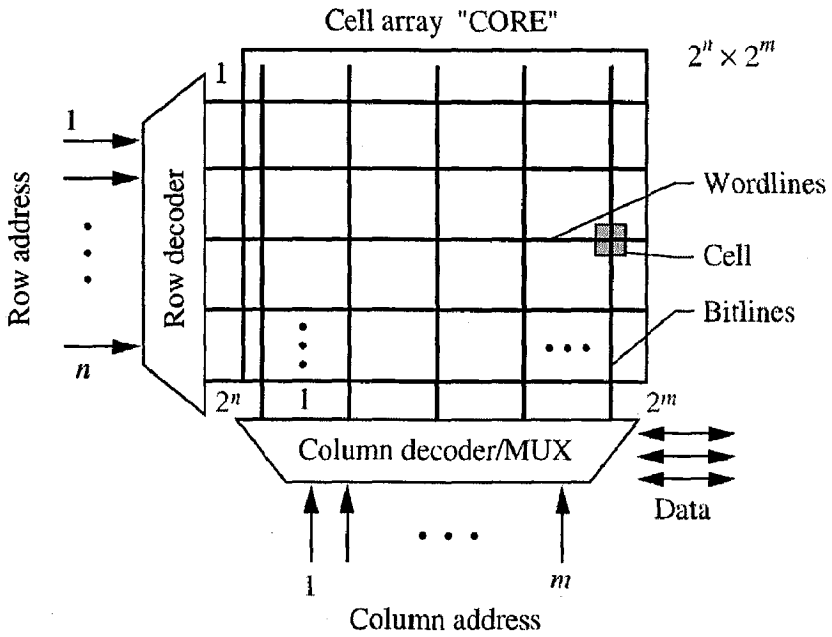


Figure 2.1: A 2^{m+n} bit memory chip organized as an array of 2^n rows X 2^m columns. [21]

For example, an n-bit decoder for row selection, as shown in Figure 2.1has 2^n output lines, a different one of which is enabled for each different n-bit input code. The column decoder takes m inputs and produces 2^m bitline access signals, of which 1, 4, 8,16, 32, or 64 may be enabled at one time. The bit selection is done using a multiplexer circuit to direct the corresponding cell

outputs to data registers. In total, 2^n X 2^m cells are stored in the core array. Along with the core peripheral plays important role in SRAM performance and lowering the power. [22]

2.1 Row/Column Decoders

The row and column decoders identified in Figure 2.2 are essential elements in all random-access memories. Access time and power consumption of memories may be largely determined by decoder design. Similar designs are used in read-only and read-write applications. Row decoders take an n-bit address and produce 2^n outputs, one of which is activated. An n-bit decoder requires 2^n logic gates, each with n inputs. For example, with n=6, we need 64 NAND gates, with 6 inputs, driving 64 inverters to implement the decoder. Gates with more than 3 or 4 inputs create large series resistances and long delays. Rather than using n-input gates, it is preferable to use a cascade of gates. Typically two stages are used: a predecoder stage and a final decode stage. The predecoder stage generates intermediate signals that are used by multiple gates in the final decode stage. The main advantage of two-level decoding is that a large number of intermediate signals can be generated by the predecoder stage and then reused by the final decoding stage. The result is a reduction in the number of inputs for each gate.

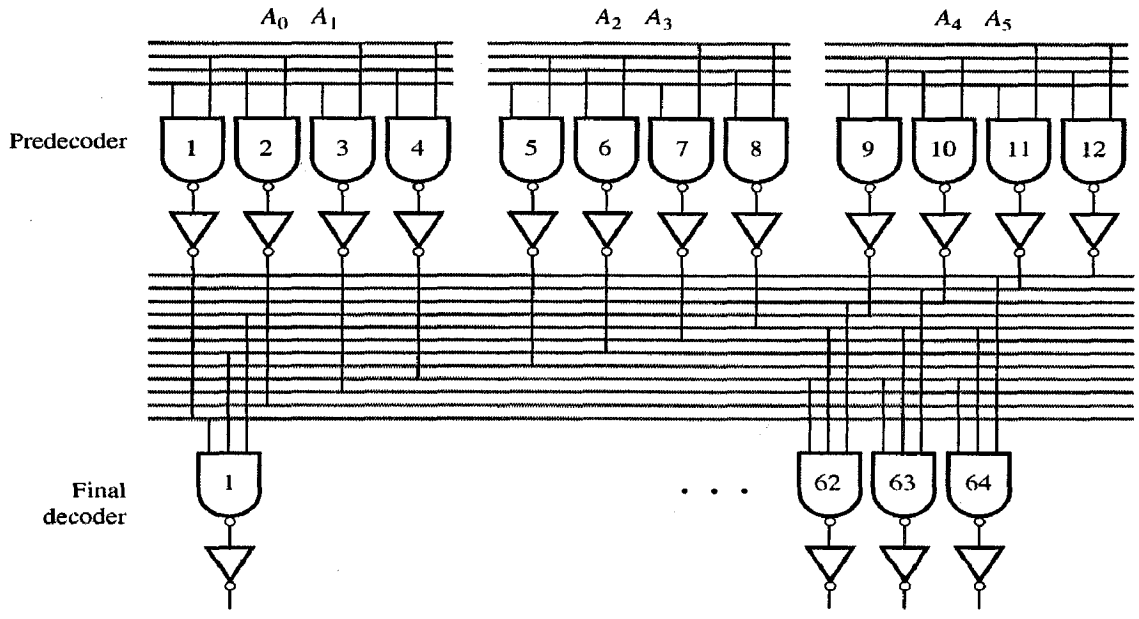


Figure 2.2: Two-level decoder for 6-bit address. [23]

Consider an example for n=6 address bits. In the predecoder, a total of 12 intermediate signals are generated from the address bits and their complements. The signals that are generated in the predecoder are as follows: $A_0A_1, A_0\overline{A}_1, \overline{A}_0A_1, \overline{A}_0\overline{A}_1, A_2A_3, A_2\overline{A}_3$,etc. These signals may now be used by the final decoding stage to generate the 64 required outputs using 3 input NAND/inverter combinations. This corresponds to the configuration shown in Figure 1.7. Each predecoder output drives 16 NAND gates. The delay through 2 input NAND-inverter-3 input NAND- inverter stages can be minimized by sizing the gates. It is important to minimize the delay through the decoder as it may constitute up to 40% of the clock cycle.

2.2 Read Sense Amplifier

A Sense Amplifier is an essential circuit in designing memory chips. Due to large arrays of SRAM cells, the resulting signal, in the event of a "read" operation, has a much lower voltage swing. To compensate for that swing a sense amplifier is used to amplify voltage coming off BL and BL_BAR. The voltage coming out of the sense amplifier typically has a full swing of few hundreds millivolts. Sense amplifier also helps reduce the delay times and power dissipation in the overall SRAM chip.

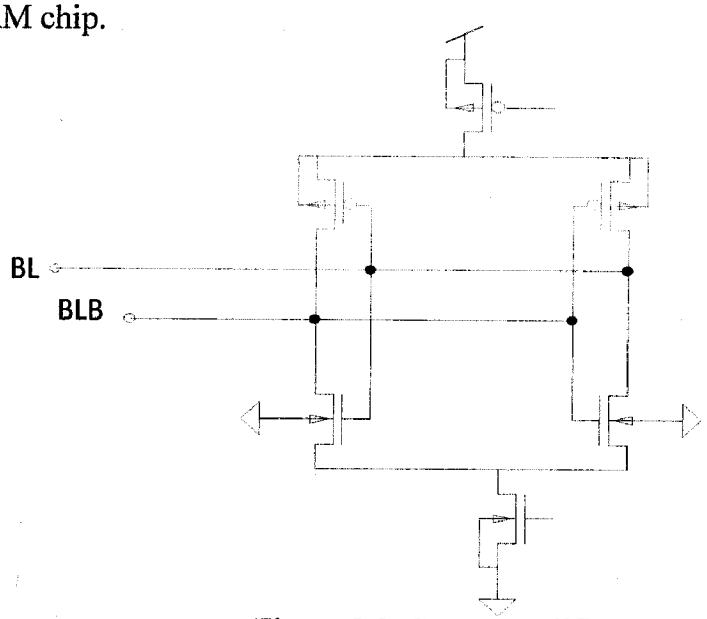


Figure 2.3: Sense amplifier

There are many versions of sense amplifiers used in memory chips. The current mirror type sense amplifier is popular and has a good common mode rejection ratio. But the relatively large area required for large transconductance of input transistors and the large power consumption can become limiting factors for this circuit. When SRAM operates in read mode, there are two

key points to determine whether the operation is correct or not. First, memory cells must provide enough voltage differences on bit line pairs. Second, when sense amplifier is enabled, it must have the ability to amplify the small differential signals at required time. Considering the process deviations, the impact on device parameters cannot be avoided. For memory cells, the variations of threshold voltage and transistor size will cause the changing of voltage difference on bit line pair when memory cell is opened. For sense amplifier, the variation of transistor size will cause the input mismatch and affect the valid sensing region. Based on these impacts from process deviations, analyzing the input signals of sense amplifier is required.

2.3 Precharge Circuit

Safe read and write operations require a modification of the memory array and timing sequence, based on a precharge circuit. The usual voltage of precharge is $V_{DD}/2$. Before reading or writing to the memory, the bit lines are tied to $V_{DD}/2$ using appropriate pass gates. When reading, the BL and BL_BAR diverge from $V_{DD}/2$ (Figure 2.4) and reach the "1" and "0" levels after a short time. As the static RAM cells are based on active devices (Two ring inverters), the SRAM memories usually provide the fastest read and write access times.

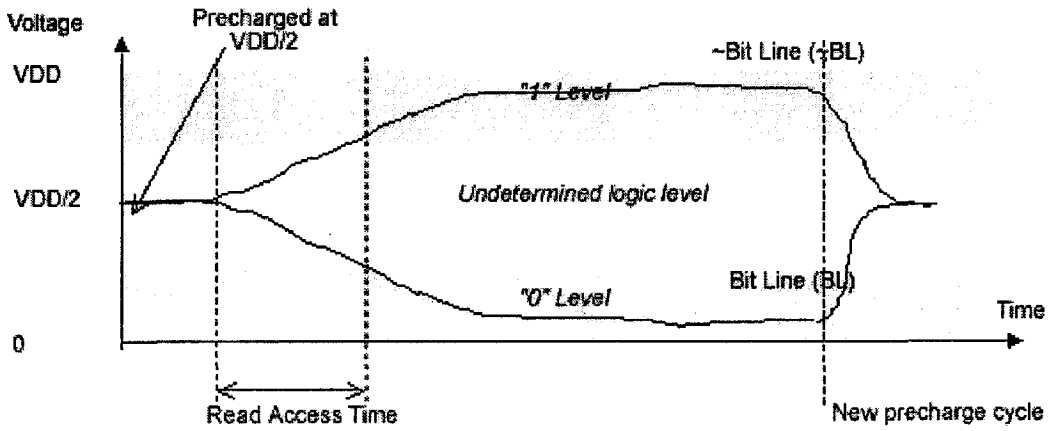


Figure 2.4: Read cycle using a precharge circuit [21]

A simple precharge circuit consists of a n-channel MOS or p-channel MOS (Both switch the voltage $V_{DD}/2$ without degradation). The drain is connected to $V_{DD}/2$, the source to the bit line.

2.4 SRAM Architecture

The overall memory architecture can now be described. In Figure 2.5, we illustrate a high-level layout of the memory array. The core array containing the cells is the largest block. The bitline precharge circuits are positioned above the core. The row decoder is placed on the left side and the column multiplexer and bit I/O are located below the core array. The row/column decoder is comprised of a predecoder and a final decoder. The row decoder drives the wordlines horizontally across the array while each pair of bitlines feeds a 2:1 column decoder (in this case) which is connected to the bitline I/O circuits such as sense amplifier and write drivers. Each memory cell is mirrored vertically and horizontally to form the array, as indicated in the Fig 2.5. Transmission Gates are used as column multiplexers to switch between the column dercoder and sense amplifier.

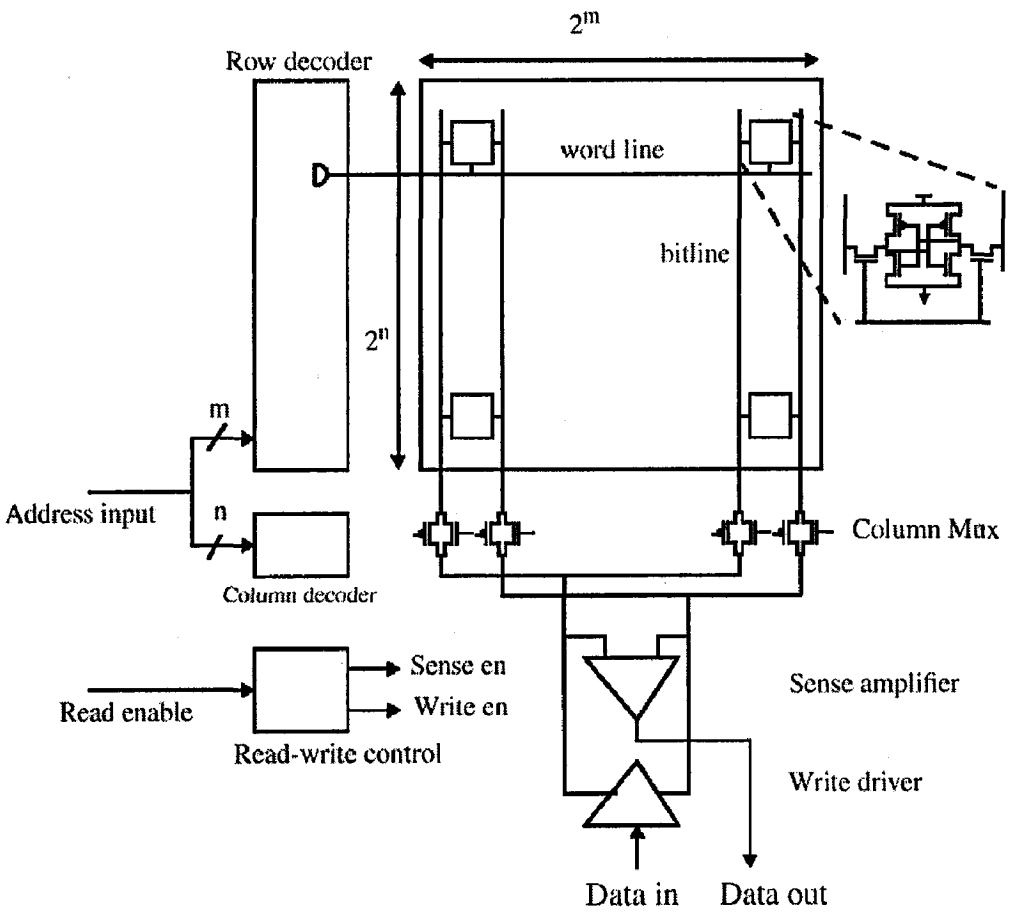


Figure 2.5: SRAM architecture for 2^{m+n} bit organized as an array of 2^n rows X 2^m columns. [23]

2.5 Power Dissipation Analysis Using Stack Effect

At nanometer regime each leakage component plays important role in determining leakage. The subthreshold leakage is the weak inversion current between source and drain of an MOS transistor when the gate voltage is less than the threshold voltage and is given by, [21]

$$I_{sub} = \mu_0 C_{ox} \frac{W_{eff}}{L_{eff}} V_T^2 \exp\left(\frac{V_{gs} - V_{th}}{nV_T}\right) \cdot \left(1 - \exp\left(\frac{-V_{ds}}{V_T}\right)\right) \qquad (2.1)$$

where μ_0 is the zero bias electron mobility, n is the subthreshold slope coefficient (for single gate devices n= $(1+C_{depletion}/C_{ox})$), V_{gs} and V_{ds} are the gate-to-source voltage and drain-to-source voltage, respectively, V_T is the thermal voltage, V_{th} is the threshold voltage, C_{ox} is the oxide capacitance per unit area, and W_{eff} and L_{eff} are the effective channel width and length, respectively. Due to the exponential relation between V_{th} and I_{sub}, an increase in V_{th} sharply reduces the subthreshold current. Subthreshold leakage is a strong function of the threshold voltage V_{th} and temperature T, since they both appear in exponential terms. Current in this regime is undesirable in digital designs, because it results in a leakage current when an ideal transistor would be completely cutoff. This leakage is especially egregious when multiplied by the millions of leakage paths present in modern designs.

Increasing the source voltage of NMOS transistor reduces the sub threshold leakage current exponentially due to negative V_{GS}, lowered signal rail (V_{CC}-V_S), reduced Drain Induced Barrier Lowering (DIBL) and body effect. This effect is also called self- reverse biasing of transistor. The self-reverse biasing effect can be achieved by turning off a stack of transistors. Turning off more than one transistor raises the internal voltage (source voltage) of the stack which acts as reverse biasing the source [24].Thus maximizing the number of off transistors by stacking and applying proper input vectors can reduce the standby leakage of a functional block.

Stacking principle had been implemented to reduce the leakage power in gates and logic circuits [25] [26] [27].The leakage current flowing through transistors connected in series depends upon the number of 'off' transistors in the stack. Turning off the stacked transistors raises the intermediate voltage to a positive value due to a small drain current.

From Fig.2.6, it can be seen that, with increasing supply voltage, leakage in single nMOS is higher as compared to stacked nMOS. It is also observed that variation of current at process corners is less for stacked nMOS. This feature provides more process tolerance to stacked structures. When an input vector "00" is applied to the gate both the transistors are turned off. As discussed earlier a small positive potential develops at the intermediate node N. This potential has the following effects:

1) The gate-to-source junction becomes reverse biased since V_{GS} is negative. As the subthreshold current is exponentially proportional to V_{GS}, it is also reduced.

2) There is an increased body effect in the top transistor due to a negative body- to-source potential and V_T is increased. Since the subthreshold current is exponentially proportional to V_T also, it is reduced.

3) Drain-to-source potential of top transistor decreases due to increase in source potential. This results in lesser Drain-Induced Barrier Lowering (DIBL). As a result the subthreshold leakage is further reduced.

This phenomenon is called stacking effect.

In [26] the effect of stacking on leakage current was extensively discussed. It was shown that the power consumption depends upon the input vectors applied to the gates. At the 90nm node applying "10" vector at the gate, reduces the gate leakage current and applying "00" vector input to a two transistors stack only reduces subthreshold leakage and does not change the gate leakage component. However the total standby power was found to be minimum when the input vector is "00" as the subthreshold leakage is dominating at the 90nm node [26][28]. This fact is exploited to reduce the standby power consumption of SRAM. However this reduction is achieved at the expense of delay penalty as the effective width of the transistor becomes W/N^2 (where W is width of the transistor before stacking and N is the number of transistors after stacking) after stack forcing. It is similar to replacing a low-V_T device with a high-V_T device in a multiple-V_T design. [25]

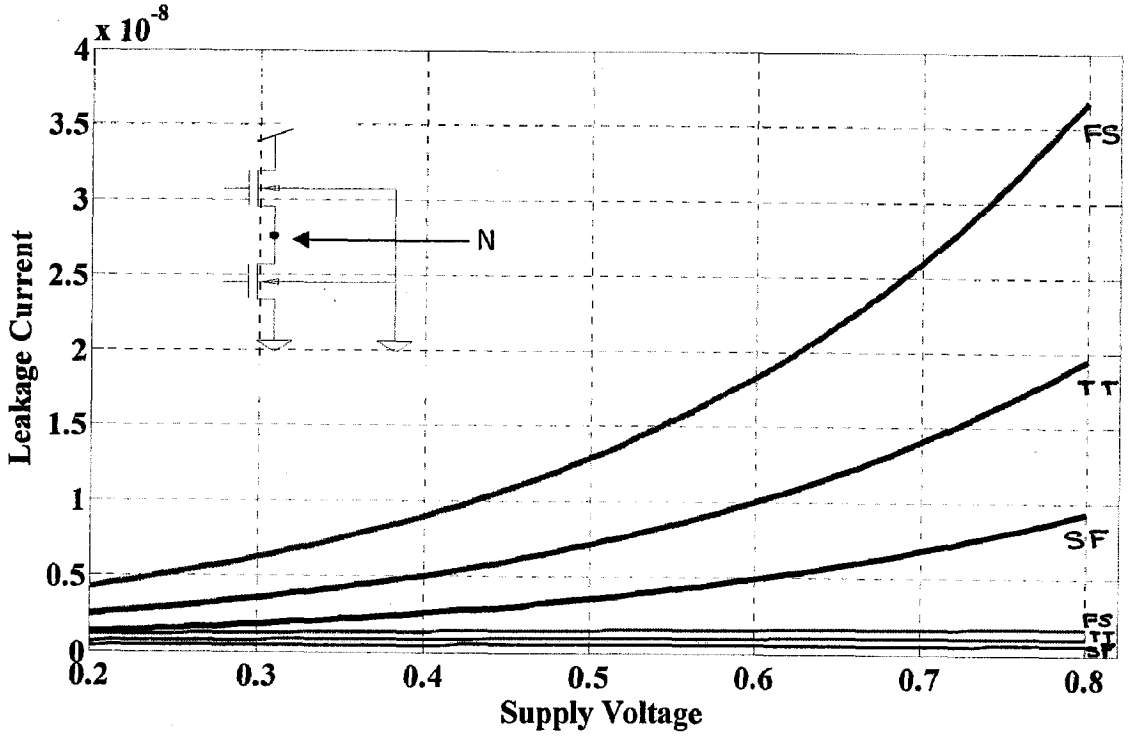


Fig 2.6: Stacked nMOS (lower three) and single nMOS(upper three) leakages at various process corners

CHAPTER 3

DATA RETENTION VOLTAGE

Data Retention Voltage (*DRV*) is the minimum value of V_{DD} required to reliably preserve data in SRAM cell. In SRAM cell, it is a measure of its state-retention capability under very low voltages. The stability of SRAM cell is also indicated by the static-noise margin (SNM) [29] under the condition shown in Fig.3.1. The *SNM* can be graphically represented as the largest square between the voltage transfer characteristic (VTC) curves of the internal inverters as shown in Fig.3.2. When V_{DD} scales down to DRV, the VTC of the cross-coupled inverters degrade to such a level that the loop gain reduces to one and SNM of the SRAM cell falls to zero.

3.1 DRV Theoretical Lower Bound

The *DRV* of a SRAM cell can be determined by solving the subthreshold VTC equations of the two internal data-holding inverters, since all the transistors conduct in weak inversion region when V_{DD} is around DRV. When an SRAM cell (Figure 3.1) is in standby mode, the currents in each internal inverter are balanced:

$$\text{Node } V_1: I_1 + I_5 = I_2 , \qquad (3.1)$$

$$\text{Node } V_2: I_3 + I_6 = I_4 \qquad (3.2)$$

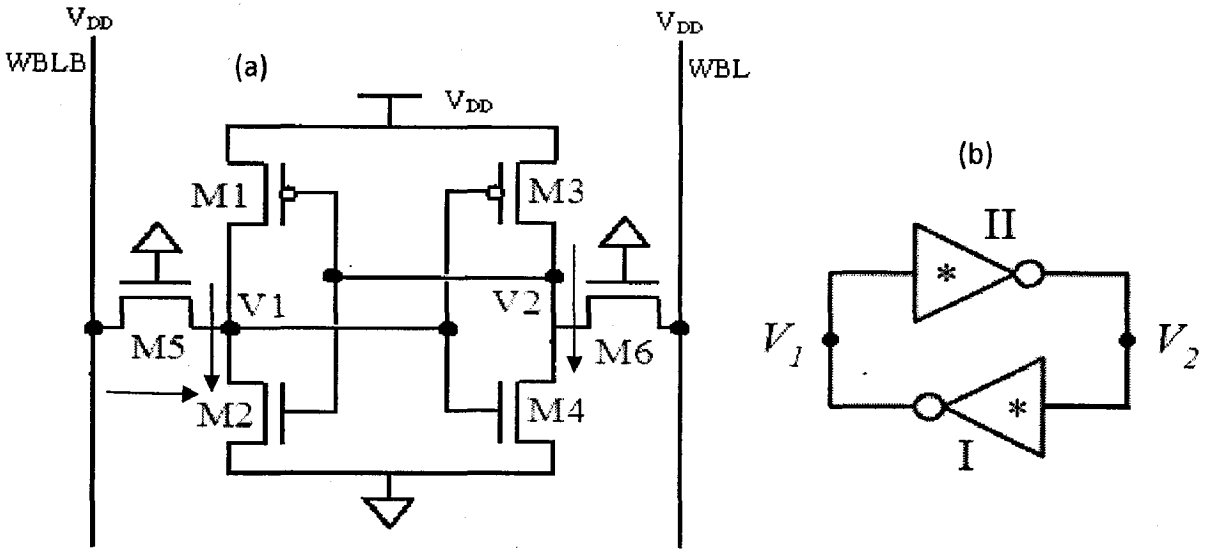


Figure 3.1: Standard 6T SRAM cell structure. (a) 6T SRAM cell in standby (assuming $V_1 \approx 0$ and $V_2 \approx V_{DD}$.). (b) Flip-flop representation of the same SRAM cell

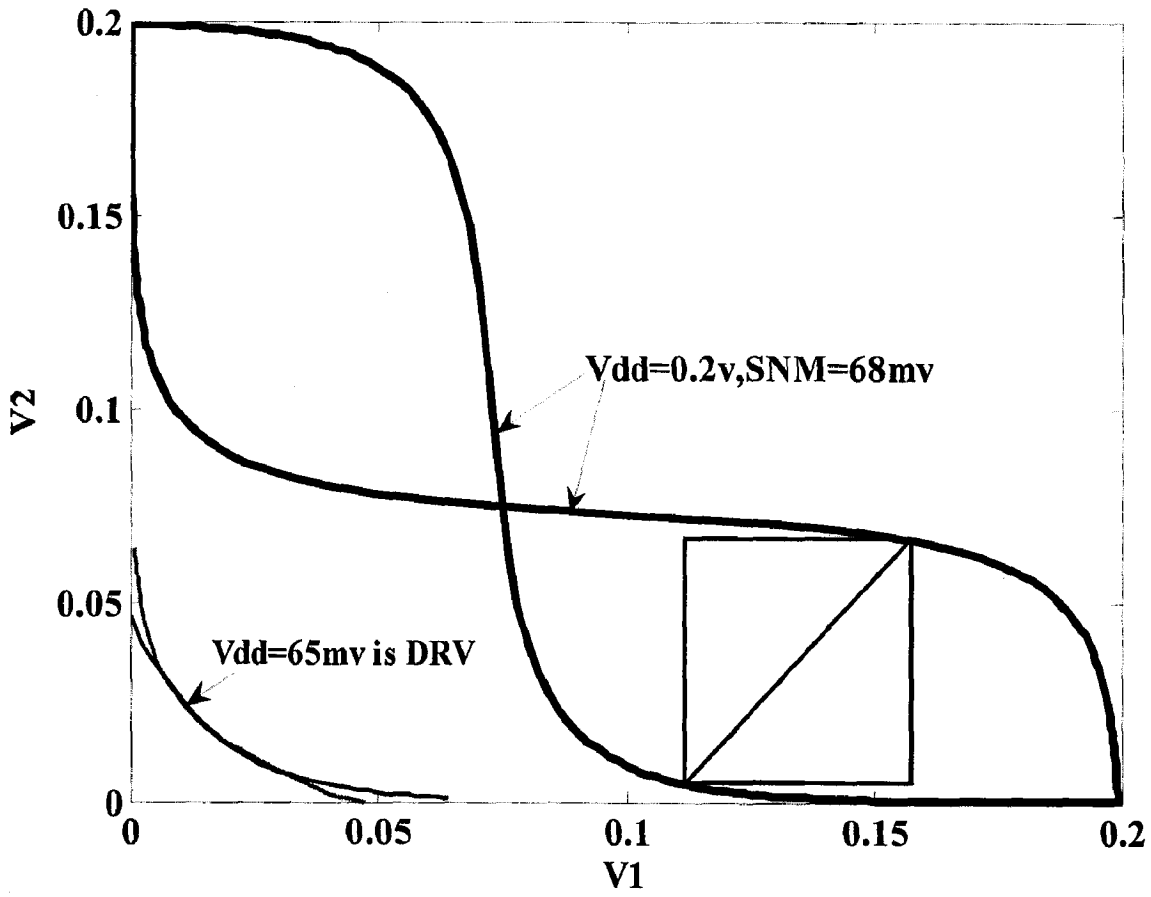


Figure: 3.2 An illustration of SRAM inverter VTC deterioration under low-V_{DD}. The SRAM cell noise margin is zero at DRV.

We may assume that the original state stored in SRAM cell is:

$$V_1 \approx 0 \text{ and } V_2 \approx V_{DD}. \quad (3.3)$$

In order to minimize mismatches and maximize the data-retention noise margin, in a theoretical DRV limit analysis we assume that the SRAM cell is manufactured with ideal process conditions, i.e., NMOS and PMOS have symmetrical V_{th} and sub-threshold slope factor, and there are no process variations. Furthermore, the leakage of access transistors (I_5 and I_6) is assumed to be totally eliminated by aggressive design optimization, e.g. reversed body bias on M_5 and M_6 during standby mode to increase V_{th}. Then, Eq.3.1 and Eq.3.2 simplifies to:

$$I_1 = I_2, I_3 = I_4. \quad (3.4)$$

Ii is the sub-threshold current of the i_{th} transistor (Figure 3.1). Assuming room temperature standby operation, I_i can be considered as dominated by the drain-source leakage. This is because at sub-threshold V_{DD} and at 90nm technology the gate leakage and other leakage mechanisms have minor contribution compared to the sub-threshold current. I_i is modeled as

$$I_i = \beta_i I_0 \exp(-V_{th,i}/n_i v_T).\exp(V_{gs,i}/n_i v_T).\left(1- \exp(-V_{ds,i}/ v_T)\right) \quad (3.5)$$

where $v_T = kT/q$ is the thermal voltage, equal to 26mV when T = 27°C; β_i is the transistor (W/L) ratio; I_0 is the leakage current of a unit sized device at $V_{gs} = 0$ and $V_{ds} >> v_T$; T is the chip temperature; and n_i is the sub-threshold factor, (sub-threshold swing divided by 60mV at room temperature). If we further define: [32]

$$I_{off,i} = \beta I_0(\frac{-V_{th,i}}{n_i v_T}) \quad (3.6)$$

The V_{thi} in Eq.3.6 can be accurately modeled as following, with the second and third terms representing the body bias effect and the drain-induced-barrier-lowering (DIBL) effect. [33]

$$V_{th,i} = V_{thi,0} + \gamma(\sqrt{\text{mod}(-2\phi_i + V_{sb,i})} - \sqrt{\text{mod}(-2\phi_i)}) - V_{sd,i}.\exp(-\alpha l_i) \quad (3.7)$$

Since all the SRAM cell transistors conduct in weak inversion region when V_{DD} is around DRV, the DIBL effect can be ignored in a DRV analysis. Substituting these current models, which are functions of V_1, V_2, V_{DD}, T, and other technology parameters, Eq.3.4 can be expanded into:[33]

$$\exp(\frac{V_{DD}-V_2}{nv_T}).[1-\exp(\frac{V_{DD}-V_1}{v_T})] = \exp(\frac{V_2}{nv_T})[1-\exp(\frac{V_1}{v_T})] \quad (3.8)$$

$$\exp(\frac{V_{DD}-V_1}{nv_T}).[1-\exp(\frac{V_{DD}-V_2}{v_T})] = \exp(\frac{V_1}{nv_T})[1-\exp(\frac{V_2}{v_T})] \quad (3.9)$$

In Eq.3.8 and Eq.3.9, $I_{off,N} = I_{off,P}$ and $n_N = n_P$ are assumed based on the symmetry requirement to maximize the data-retention noise margin. The $I_{off,N} = I_{off,P}$ condition represents a balanced

PMOS-to-NMOS (*P/N*) leakage strength ratio. Then, by solving (V_1/V_2) from Eq.3.1 respectively and using the condition of Eq.3.4, the theoretical limit of *DRV* is solved as:

$$DRV_{ideal} = 2*V_T \ln(1 + n) \qquad (3.10)$$

For an ideal CMOS technology n = 1 (i.e., 60mV/decade as the swing), which provides DRV_{ideal} = 36mV.For a typical 90nm technology with n = 1.5, DRV goes up to 50mV. These results were confirmed with SPICE simulation result from an industrial 90 nm technology.

3.2 Low-Voltage SRAM Standby Stability Analysis

In order to reliably preserve data in an SRAM cell at a low-voltage standby mode, an adequate SNM is necessary. Usually a positive SNM is created by setting the SRAM standby V_{DD} at a level higher than the DRV. The difference between the standby V_{DD} and the DRV is called the standby guard band voltage.

The SNM of an SRAM cell can be calculated in many different ways: the maximum square between the normal and mirrored VTC, small-signal loop-gain, Jacobian of the Kirchoff equations, coinciding roots [30]. These methods are well researched and it has been shown that they are all equivalent [30]. Similar to [31], taking the loop-gain approach of analyzing the SNM as the maximum value of noise that can be tolerated by the flip-flop before changing states. As shown in Figure 3.3, two noise sources, V_n, are inserted to assure the worst-case noise scenario when the noise is present in both gates in the same way [31]

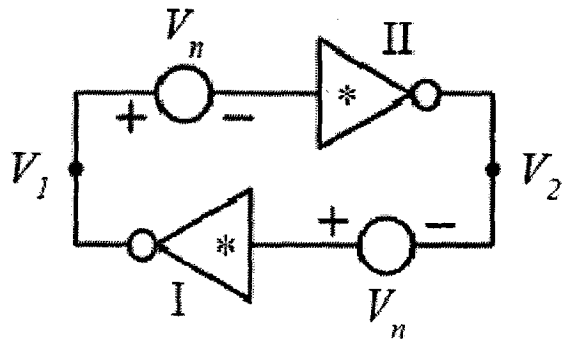


Figure 3.3: Flip-flop representation of SRAM cell with inserted static noise, V_n.

Following the methodology of DRV derivation but this time with inserted static noise V_n,

$$V_{GS2} + V_N = V_2 \quad (3.11)$$

$$V_{GS4} - V_N = V_1 \quad (3.12)$$

we obtain a zero-order approximation for *SNM* from the condition of marginal stability, that is the unity loop-gain. The maximum noise corresponding to the unity gain is given by:[33]

$$SNM = \frac{2}{3}V_{DD} - \frac{n.kT/q}{3}.\ln\left(\frac{2I_{off,4}}{n.I_{off,2}.I_{off,3}}\right) - \frac{n.kT/q}{3}.\ln\left(\frac{I_{off,5}}{n} + \frac{I_{off,1}}{n}.\exp\left(\frac{SNM}{n.kT/q}\right)\right) \quad (3.13)$$

It is assumed that there is equal sub-threshold slope factors have been assumed for NMOS and PMOS transistors. The above formula does not have a closed-form solution, but can be solved iteratively. It can be shown that this zero order SNM can be expressed as [33]

$$SNM = K \cdot (V_{DD} - DRV), \quad (3.14)$$

Further expansion of Eq.3.13 and comparison with Eq.3.14 yield the following approximation of the K factor:

$$K = 2/(3 + n), \quad (3.15)$$

Where $I_{off,5}$ from Eq. 3.13 is neglected due to exponential nature of the other term under the logarithm. This approximation is valid for SNM > nkT/q. The result in Eq. 3.15 means that a smaller sub-threshold factor is desirable for higher noise tolerance in standby mode. This linear correlation of SNM and the standby guard band voltage facilitates the SRAM design for reliable data retention under low voltage. For example, in order to achieve a 50mV SNM under 3σ local process variations, the SRAM standby V_{DD} needs to be 100mV higher than the corresponding DRV.[32]-[34]

PROPOSED 10T SRAM DESIGN

The key to the microprocessor cache market is high performance, high stability, low power consumption. With the excellent performance and stability of the 6T SRAM, it has been dominating. Here it is shown that with some modifications of the cell, we can design the cell that can work efficiently in super threshold as well as sub threshold region and is more robust to process variation.

4.1 Cell Structure

In a normal 6T cell both storage nodes are accessed through NMOS pass-transistors. This is necessary for the writing of the cell since none of the internal cell nodes can be pulled up from a stored '0' by a high on the bitline. If this was not the case an accidental write could occur when reading a stored '0'.However, if separate buffer is used to read data from the bit lines this is no longer true, as shown in Fig. 4.1. Here stacking is used to prevent leakage current during standby voltage and in addition to it the buffer keeps the stability and performance of the cell high.

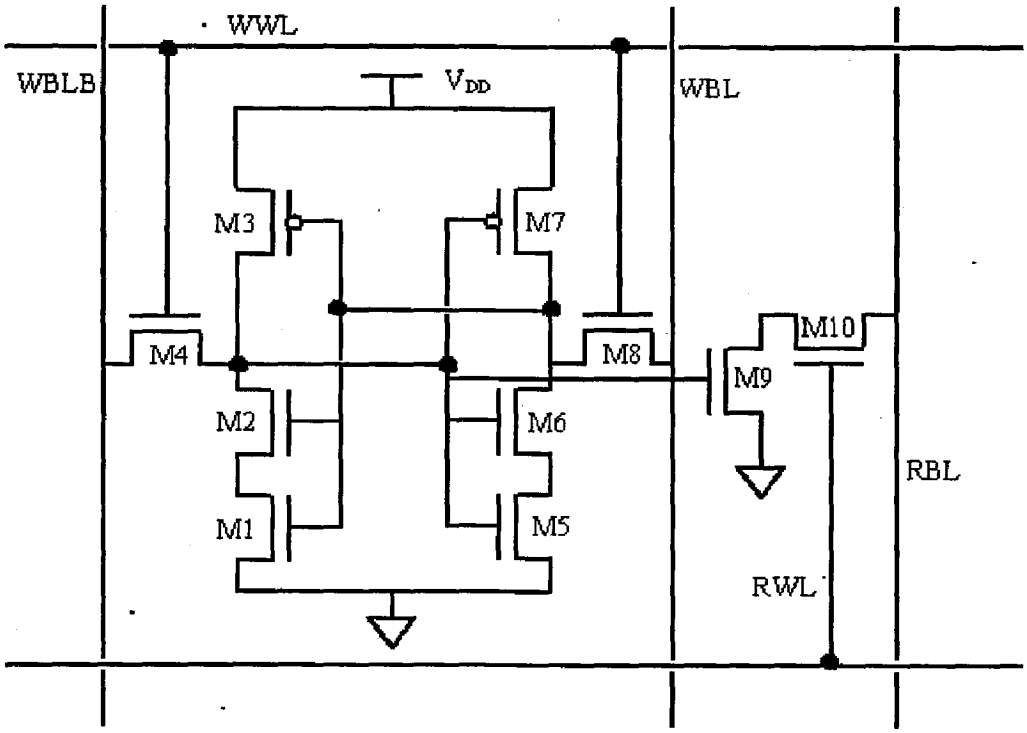


Figure 4.1: 10T SRAM cell design with M9 and M10 MOSFET used as buffer and M4-M6/M1-M3 stacked for leakage reduction.

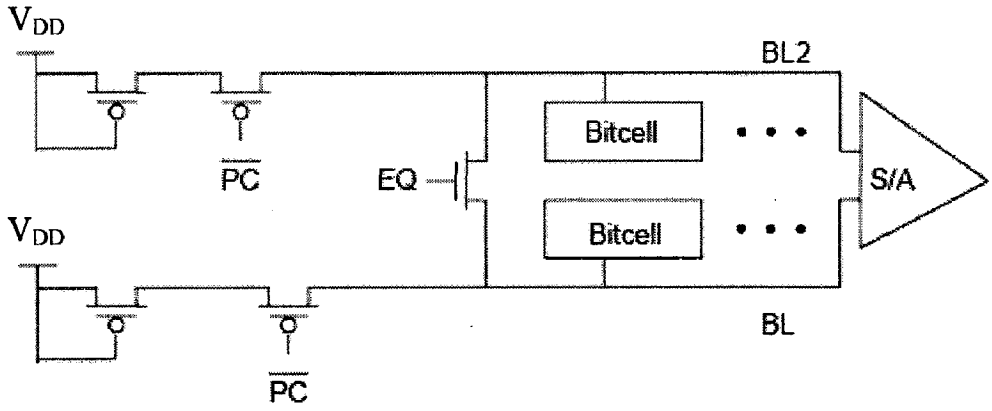


Figure 4.3: Open Bit line architecture for single ended sensing

As single read bit line is associated to each cell hence single ended sensing is converted to differential by the approach of open bit line architecture as seen in the Fig.4.3. Here the cells in single column are divided into two equal parts with differential amplifier placed in middle. On each side the bit line capacitance will store the charge once the precharge operation is done [21]. The delay was measured when the voltage levels had reached 90% of their steady state values of supply voltage and 0V. The differential sensing proves faster then single ended sensing [9]. The basic concept behind the single to differential conversion is used here. The memory array is divided into two halves, with the differential amplifier placed in the middle. When the EQ signal is raised both the BL and BL2 are precharged to $V_{DD}/2$. Now let us assume that the cell in BL is accessed and it leads to pull down of BL, while BL2 remains at $V_{DD}/2$. Differential voltage is generated at the inputs of the amplifier resulting in sense latch to toggle. It also has a advantage that dividing the bitline reduces the bit line capacitance. This doubles the charge transfer ratio and improves the signal to noise ratio.

So during active reading time, the buffer transistors are accessed, hence there is no deterioration in performance as compared to the 6T, along with which we get high read stability. While stacking of driver transistor is used to reduce leakage power during standby mode.

4.3 Writing Operation

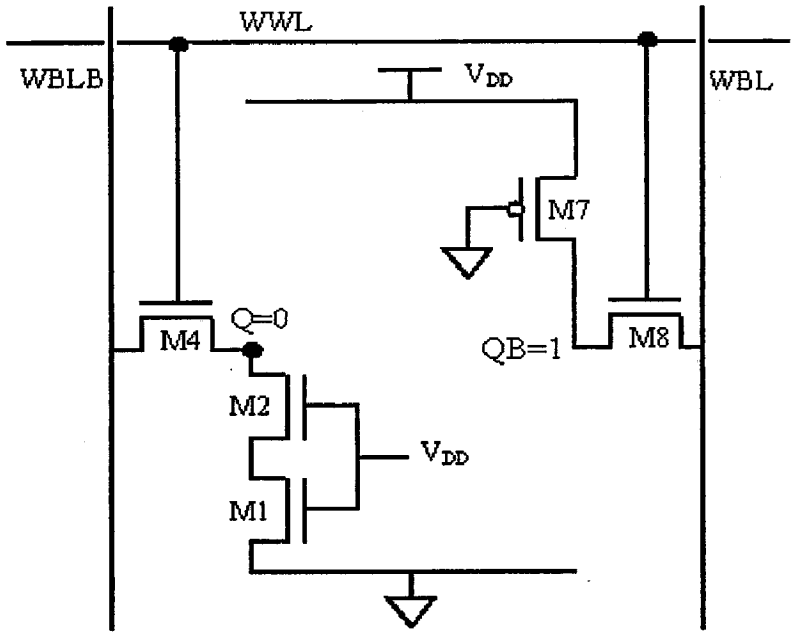


Figure 4.4: Six-transistor SRAM cell at the onset of write operation (writing '0'→'1').

Similar to standard 6T SRAM cell, writing is done by lowering one of the bitlines to ground while asserting the wordline. To write a '0' *BL* is lowered, while writing a '1' requires *BL* to be lowered. The simplified diagram during write operation is shown in the Fig.4.4. Unlike the read, the bitlines are held at V_{DD} and gnd respectively during write operation since both bitlines are now held at their respective value, the bitline capacitances have been omitted. During the discussion of read the storage nodes were not disturbed, it was concluded that driver transistor need not be stronger than access transistor to prevent accidental writing. Hence size of driver transistor can be kept equal to load transistor hence some area could be saved. Along with area saving in the write case, this feature actually helps in desired write operation. Even when transistor *M5* is turned on and current is flowing from *BL* to the storage node, the state of the node will not change. In traditional 6T cell as the node is raised, driver transistor will sink current to ground, and the node is prevented from reaching even close to the switching point. This is not the case with 10T configuration where stacking makes the pulldown path weaker which is helpful in reaching closing near to switching point. So on left side of circuit '1' is written and at right side of the cell BL is held at gnd and when the wordline is raised M8 is turned on and current is drawn from the inverse storage node to *BL*. At the same time, however, M6 is turned on and, as soon as the potential at the inverse storage node starts to decrease,

current will flow from V_{DD} to the node. So M8 is kept slightly stronger then M6 as inherently NMOS transistor is stronger then PMOS (the mobility is lower in PMOS than in NMOS). Therefore making both of them minimum size, according to the process design rules, will assure that M8 is stronger so that writing is possible. When the inverse node has been pulled low enough, the transistor M1-M2 will no longer be open and the normal storage node will also flip, leaving the cell in a new stable state. Hence in case of stacked cell, writing becomes faster due to weaker pull down path and turning off of M1-M2 by falling voltage of node Q during write as seen in Fig.4.4.

4.4 Static Noise Margin

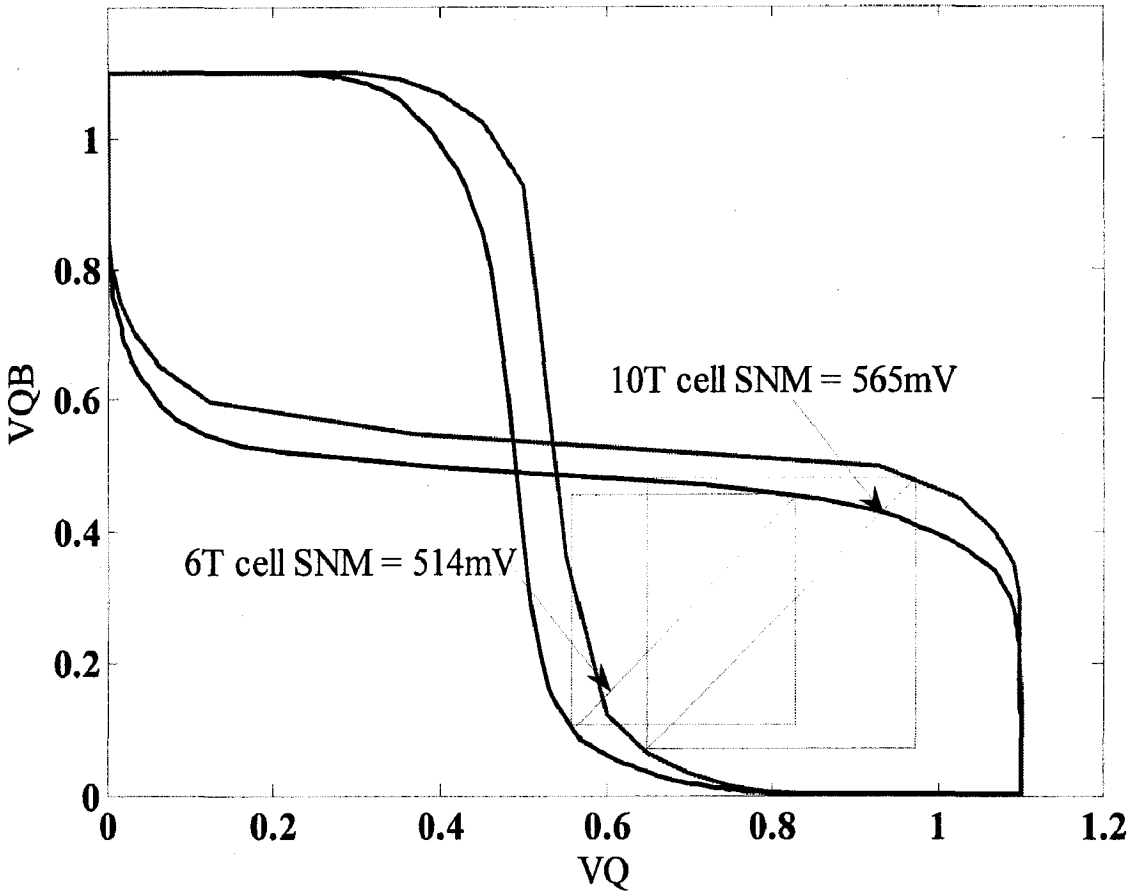


Figure.4.5 Shows the Butterfly curve at supply voltage of 1.1V

Noise margin is the maximum amount of voltage noise that can be introduced at the outputs of the two cross coupled inverters such that the cell retains its data. In the standby mode this is referred to as Static Noise Margin (SNM), during the read operation as Read Noise

Margin(RNM) and during write operation as Write Noise Margin (WNM) [31][35][36]. Figure.4.5 shows the graphical method to calculate SNM. There is an improvement of about 50 mV in SNM value in stacked (565mV) as compared to the conventional cell (514mV) at 1.1V. However earlier works have reported an SNM value of 300mV at 130nm at V_{DD} of 1.5V [37], 200mV at 130nm at 1.2 V_{DD} [38], 110mV at 50nm at V_{DD} of 1V [39]. Fig.4.6 shows that for the given voltage range the stacked SRAM has superior noise margin because in conventional cell the load transistor are kept weaker and driver transistor is made stronger to increase the drive current high during read operation. This adversely affects the SNM. In butterfly curve two lobes will become unequal reducing SNM in conventional SRAM. In case of 10T cell pull-down and pull-up paths remain equally strong and hence we get maximum SNM.

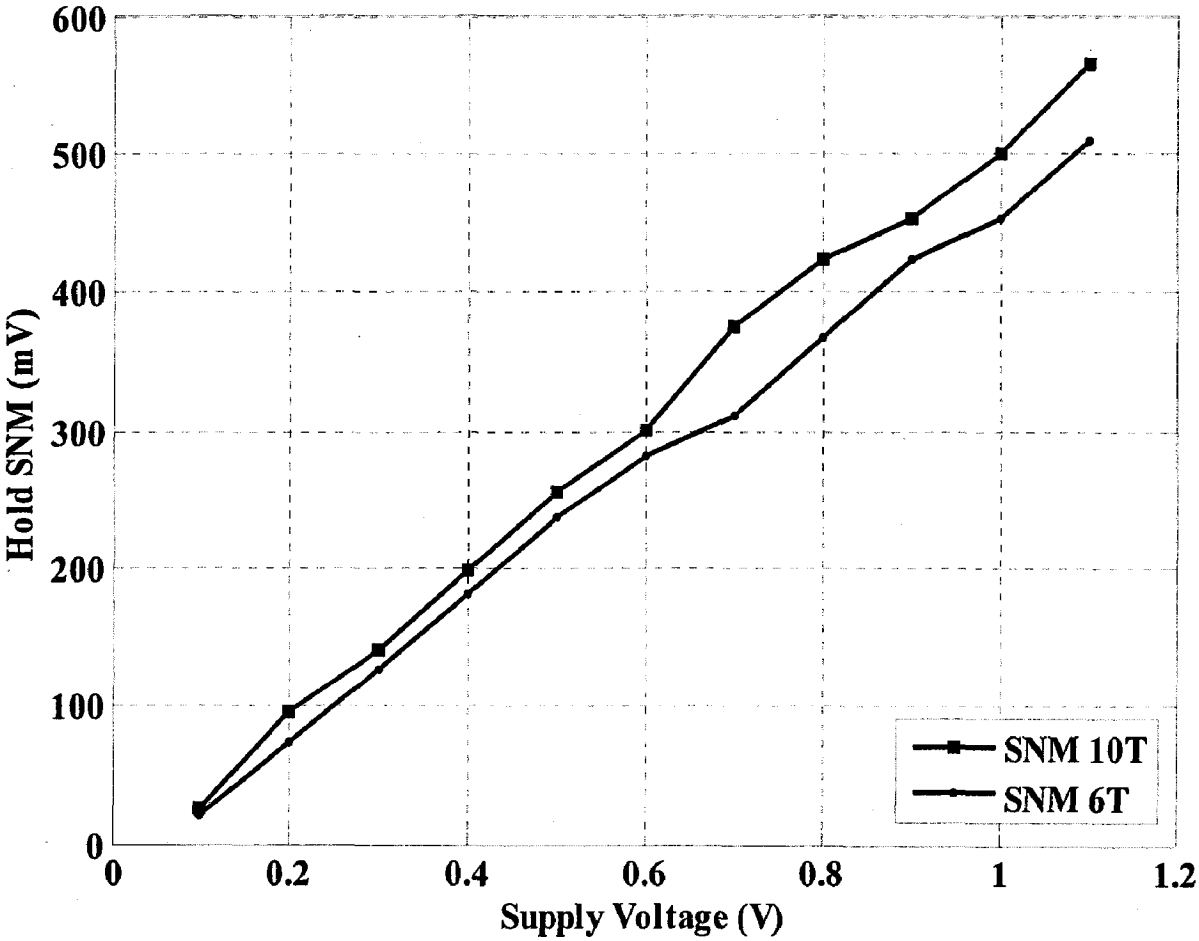
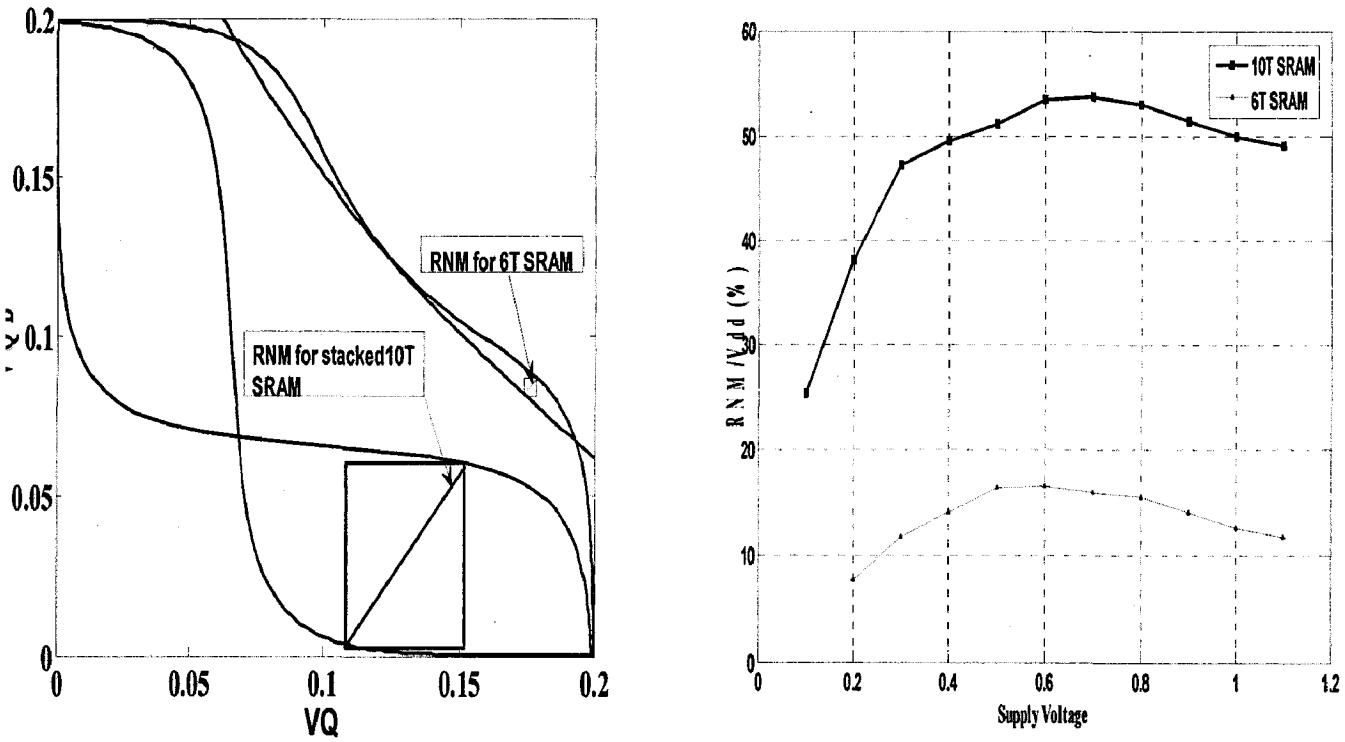


Figure.4.6 Shows the SNM comparison of 6T and 10T cells by varying the supply voltage

4.5 Read Noise Margin

Near subthreshold voltages, during read operation the RNM is too less to be of some practical use. Hence use of 6T cell configuration at subthreshold voltages leads to data flipping and losses. In 10T cell extra buffer keeps the cell robust during read operation. Fig.4.7 (a) shows that at 200mV the RNM for conventional SRAM falls to 15mV, while that for the 10T cell remains near to 75mV. In Figure 4.7(b) normalized RNM is drawn which indicates the RNM per unit supply voltage. It is the measure of measure of stability of cell with respect to supply voltage. Around 700mV 10T cell has maximum stability. It can be observed that RNM in conventional cell is very less in comparison to the new 10T structure.



(a) (b)

Figure.4.7 (a) RNM comparison between 6T and 10T cells at 200mV, (b) Variation of normalized RNM with supply voltage

4.6 Voltage Scaling

Another important issue in today's IC design is voltage scaling i.e. will it be possible to work with the same design at lower voltages. This will in be an indication of how well the design can be used in a smaller technology since the supply voltage generally decreases with every new generation of processes. Power dissipation is a large problem in today's microprocessors. It limits the battery life of mobile applications and it also makes the chips so hot that malfunction

can occur. To deal with these issues the supply voltage is often lowered. With lowering supply voltages the access to cell during read and write becomes slower. It is observed in Fig.4.8 that in subthreshold region the change in maximum access frequency is exponential, its due to the fact that in subthreshold region drain current depends exponentially on V_{DS}. Above subthreshold the dependence is follows square law.

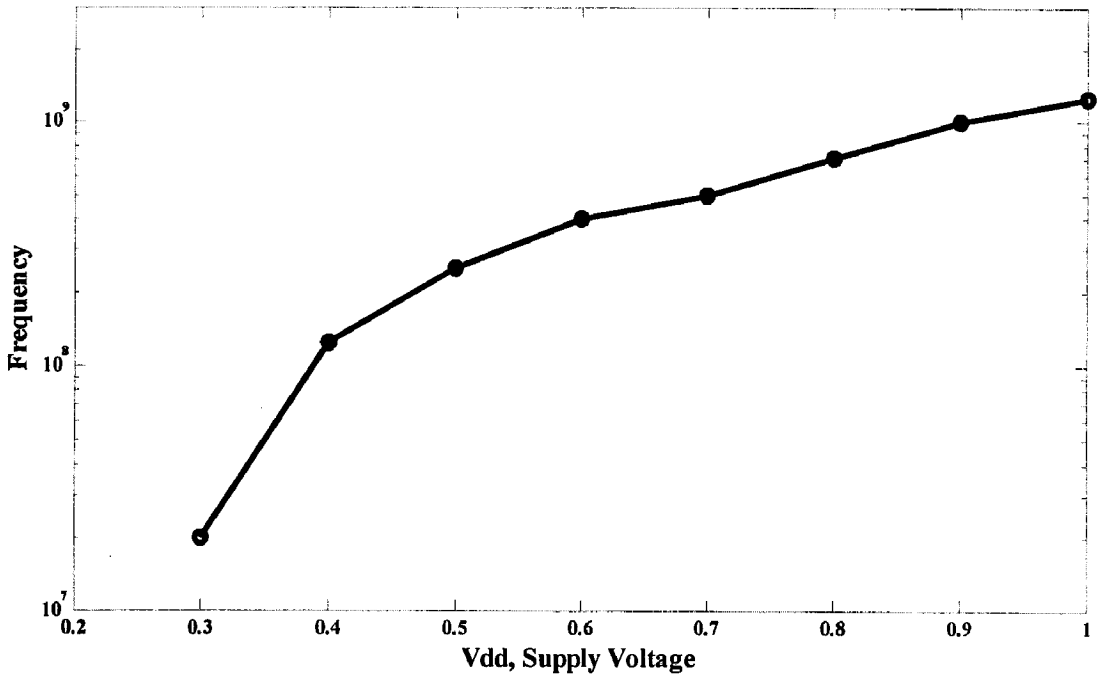


Figure.4.8 Maximum access frequency attained for different supply voltage

4.7 Sensitivity to Process Variations and Mismatch

Random dopant fluctuation (RDF) and processing variation are dominating effects in modern nanometer technologies. Both prominently change the resulting threshold voltage of devices. In sub-Vt, where Vt has an exponential effect on the drain current, the resulting impact is overpowering [40]. An additional consideration is geometric variation, particularly in effective channel length, which impacts the drift mechanism as well as short-channel characteristics like DIBL. In sub-V_t, however, the reduced V_{DS} mitigate the strength of DIBL [41]. Consequently, RDF, due to its exponential impact through V_t, is the dominating source of variability affecting functionality, performance, and energy efficiency. The exponential variation in sub-Vt drain current is particularly problematic in the face of severely reduced I_{ON}/I_{OFF}. Nominally, the

I_{ON}/I_{OFF} of devices in a circuit operating at the minimum energy voltage is between10^3–10^4, whereas that in strong inversion is approximately 10^7 [42]. Degradation in drain current, due to variation, however, can severely reduce this ratio even further. This introduces a very relevant failure mechanism in SRAMs.

Random and systematic fluctuations in channel length, doping concentration, and gate-oxide thickness cause variations in MOSFET characteristics. The read stability of a 6T SRAM cell is determined by the ratio of the current produced by the access transistors and the nMOS transistors in the cross-coupled inverters. The relative strength of the nMOS transistors in the inverters can vary as compared to the access transistors due to process parameter fluctuations. The read stability, therefore, fluctuates with the process parameters. In this section, read stability and leakage power variations of the SRAM cells due to process fluctuations in gate length (L_{gate}), channel doping concentration (N_{ch}), and gate-oxide thickness (t_{ox}) are evaluated. L_{gate}, N_{ch}, and t_{ox} are assumed to have normal Gaussian statistical distributions. Each parameter is assumed to have a three sigma (3σ) variation of 10%.

As seen in figure 4.9, process variations can have a large impact on stability. But 10T cell is showing more robustness towards process variation.These process variations however only take into account differences between types of transistors, not differences between transistors of the same type. For example, the corner N Fast/P Slow means that all NMOS transistors are in the fast process corner and all PMOS transistors are in the slow corner. Today another type of variation is becoming increasingly important. It is the so called *mismatch* between transistors. This means that two transistors of the same type can have different properties. For instance the lengths can vary slightly, or the size of the drain area. To simulate the effects of mismatch a *Monte Carlo* simulation is usually done. Monte Carlo simulation is a way of using given process variations and applying statistical spread. The same simulation is performed very many times and for every time, slightly different parameters are used. The statistical spread between transistors is supplied by the manufacturers, and the values used for each simulation are determined according to Gaussian distribution

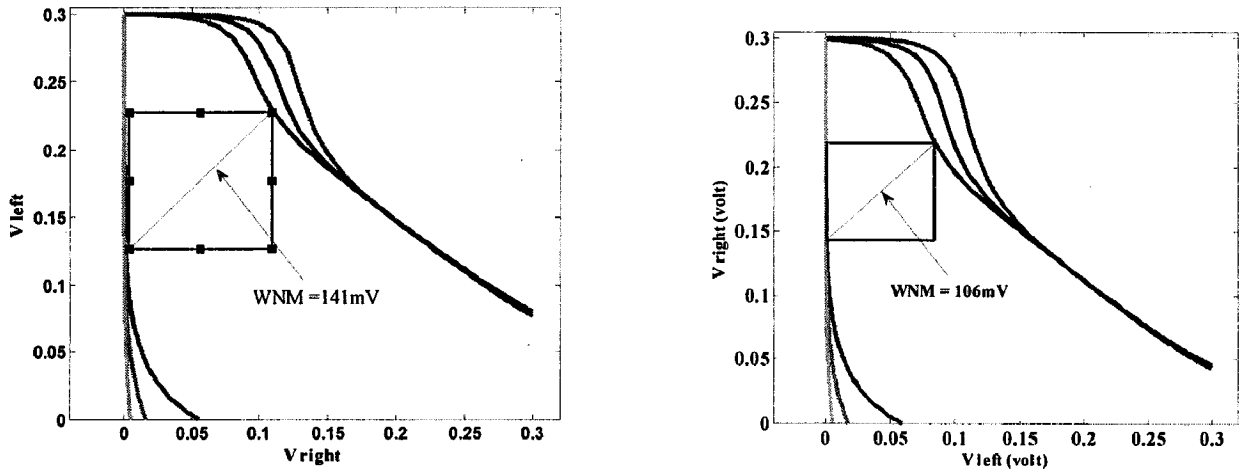


Figure 4.9: Shows WNM of 10T cell and (b) for 6T SRAM at 300mV

4.8 Sizing and Layout

A 6T SRAM cell is absolutely symmetric in the layout. Since the cell is read differentially this is very important. Both storage nodes must have the same capacitances and the same sizing of the transistors connected to them. Fig. M1-M6 are kept minimum sized as reduced drive transistor width will aid in :-

1. Faster write operation as weaker drive transistor means faster pull up of the node storing '0' when data is written into the cell.
2. The access transistor width can be decreased because drive transistor are now weaker. Hence we can save area of M1-M8 transistor.
3. The access NMOS and buffer NMOS are kept wider so as to have the faster write and read operation.

The transistor Width were optimized and found to be as follows

Transistor Type	Transistor Width (at 90nm)
Access-NMOS	220nm
Drive-NMOS/Load-PMOS	180nm
Buffer-NMOS	220nm

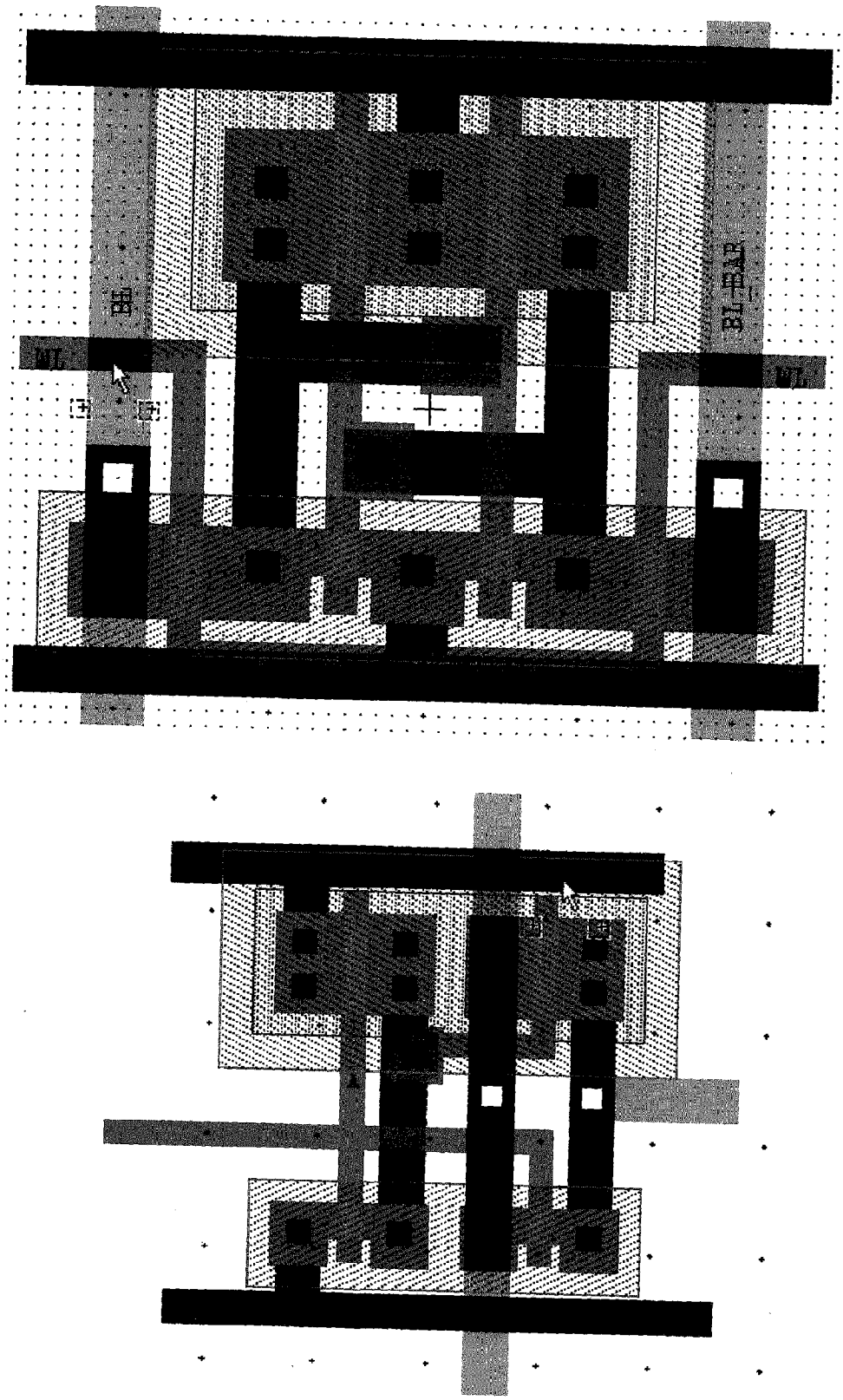


Figure.4.10 Layout of TG with metal 2 and Conventional 6T SRAM

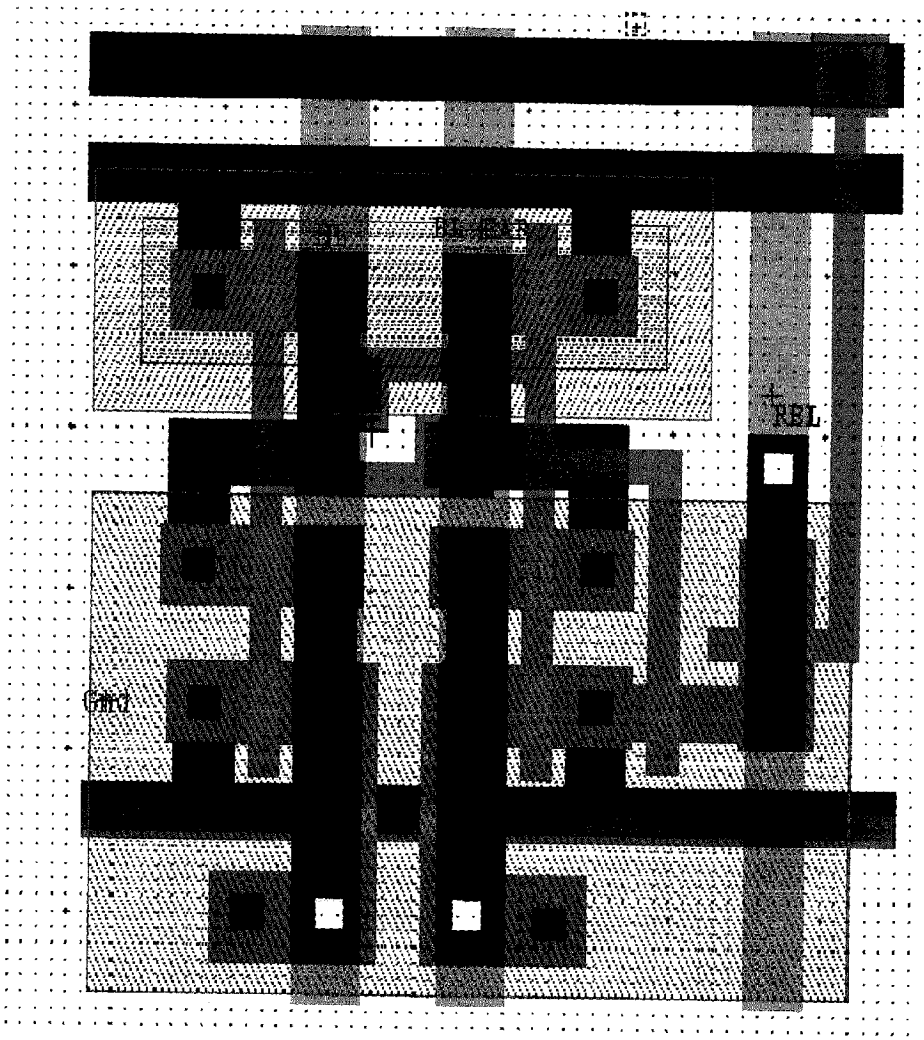


Figure.4.11 Layout of 10T stacked SRAM

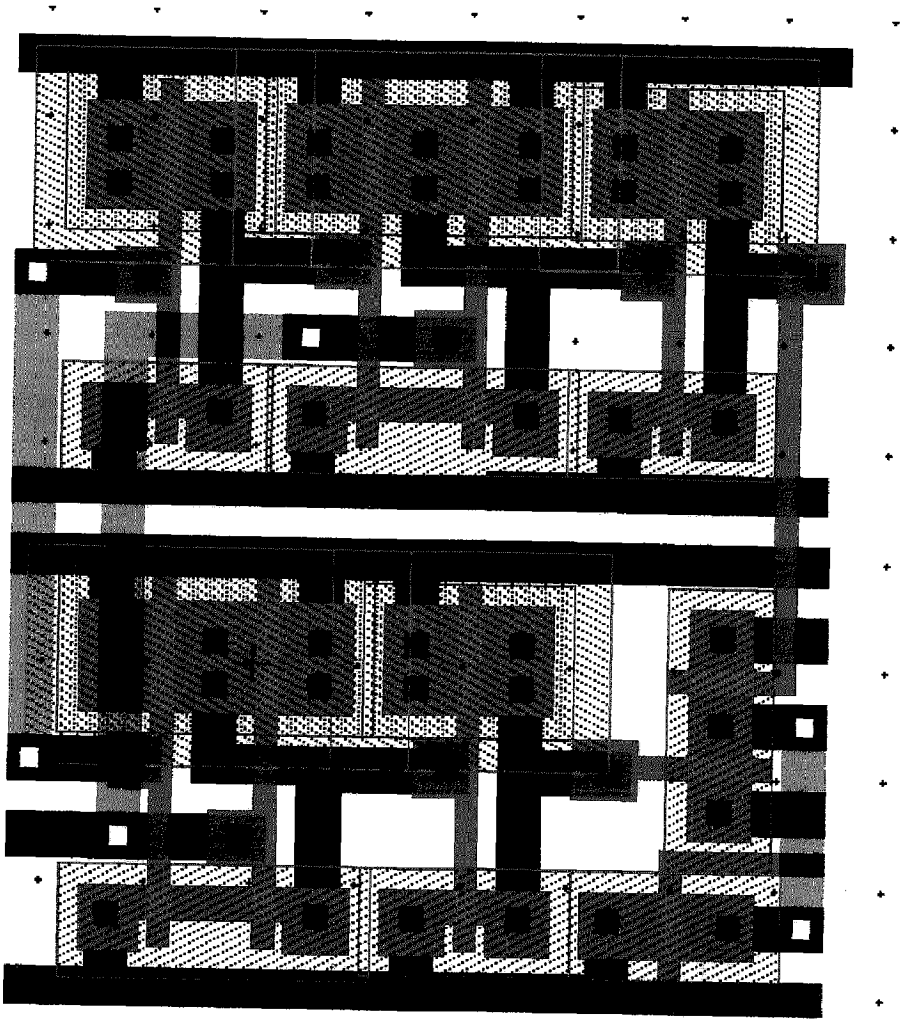


Figure.4.12 Layout of Write Driver

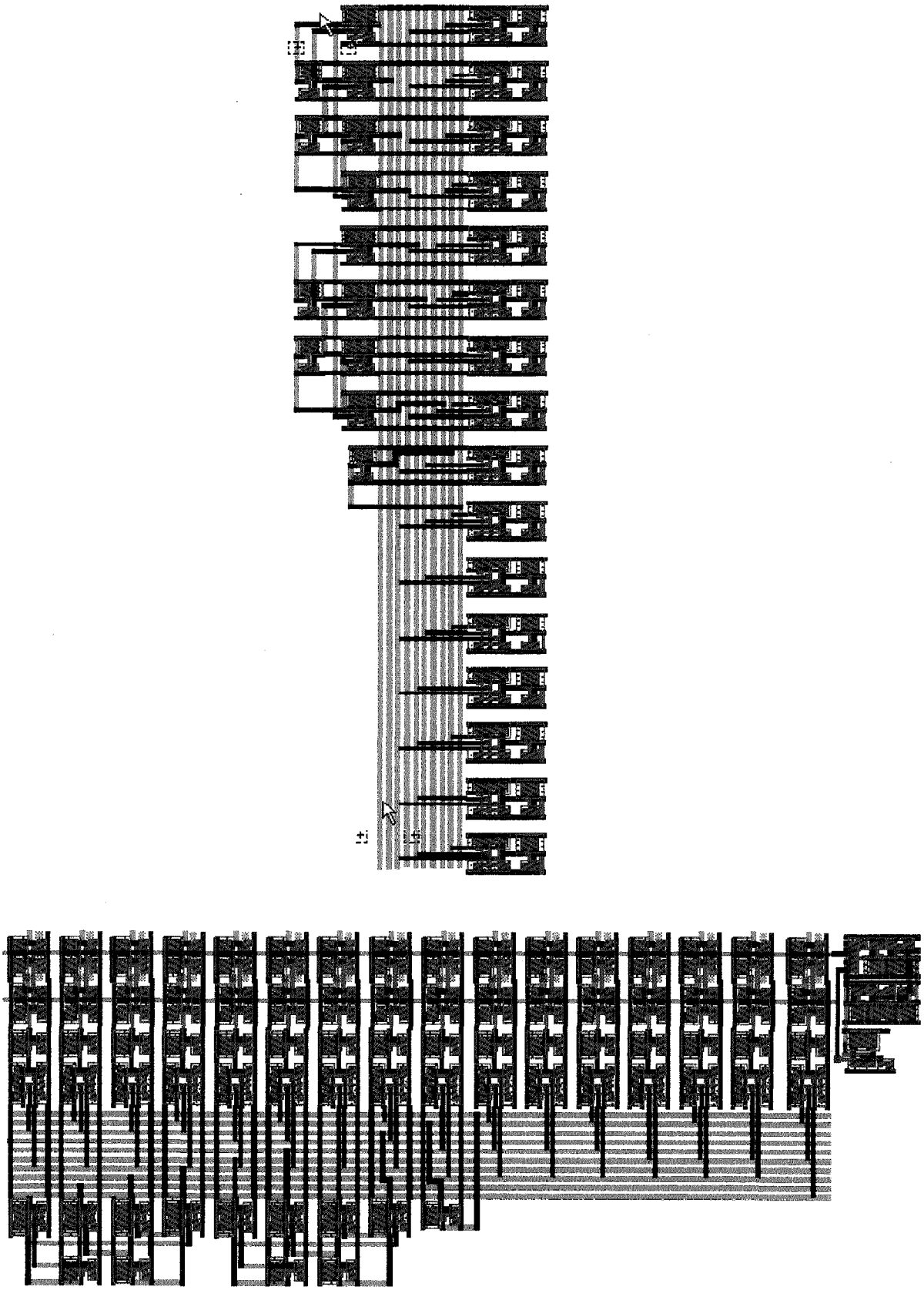


Figure.4.12 Layout of Row and Column Decoder

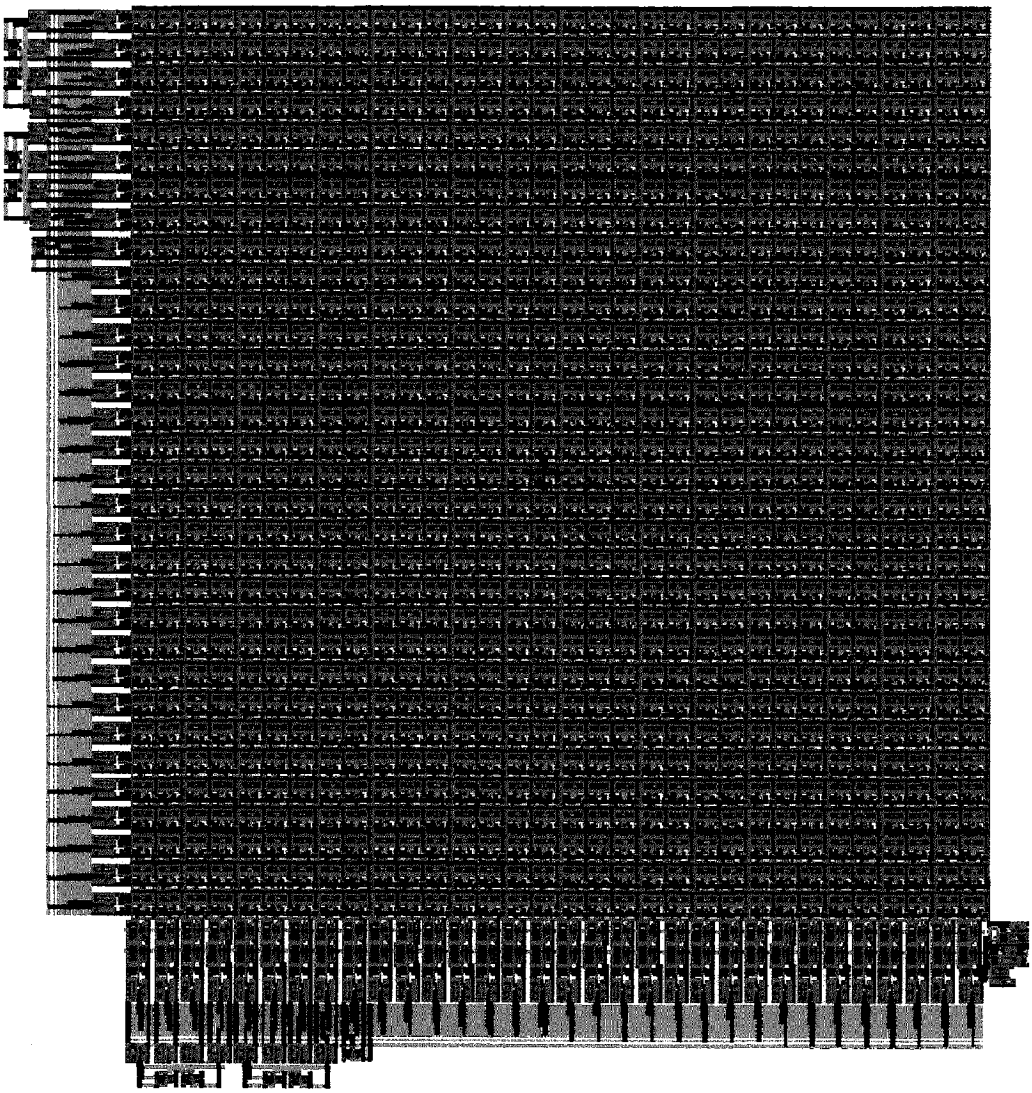


Figure.4.13 Layout of 10T SRAM with Peripheral Circuit

4.9 Current Leakage

The leakage of a memory cell can be divided into two separate discussions: leakage to the bitline(s) and total leakage through the latch. The total leakage is interesting from a power dissipation point of view. The total power dissipation should be held down to stop the chip from Over heating and increase battery life of mobile applications. There is a significant decrease in the power losses in latch due stacking of driver nMOS, which have following effects

1) The gate-to-source junction becomes reverse biased since V_{GS} is negative. As the subthreshold current is exponentially proportional to V_{GS}, it is also reduced.

2) There is an increased body effect in the top transistor due to a negative body- to-source potential and V_T is increased. Since the subthreshold current is exponentially proportional to V_T also, it is reduced.

3) Drain-to-source potential of top transistor decreases due to increase in source potential. This results in lesser Drain-Induced Barrier Lowering (DIBL). As a result the subthreshold leakage is further reduced.

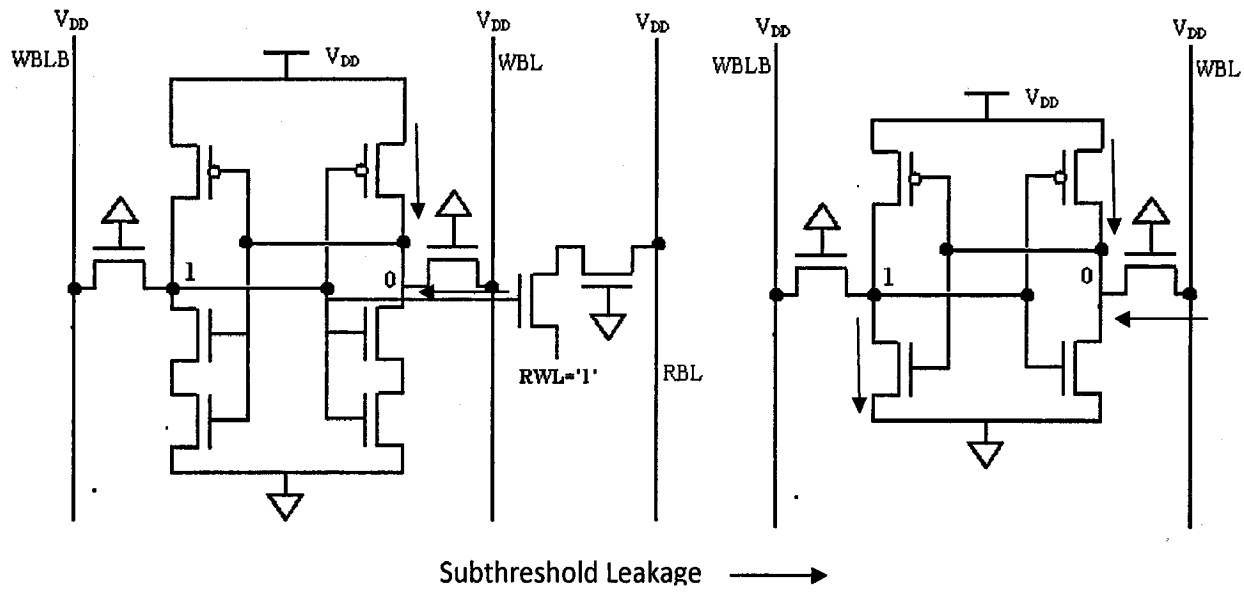


Figure.4.17 Subthreshold Leakage paths in 6T cell = 3 paths and in 10T = 2 paths

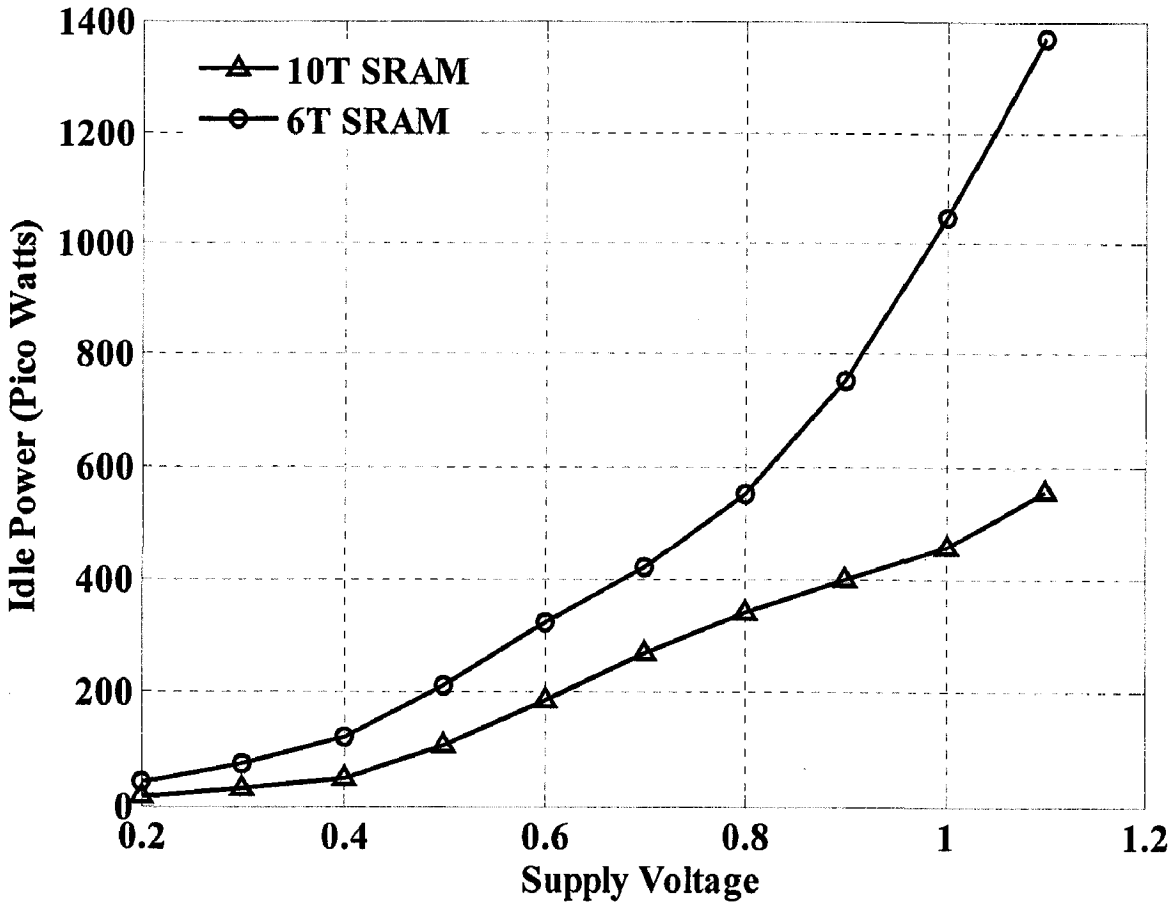


Figure.4.18 Shows the leakage power during standby mode in single bit of SRAM

Bitline plays the important role in leakage is that it contributes to the total power dissipation. The other is a performance issue. In 6T cell a stored value is read by amplifying a difference between one bitline (BL) and a reference (BLbar or BL2). To ensure a correct evaluation, the bitline and the reference must have a significant voltage difference. If we consider the case when all cells on a particular bitline (6T SRAM), except for the cell being read, has a '0' stored, an interesting situation occurs. After the precharge both bitlines are high (V_{DD}), but as soon as the precharge is lifted the sub-threshold leakage of the cells start affecting the bitline. If all the cells are storing a '0' all the leakage currents will discharge BL whereas the reference BLbar will remain high (a '1' is stored on the inverse node). When wordline is asserted the cell to be read (storing a'1') starts discharging through BLbar. and charging BL, while the other cells are still leaking. It now takes a bit of time before the voltages of BL and BLbar pass each other; time that has essentially been wasted since the same voltage difference as before must be developed for reliable read.

This is a worst case scenario, but since there is no way of knowing what values are stored in the memory, the timing has to be designed for the worst case. Another implication of this is that more cells on a bitline increase this problem. Therefore, if there is a lower leakage current from each cell, more cells can be put on the same bitline. In 10T cell leakages are lower Fig 4.17.

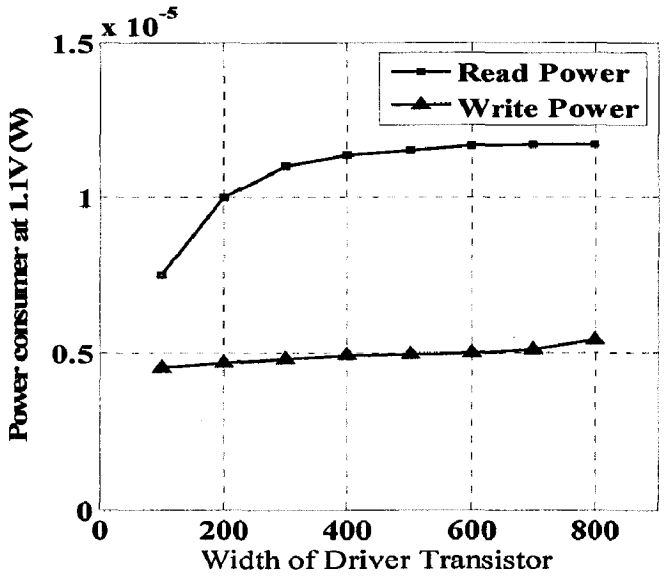


Figure.4.19 Variation of Idle read/write power with width of Driver transistor

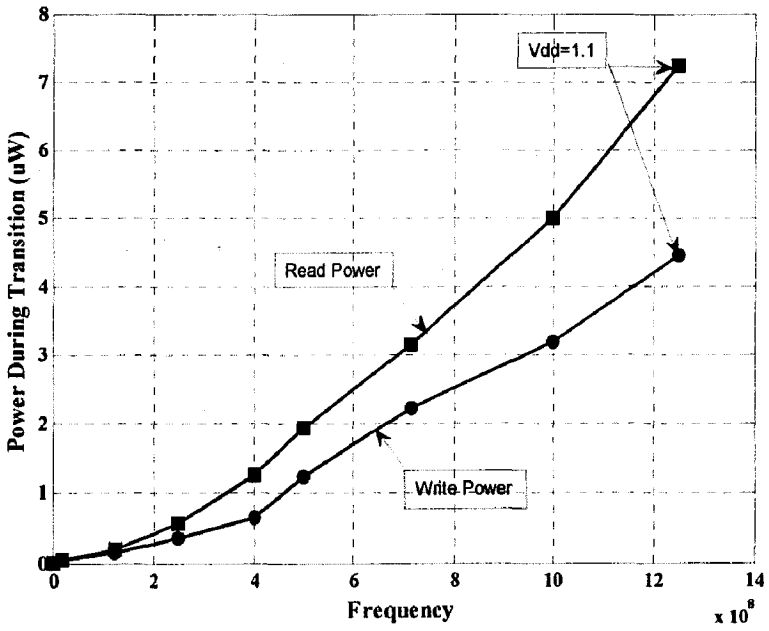


Figure.4.20 Variation of read/write power with access frequency

In Fig.4.18 shows the idle power consumed by memory cell when it is not accessed, as seen in subthreshold region the less then 100 PicoWatts and it increases linearly with voltage in this region due to the fact that power leaked will depend linearly on the supply voltage. Beyond

it the leakage power increases in parabolic way. In Fig.4.19 shows the effect of transistor sizing on power consumed. As we are using minimum sized transistor in drive NMOS, hence saves power consumed during read operation. It can be seen that read power remains higher then the write power, its due to the fact that during read the differential voltage has to be developed over the bit lined whose capacitance is hundred times that of the single bit. In case of write operation cell is flipped which is much smaller then the bit line's capacitance, hence need less power to be written. Although the stacking was done primarily to reduce the leakage power the figures show a reduction in the active power too.

In Fig.4.20 it is observed that as the voltage is decreased the maximum access frequency decreases i.e. it take more time to read the data and to write upon it, but power consumption is saved due to the reduced V_{DD}. At higher voltages the power consumption increases drastically. Fig.4.21 describes read/write power at given voltage when access frequency is kept at maximum value.

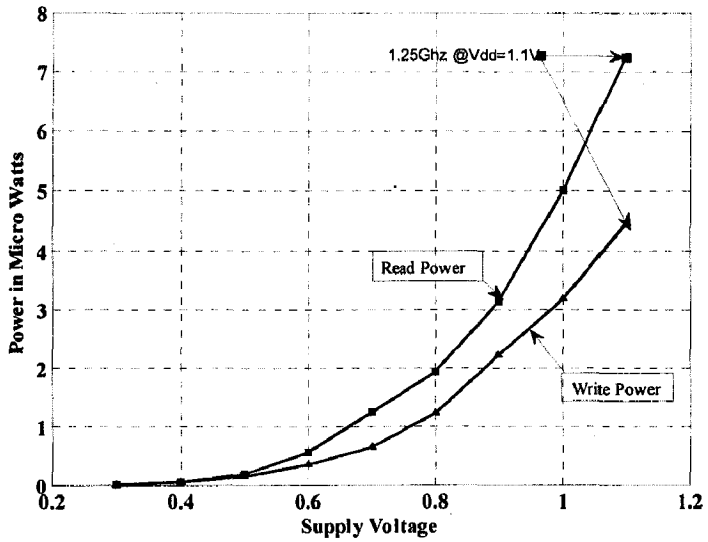


Figure.4.21 Variation of read/write power with supply voltage

The access time for the SRAM consists of the sum of the delay times of the four circuit stages- address buffers, decoders, memory cell arrays and sense amplifier circuits. However the major contributors to the access time are the decoder and the sense amplifier circuits [22] as seen in Fig.4.3.

Subthreshold operation holds significant promise of ultra low energy operation for emerging applications such as environmental and biomedical sensing and supply chain management. The key obstacle in the subthreshold design is the robustness of SRAM design.

At high supply voltages, the SRAM memory cell sizing is determined by the read/write conditions. In our subthreshold analysis, the read/write operation conditions are performed at the worst case process corners to account for the large subthreshold idle currents[3].

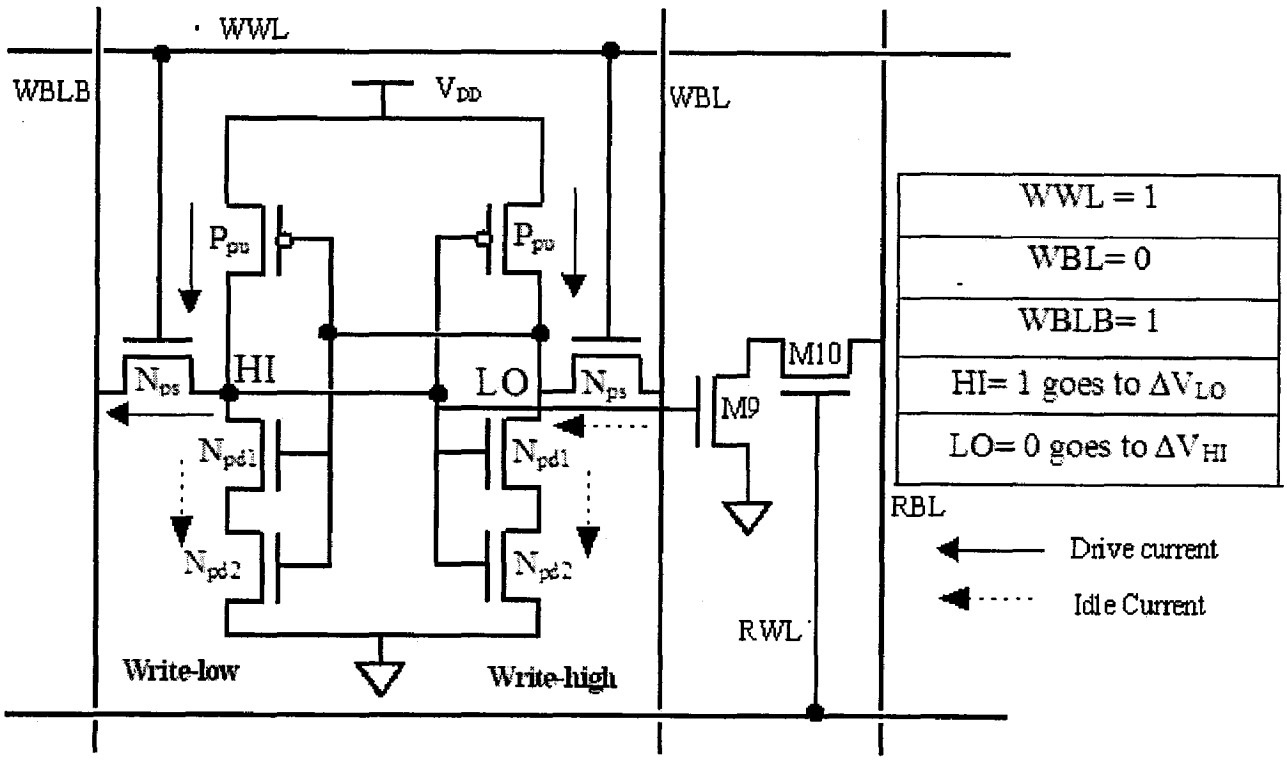


Figure.5.1 Write opteration in Sub-V_t region

The read upset condition does not arise in 10T case because of the buffer connected. As seen in the Fig.5.1 write condition is analyzed for two cases. The Write-Low case is when a '0' is written into HI, and the Write-High case is when a '1' is written into LO. The Write-Low case determines the minimum width for N_{ps} to pull HI down to ΔV_{LO} and is performed at the SF corner. In the Write-High case, the analysis places a constraint on the maximum width of N_{pd1} N_{pd2} and N_{ps}. Large idle current through N_{pd1} N_{pd2} and N_{ps} causes a voltage divider that

overpowers the drive current of P_{pu} used to pull LO up to ΔV_{HI}. It is taken that node HI is pulled down to $\Delta V_{LO} = 20\%* V_{DD}$ and node LO is pulled up to $\Delta V_{HI} = 80\%* V_{DD}$ and $P_{pu} = N_{ps}$.

The major difficulties with traditional 6T SRAM design under sub-threshold operation includes,

1. Exponential dependence of drive current on V_t in subthreshold region which is even badly effected by the process variations.
2. Reduced I_{ON}/I_{OFF} ratio in subthreshold operation compromises the robustness
3. The sensing voltage of sense amplifier does not scale below V_{DD}.

5.1 Writeback Scheme for Row Data Preservation

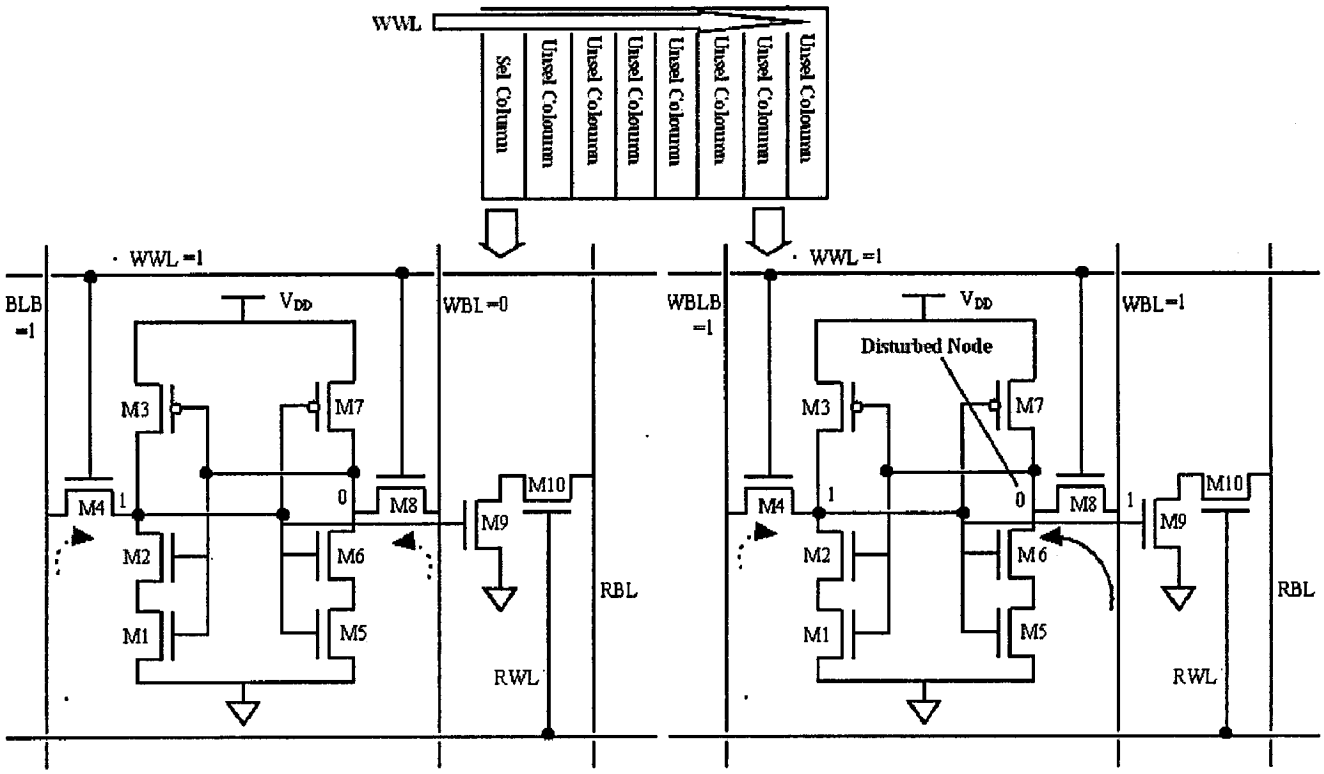


Figure.5.2 Shows the Pseudo Write in unselected column

In a column MUXed array, the write operation still has stability problems because the enabled write wordline is also shared by the unselected columns. This is also referred to as the pseudo-write (or pseudo-read) problem in conventional 6-T designs. Fig.5.2 illustrates this issue where the unselected cells can undergo a write when the WWL signal is asserted while the write

bitlines (WBL, WBLB) are precharged to V_{DD}. This is exactly the same condition as the worst case read stability in conventional 6-T SRAMs. [1]

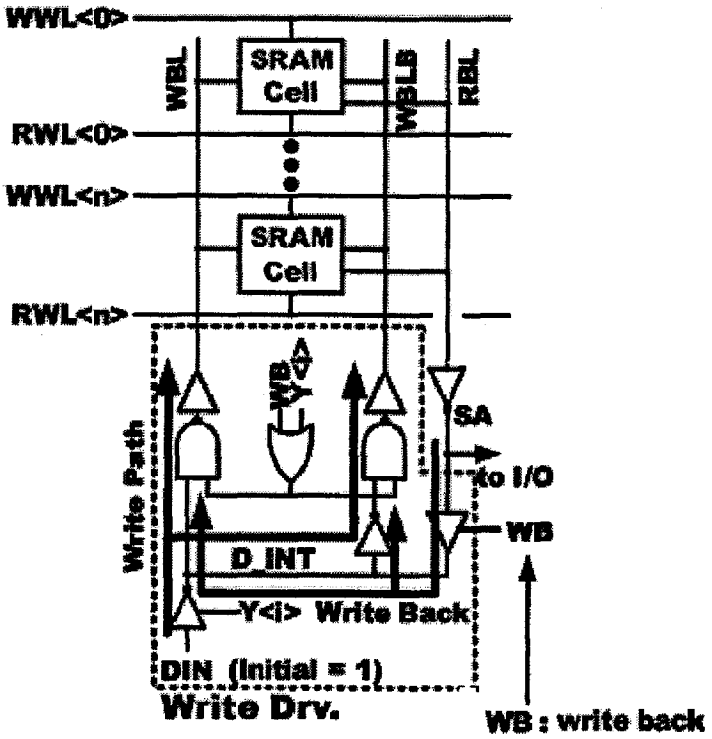


Figure.5.3 Write back scheme for preserving row data during write operation [1]

A writeback scheme shown in Fig.5.3 is applied to resolve the pseudo-write problem [11]. The write driver consists of a conventional write path and the writeback path. During write operation, read wordline (RWL) and write wordline (WWL) are enabled simultaneously. If the column is not selected for access Y(i)=0 , the write bitlines are kept to V_{DD} and read operation is executed. The writeback signal (WB) is enabled from the rising edge of RWL with additional delay enabling the writeback path and the read data from the read buffer is transferred to D_INT and written back to WBL and WBLB. By rewriting the read data back to WBL and WBLB, there is no voltage difference between write bitlines (WBL, WBLB) and the cell nodes, eliminating the contention current.

5.2 Floating Bitline and Back Gate Biasing of pMOS

During the standby mode, junction leakage contributes to array leakage on both the internal and bitline nodes when bitlines are pre-charged to V_{DD}. While junction leakage at internal node can

be reduced by lowering V_{DD}, the junction leakage at bitline node remains high with bitline held at V_{DD}. To reduce junction leakage, bitline is allowed to float during standby mode in this design. The subthreshold leakage of pMOS can become a significant contributor to the cell leakage when SRAM V_{DD} is scaled to very low level. pMOS back-gate bias is applied to reduce this leakage component.

5.3 Data-Dependent Bitline leakage

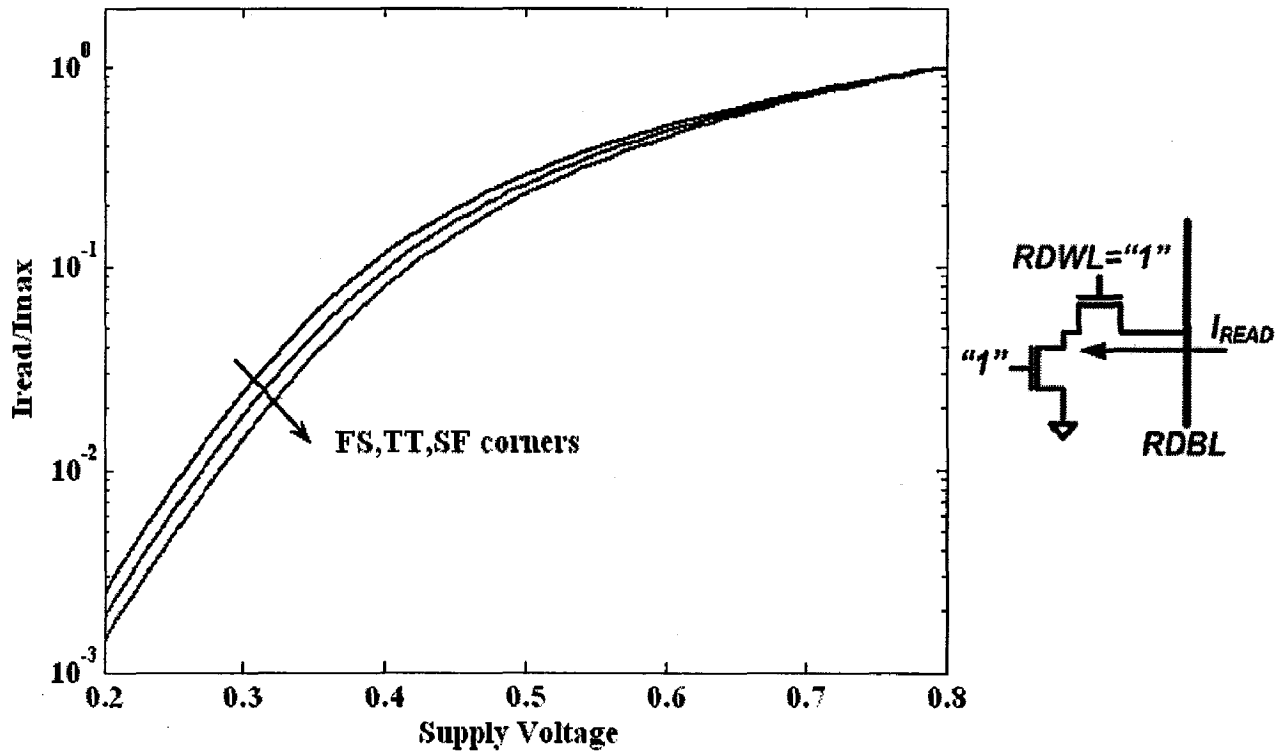


Figure 5.4 Normalized current under typical and slow case of nMOS transistor

The small I_{ON}-to- I_{OFF} ratio in the subthreshold region limits the number of cells per bitline and negatively impacts theSRAM density. As the number of cells in a bitline increases, bitline leakage from the unaccessed cells can rival the read current of the accessed cell making it difficult to distinguish between the bitline high and low levels. A significantly reduced read current is expected in sub-Vt due to the lower gate drive. However, the exponential impact of variation further degrades I_{READ}.As seen in the Fig.5.4 the typical case and the worse case with 3σ making the I_{READ} to go even less , this effect is particularly severe in sub-Vt, where the weak-cell read current can easily be a couple of orders of magnitude worse than the typical current. The combination of variation on top of drastically reduced mean read current implies that the read access time can extend almost arbitrarily. This is undesirable from a performance point of

view, but more importantly, it affects the ability to correctly sense data. Specifically, all of the unaccessed realization cells that share the read bit-line impose a leakage current that depends on their stored data i.e. data dependent leakage. In Fig.5.5, the aggregate leakage current is normalized by the maximum read current, assuming **128** cells per bit-line. As shown in Fig.5.6 and 5.7, the leakage can exceed the read signal, making the accessed data indecipherable. It is seen that due to process variations, at low supply voltages, variation in drain current is particularly problematic in the face of severely reduced I_{ON}/I_{OFF} ratio. Curves represent the I_{READ} to I_{LEAK} current when 64 bits are connected to the same bitline. Read current through accessed buffer at FS,SF and TT corners was normalized with respect to the worst case corner for I_{LEAK} i.e. FS. At FS corner the unaccessed read buffer line drives in maximum current. $I_{READ\ SF}/I_{LEAK\ FS}$ is most concerning case, as it has least value. To make improve upon I_{ON}/I_{OFF} ratio foot-driver is connected to foot of the buffer

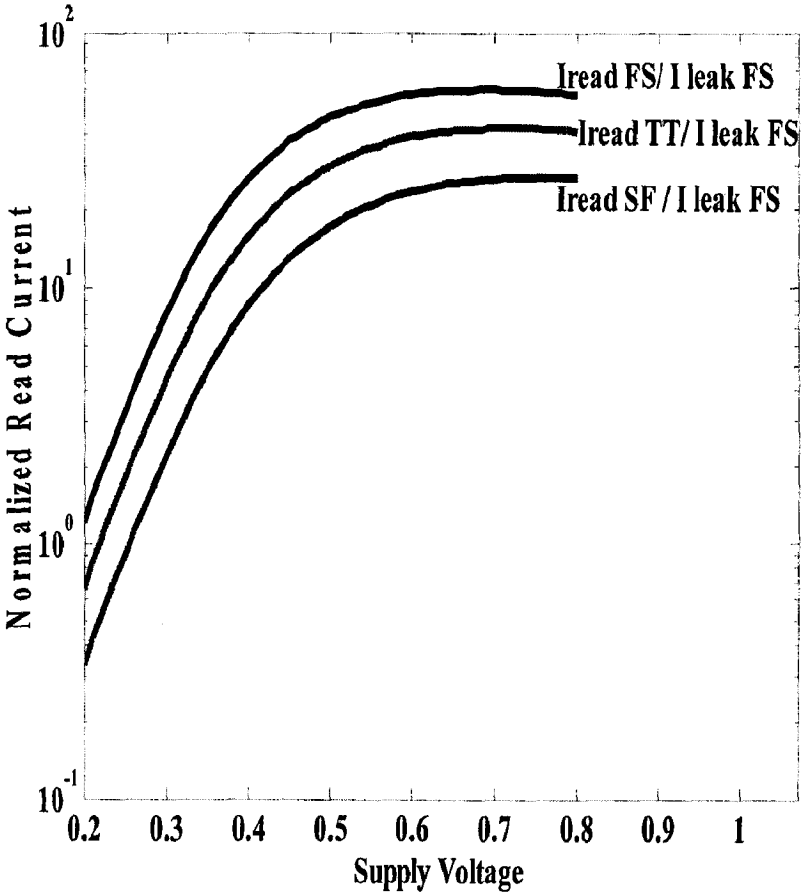


Figure.5.6 I_{ON}/I_{OFF} ratio and its degradation with process variation

To increase I_{READ} driver transistor width can be increased but it affects cell area as pass transistor width has to be increased to make write operation possible. Also standard deviation of V_t is inversely related to the square root of device areas [42], the variation-induced degradation of I_{READ} is greatly reduced in sub-V_t, where the dependence on V_t is exponential. So instead of connecting the buffer foot to the ground its connected to the read buffer as shown in Fig.5.8 The buffer foot of all the cells present in the same word is shorted; foot driver is a simple inverter. While read operation the the foot of the accessed word line is driven low, while other lines remain at supply voltage. Hence leakage path through read buffer of the unaccessed cell have no voltage drop across them, hence I_{ON} to I_{OFF} ratio increases as I_{OFF} decreases.

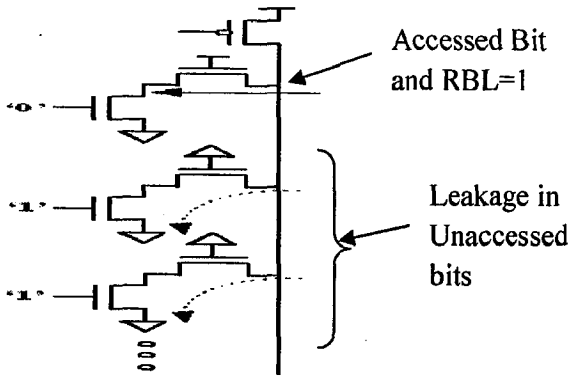


Figure.5.7 Worst case to determine the maximum number of transistor in the column

5.4 Write margin Improvement

To improve the write margin so that the SRAM works in Subthreshold region the techniques used are

- Word Line Boosting
- Reducing Supply Voltage of Cell During Write

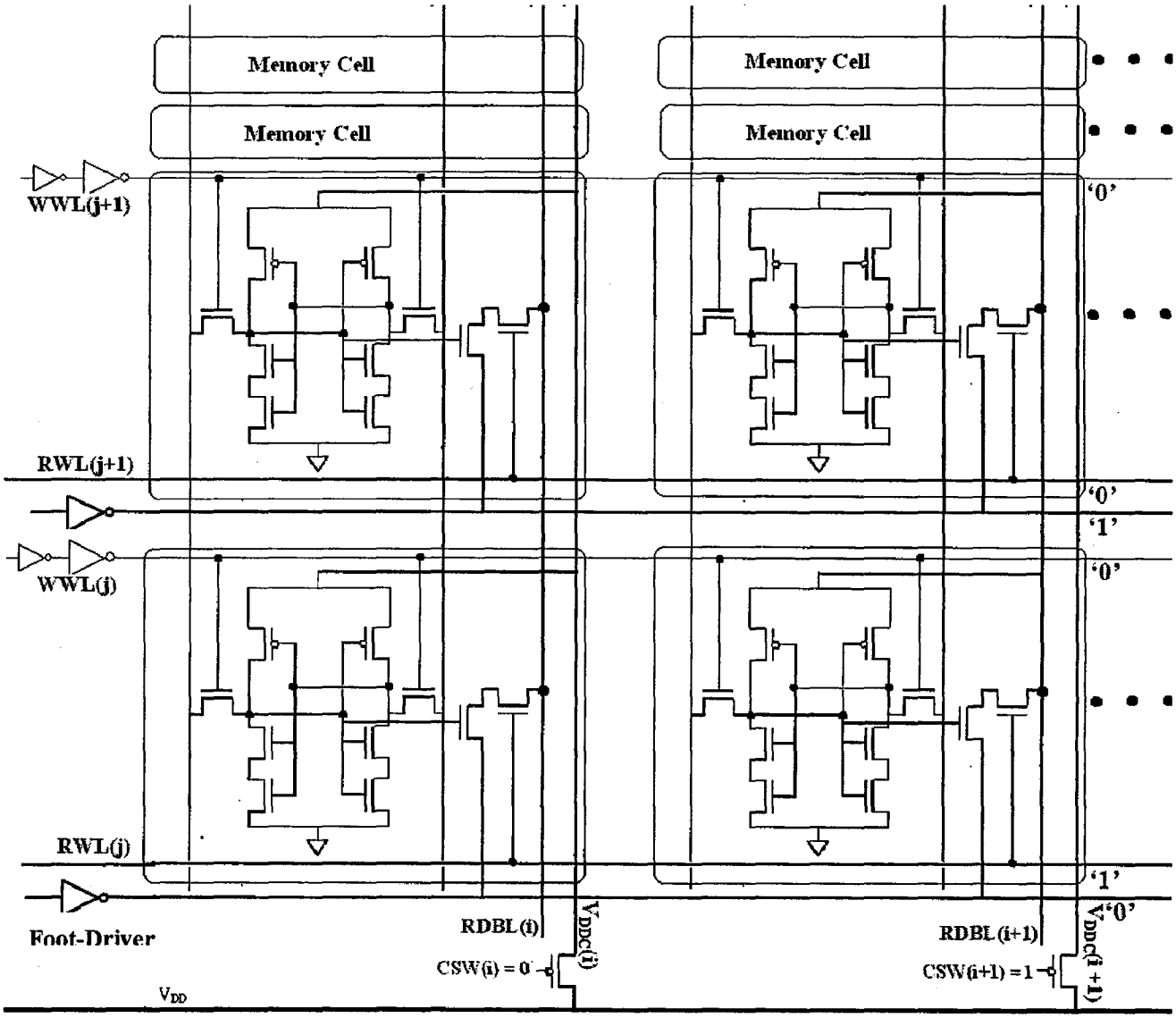


Figure.5.8 Shows controlled power supply, Foot Driver and Boosted Word line

During write the the cell is made monostable only at desired value. Then the local feedback generates it to the correct state. The write operation is ratioed since the nMOS pass transistor overpowers the pMOS load device in order to over write the data. So flipping of data will be easier if the pMOS is made weaker during write and its done by switching off the power supply of the column to be accessed. When column is accessed for write operation the control switch is turned off thus making the power supply to write accessed cell is floating. The supply voltage falls as write current flows through the node holding HI i.e. write low case in Fig.5.8, it further helps in improving write as pMOS becomes weaker during write time. Both the methods are implemented in Fig.5.8. Controlling the supply voltage of the cell to be written improves the

writing capability but degrades data retention in the cells whose word line is asserted. So to account for this floating voltage is kept on the column cells where only cell to be written is asserted. If the floating supply line is kept row wise then all cells in the row will become largely unstable by controlled supply voltage and WL activation and noise equal to RNM will be sufficient to flip the data in other cells of the row which are not to be written, and the data in these memory cells are easily destroyed. The voltage is controlled only in selected column and this maintains high retention characteristics in the memory cells that are not selected to be written and is equal to the SNM at operating voltage.

6.1 Implementation Results and Schematics

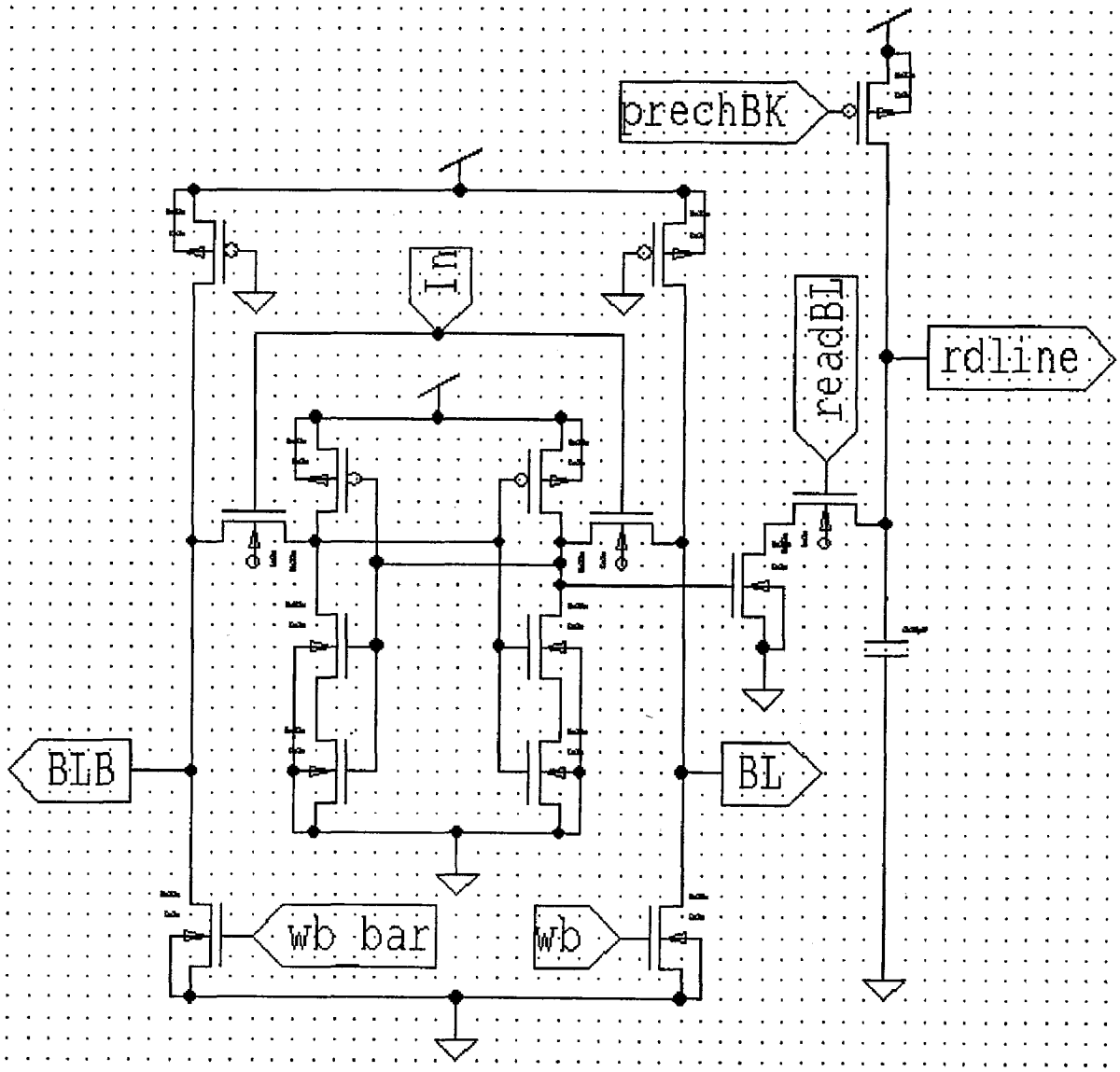


Figure.6.1 Schematic View for idle power, write operation and write power simulation ckt.

Power Results in idle mode

Following results are obtained when cell is in inactive mode i.e. data is written onto the cell and its left unaccessed. Hence in Idle mode there is leakage current flowing through the power supply. This power is calculated once the data is set on the nodes of the cell. It is observed that

Vdd=1.1v

vdd from time 4e-008 to 1e-007

Average power consumed -> 5.574488e-010 watts

Vdd=1v

vdd from time 4e-008 to 1e-007

Average power consumed -> 4.604574e-010 watts

Vdd=0.9v

vdd from time 4e-008 to 1e-007

Average power consumed -> 3.885223e-010 watts

Vdd=0.8v

vdd from time 4e-008 to 1e-007

Average power consumed -> 3.535041e-010 watts

Vdd=0.7v

vdd from time 4e-008 to 1e-007

Average power consumed -> 3.087842e-010 watts

Vdd=0.6v

vdd from time 6e-008 to 1e-007

Average power consumed -> 1.864991e-010 watts

Vdd=0.5v

vdd from time 1.1e-007 to 1.4e-007

Average power consumed -> 1.052504e-010 watts

Vdd=0.4v

vdd from time 4e-008 to 1e-007

Average power consumed -> 4.862380e-011 watts

Vdd=0.3v

vdd from time 4e-008 to 1e-007

Average power consumed -> 2.975719e-011 watts

Vdd=0.2v

vdd from time 1.4e-007 to 1.5e-007

Average power consumed -> 1.802838e-011 watts

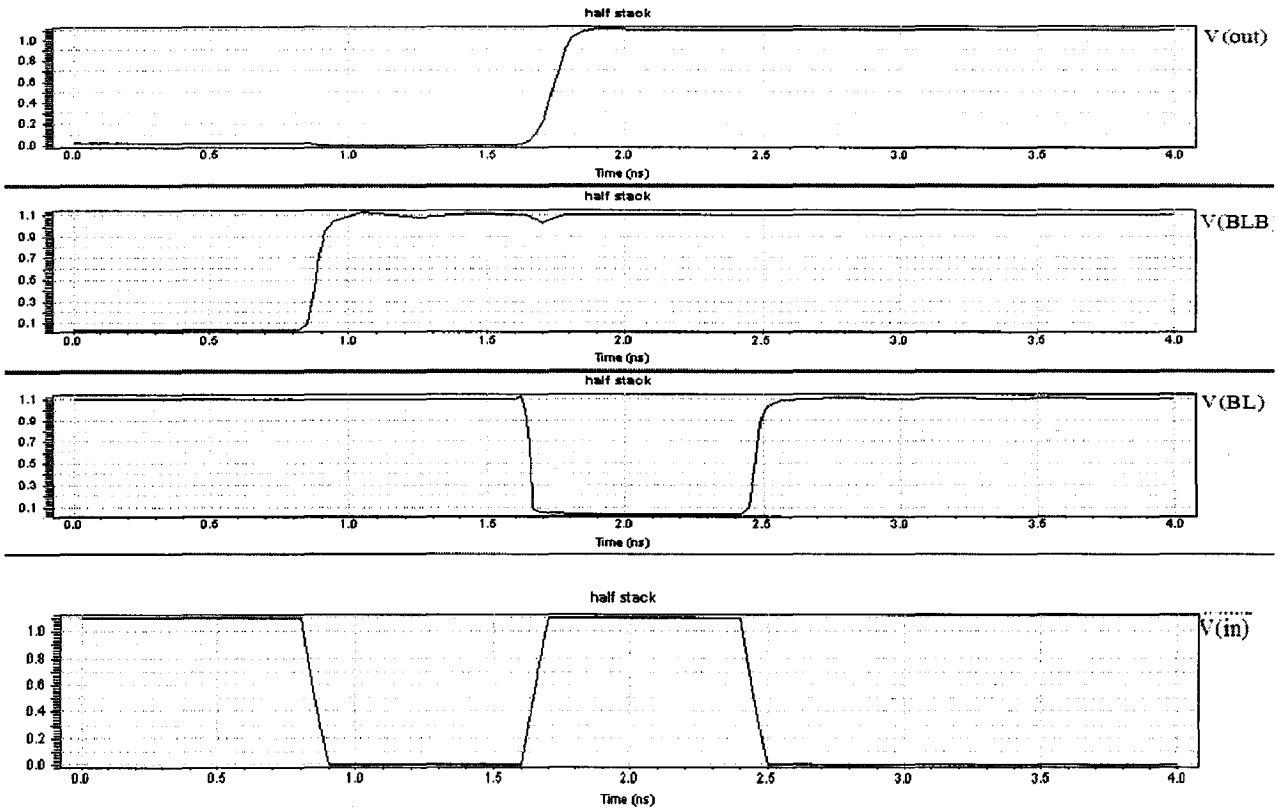


Figure.6.2 Write operation at 1.1V with writing time of 0.2 ns

Fig. 6.2 shows write operation. V(wb) is made high and V(wb_bar) is still kept low hence Bit line and Bitline bar develops '0' and '1' volts respectively. Now access transistors are made high. Immediately after this nodes of the cell starts to develop voltage. Then access transistors are made off and cell through regenerative action attains one of the desired bistable states. Access time was calculated as the time for which access transistors are kept on. Write power is average power drained out of the cell during transition. In Fig.6.3 it's observed that as voltage falls it take more time to write but power consumed during transition decreases. In Fig. 6.4 read sensing is done. ReadBL is made high sensing data. When enough voltage is build on ReadBL then Sense, Sense_Bar and Access signal are made high for sense amplifier to sense the data.

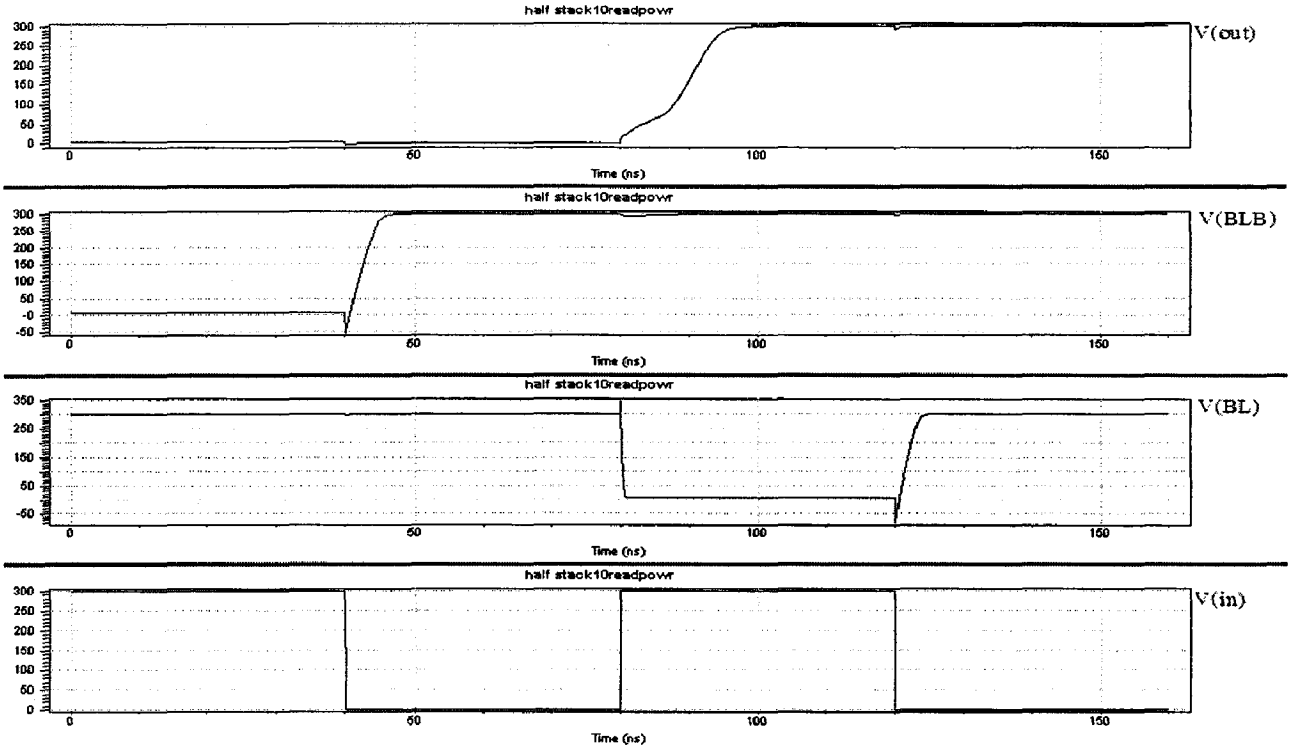


Figure.6.3 Write operating on at 300mV with writing time of 40 ns

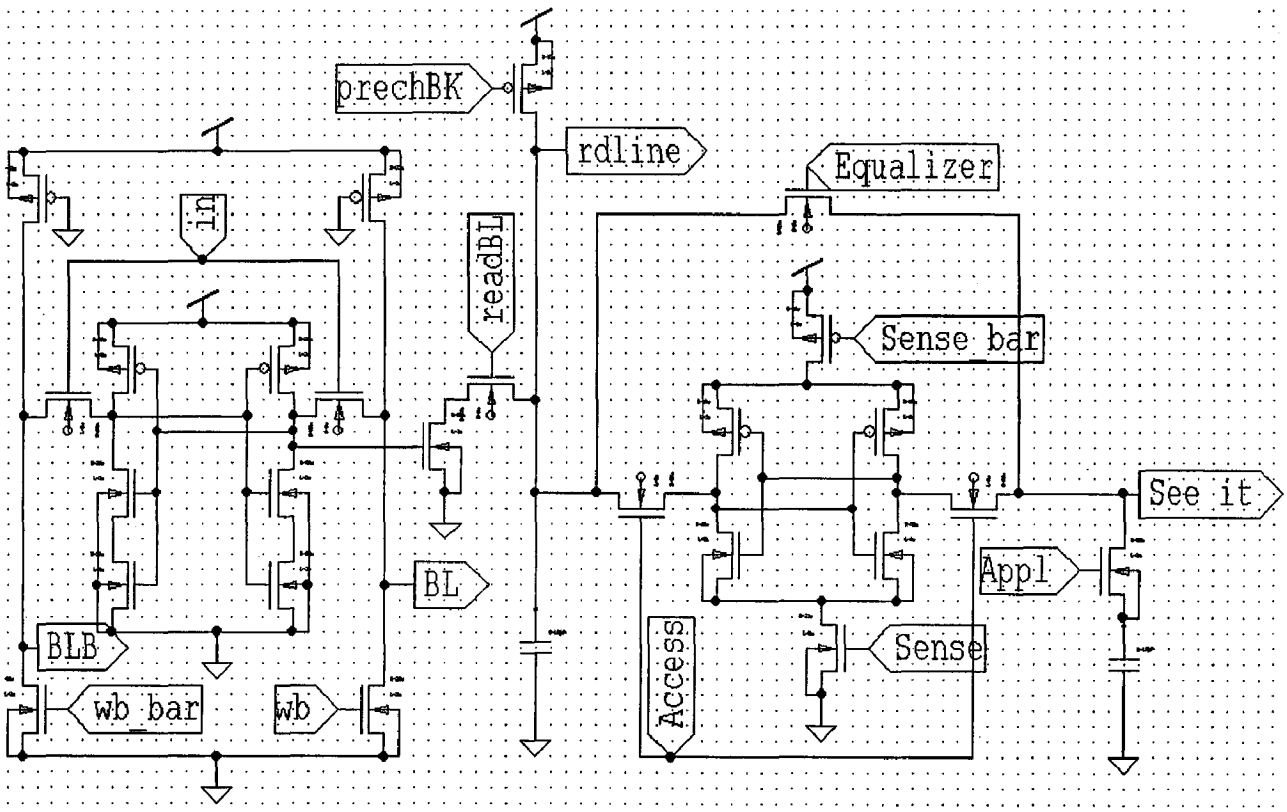


Figure.6.4 Schematic for read time and read power simulation ckt

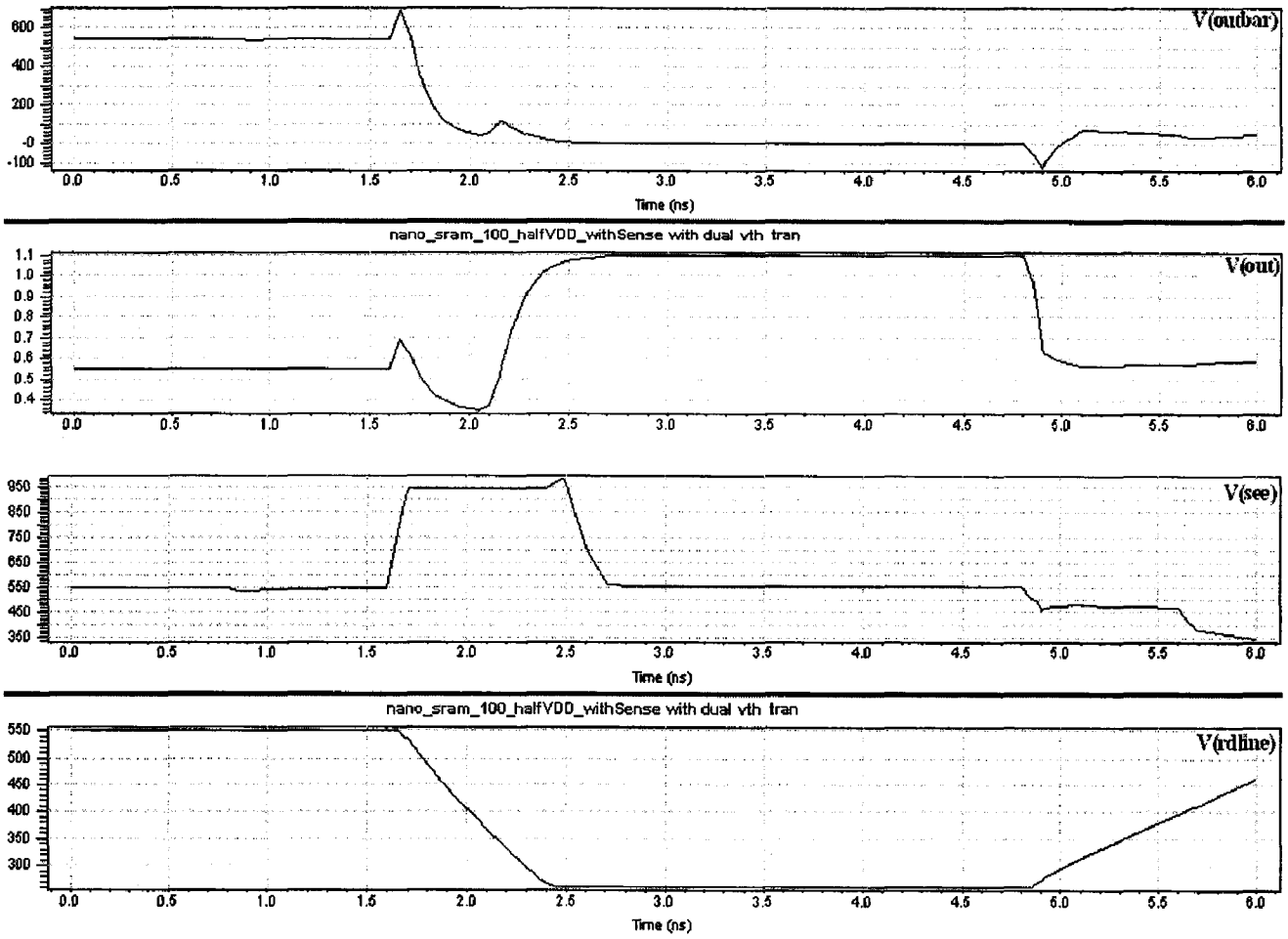


Figure.6.5 Read time and read power simulation waveforms V_{DD}=1.1

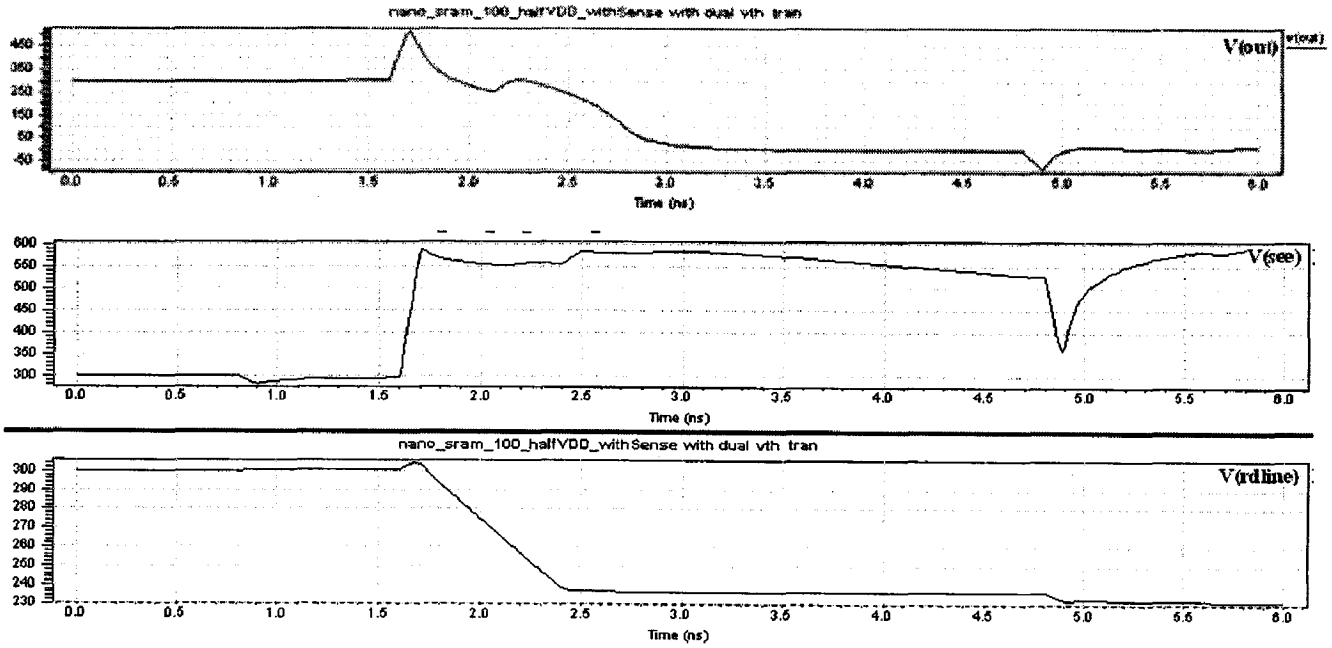


Figure.6.6 Read waveforms V_{DD}= 600 mV

Write power Results

V_{dd} =1.1 volts (1.25Ghz)

vdd from time 1.6e-009 to 2.4e-009

Average power consumed -> 4.450003e-006 watts

V_{dd}=1.0 volts (1.11Ghz)

vdd from time 1.8e-009 to 2.7e-009

Average power consumed -> 3.182249e-006 watts

V_{dd}=0.9volts (1Ghz)

vdd from time 2e-009 to 3e-009

Average power consumed -> 2.233774e-006 watts

V_{dd}= 0.8 volts (0.7142 GHz)

vdd from time 2.8e-009 to 4.2e-009

Average power consumed -> 1.239853e-006 watts

V_{dd}= 0.7 volts (500GHz)

vdd from time 4e-009 to 6e-009

Average power consumed -> 6.468889e-007 watts

V_{dd} 0.6 volts (400 MHz)

vdd from time 5e-009 to 7.5e-009

Average power consumed -> 3.725213e-007 watts

V_{dd} 0.5 volts (250 MHz)

vdd from time 8e-009 to 1.2e-008

Average power consumed -> 1.572425e-007 watts

V_{dd} 0.4 volts (125Mhz)

vdd from time 1.6e-008 to 2.4e-008

Average power consumed -> 5.099491e-008 watts

V_{dd}=0.3 (20Mhz)

vdd from time 1.6e-008 to 2.4e-008

Average power consumed -> 3.243847e-011 watts

6.2 Conclusion

New stacked 10T SRAM cell design has been proposed for ultra low power applications. The proposed design has enhanced performance in terms of power and SNM in sub threshold and above sub threshold regions. Besides, an enhanced static noise margin of 565mV has been obtained at a supply voltage of 1.1V. The proposed design is less susceptible to process variation as well. This ensures that minor errors in the fabrication process do not drastically affect the cell performance. Moreover, the stacked cell offers better noise margins with lesser power over in comparison to that of the full stacked cell at the cost of area over head. Thus, the proposed cell architecture offers, superior performance in comparison to earlier cell designs and can be implemented for ultra low-power applications. The cell has robust noise margin at voltages low as 400mV. Further reduction in voltage and subthreshold operations need state of art read/write assist circuits. Techniques like Floating Bitline, Word Line Boosting, Read Buffer, Floating Supply Voltage, Write back circuits were used. To reduce the capacitance of single ended Bitline open Bitline architecture was used. It improved the number of cells that can be connected to the same Bitline. Write noise margin was also improved under process variation. Tolerance to leakages under process variation was improved by stacking and Foot-Buffer. Optimized layout ensured least increase in area of per cell. Every effort is made to minimize the area of the memory cell. Extensive use is made of symmetry to allow the core array to be generated by simply "tiling" the cells together vertically and horizontally. Two levels of metal and one layer of poly are used to realize this memory cell. The large number of devices connected to the wordline and bitlines gives rise to large capacitance (and resistance) values. The row lines are routed in both Metal1 and poly to reduce resistance, while the bitlines are routed in Metal2. Removing the substrate connections may further reduce the cell area on the cost of stability. The access frequency was optimized by the use of cross-coupled sense amplifier and minimizing the time of access hence improving upon the speed of the circuit. Voltage difference of as low as 70mV was sensed by the sensing circuit.

6.3 Scope for Future Work

In this work the area was traded with low leakage current and optimized performance. Near to subthreshold region read/write operation were being performed. Scope lies in designing more suitable peripherals which can work in optimized way in sub-V_t region. Metal layers upto 8 can be used to make cell more compact and symmetric. Power could be saved further by bringing the subarray not in use to DRV. Impact of process variation on static characteristics was analyzed, Its impact on read/write access time could be analyzed and tolerant circuit could be proposed. Adding of error check codes like odd/even parity or bit interleaving may help to overcome soft errors and accidental data flip.

REFERENCES

1. Y. Wang *et al.*, "A 1.1 GHz 12 μA/Mb-Leakage SRAM Design in 65 nm Ultra-Low-Power CMOS Technology With Integrated Leakage Reduction for Mobile Applications," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 173–179, Jan. 2008.
2. International Technology Road Map for Semiconductors, [Online]. Available: http://public.itrs.net
3. Alice Wang, and Anantha Channdrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.
4. T. H. Kim, J. Liu, J. Kean, and C. H. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," *Proc. IEEE Int. Solid-State Circuits Conf., Digest of Tech. Papers*, pp. 330–331, Feb. 2007.
5. B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200 mV 6T SRAM in 0.13 μm CMOS," *Proc. IEEE Int. Solid-State Circuits Conf., Digest of Tech. Papers*, pp. 332–333, Feb. 2007.
6. N. Verma, and A. Chandrakasan, "A 65nm 8T Sub-V_t SRAM Employing Sense-Amplifier Redundancy," *Proc. IEEE Int. Solid-State Circuits Conf., Digest of Tech. Papers*, pp.328-329, Feb. 2007.
7. Ik Joon Chang, Jae-Joon Kim Sang Phill Park, and Kaushik Roy, "A 32kb 10T Subthreshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90nm CMOS," *Proc. IEEE Int. Solid-State Circuits Conf., Digest of Tech. Papers*, pp. 388–390, Feb. 2008.
8. Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto, "An area-conscious low-voltage-oriented 8T-SRAM design under DVS environment," *Proc. IEEE Symp. VLSI Circuits, Digest of Tech. Papers*, pp. 256–257, Jun. 2007.
9. B. Calhoun, and A. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," *Proc. IEEE Int. Solid-State Circuits, Conf., Digest of Tech. Papers*, pp. 628–629, Feb. 2006.

10. L. Chang, Y. Nakamura, R. K. Montoye, J. Sawada, A. K. Martin, K. Kinoshita, F. H. Gebara, K. B. Agarwal, D. J. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3 GHz 8T-SRAM with operation down to 0.41 V in 65 nm CMOS," *Proc. IEEE Symp. VLSI Circuits*, pp. 252–253, Jun. 2007.

11. M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De, "Wordline and bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs," *Proc. IEEE Symp. VLSI Circuits*, pp. 9–10, Jun. 2006.

12. Z. Guo, S. Balasubramanian, R. Zlatanovici, T. J. King, and B. Nikolic, "FinFET-based SRAM design," *Proc. of International Symposium on Low Power Electronics and Design*, pp. 2-7, Aug. 2005.

13. H. Kawaguchi, Y. Iataka, and T. Sakurai, "Dynamic Leakage Cut-off Scheme for Low-Voltage SRAM's," *Digest of Technical Papers, Symposium on VLSI Circuits*, pp.140-141, June 1998.

14. K. Zhang *et al.*, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE Journal of Solid-State Circuits*, vol. 40, issue 4, pp. 895-901, April 2005.

15. C. H. Kim, J. Kim, I. Chang, and K. Roy, "PVT-Aware leakage reduction for on-die caches with improved read stability", *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 170-178, Jan. 2006.

16. M. Khellah, D. Somasekhar, Y. Ye, N. S. Kim, J. Howard, G. Ruhl, M. Sunna, J. Tschanz, N. Borkar, F. Hamzaoglu, G. Pandya, A. Farhang, K. Zhang, and V. De, "A 256-Kb Dual-V_{CC} SRAM Building Block in 65-nm CMOS Process With Actively Clamped Sleep Transistor," *IEEE Journal of Solid-State Circuits*, vol. 42, issue 1, pp. 233-242, Jan. 2007.

17. K. Itoh, "Low Voltage Memories for Power-Aware Systems," *Proc. of International Symposium on Low Power Electronics and Design*, pp. 1- 6, Aug. 2002.

18. K. Itoh, A. R. Fridi, A. Bellaouar, and M. I. Elmasry, "A deep sub-V, single power-supply. SRAM cell with multi-Vt, boosted storage node and dynamic load," *Digest of Technical Papers, Symposium on VLSI Circuits*, pp. 132–133, June 1996.

19. A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De," Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual V_t CMOS ICs," *Proceedings of International Symposium on Low Power Electronics and Design (ISLPED)*, Huntington Beach, CA , pp. 207–212, August 2001.

20. K. Flautner et al, "Drowsy caches: simple techniques for reducing leakage power," *Proc. of International Symposium on Computer Architecture*, pp. 148-157, May 2002.

21. J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Intergrated Circuits: A Design Perspective*, 2nd ed. Englewood Cliffs, NJ:Prentice- Hall, 2003.

22. A.S. Sedra, and K.C. Smith, *Microelectronic Circuits*, 4th ed. Oxford Unversity Press, 2003.

23. D.A. Hodges, *Analysis and Design of Digital Integrated Circuits*, 3rd ed. McGraw-Hill, 2004

24. Amit Agarwal, Chris H. Kim, Saibal Mukhopadhyay and Kaushik Roy, "Leakage in Nano-Scale Technologies: Mechanisms, Impact and Design Considerations", *Proc. of Design Automation Conference*, pp. 6-11, June 7-11, 2004.

25. W. Hung, Y. Xie, N. Vijaykrishnan, M. Kandemir, M.J. Irwin and Y. Tsai, "Total Power Optimization through Simultaneously Multiple-V_{DD} Multiple-V_{TH} Assignment and Device Sizing with Stack Forcing", *ISLPED '04*, pp.144-149, August 9-11, 2004.

26. Saibal Mukhopadhyay, Cassondra Neau, Riza Tamer Cakici, Amit Agarwal, Chris H. Kim and Kaushik Roy, " Gate Leakage Reduction for Scaled Devices Using Transistor Stacking", *IEEE Trans.on Very Large Scale Integr. (VLSI) Systems*, vol. 11, no. 4, pp. 716-730,August 2003.

27. Siva Narendra, Shekhar Borkar, Vivek De, Dimitri Antoniadis, and Anantha Chandrakasan, "Scaling of Stack Effect and its Application for Leakage Reduction," *Proc. of ISLPED '01*, pp. 195-200, August 6-7, 2001.

28. A. Agarwal, S. Mukhopadhyay, C.H .Kim, A. Raychowdhury and K. Roy, "Leakage power analysis and reduction: models, estimation and tools", *IEEE Proc. Computer and Digital Techniques*, Vol. 152, No. 3, pp. 353-368, May 2005.

29. E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. SC-22, No. 5, pp. 748-754, Oct. 1987.

30. J. Lohstroh, E. Seevinck, and J.D. Groot, "Worst-Case Static Noise Margin Criteria for Logic Circuits and Their Mathematical Equivalence," *IEEE Journal of Solid-State Circuits*, vol.18, no. 6, pp. 803-807, Dec 1983.

31. E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. SC-22, No. 5, pp. 748-754, Oct. 1987.

32. C. H. Kim, J. Kim, I. Chang, and K. Roy, "PVT-Aware leakage reduction for on-die caches with improved read stability", *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 170-178, Jan. 2006.

33. H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J.Rabaey, "Standby supply voltage minimization for deep sub-micron SRAM", *IEEE Microelectronics Journal*, vol. 36, pp. 789-800, Aug 2005.

34. Huifang Qin, Animesh Kumar,Kannan Ramchandran, Jan Rabaey, and Prakash Ishawar, "Error-Tolerant SRAM Design for Ultra-Low Power Standby Operation", *Proc. of 9th International Symposium on Quality Electronic Design, ISQED 2008*, pp. 30-34, March 2008.

35. Kanak Agarwal, and Sani Nassif, "Statistical Analysis of SRAM Cell Stability", *Proc of IEEE Design Automation Conference 2006*, pp. 57-62, July 24-28, 2006.

36. Benton H. Calhoun, and Anantha P. Chandrakasan, " Static Noise Margin Variation for Sub-threshold SRAM in 65-nm CMOS", in *IEEE J. Solid-State Circuits*, vol. 41, no.7, pp. 1673-1679, July 2006.

37. W.Kong, R.Venkatraman, R.Castagnetti, F.Duan and S.Ramesh, "High-Density and High-Performance 6T-SRAM for System-on-Chip in 130nm CMOS Technology", *Symp. VLSI Technology, Digest of Tech. Papers*, pp.105-106, June 2001.

38. Ramnath Venkatraman, Ruggero Castagnetti,, Olga Kobozeva, Franklin L. Duan, Arvind Kamath, S. T. Sabbagh, Miguel A. Vilchis-Cruz, Jhon Jhy Liaw, Jyh-Cheng You, and Subramanian Ramesh, "The Design, Analysis, and Development of Highly Manufacturable 6-T SRAM Bitcells for SoC Applications", *IEEE Trans. on Electron Devices*, vol. 52, no. 2,pp. 218-226, February 2005.

39. Chris Hyung-il Kim, Jae-Joon Kim, Saibal Mukhopadhyay, and Kaushik Roy, "A Forward Body-Biased Low-Leakage SRAM Cache: Device, Circuit and Architecture

Considerations", *IEEE Trans. on VLSI Systems*, Vol. 13, No. 3, pp. 349-357, March 2005.

40. Naveen Verma, Joyce Kwong, , and Anantha P. Chandrakasan, "Nanometer MOSFET Variation in Minimum Energy Subthreshold Circuits", *IEEE Trans. on Electron Devices*, vol. 55,No. 1,pp. 163-174, , January 2008.
41. B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", *Proc. Int. Symp. Low Power Electronic and Des.*, pp. 20–25, Aug.2005.
42. M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors", *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

G 14334