ANALYSIS AND DESIGN OF A DRAM CELL FOR LOW LEAKAGE

Thesis submitted in partial fulfillment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

VLSI Design & CAD

By

Rashmi Singh

Roll No. 60761016

Under the supervision of

Mr. Arun Kumar Chatterjee

Lecturer, ECED

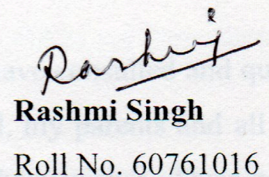
Department of Electronics and Communication Engineering
Thapar University, Patiala-147004, India

June, 2009

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Analysis and Design of a DRAM Cell for Low Leakage"** in partial fulfilment of the requirements for the award of the degree of Master of Technology in VLSI Design and CAD at the Electronics and Communication Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Mr. Arun Kumar Chatterjee, Lecturer, ECED.

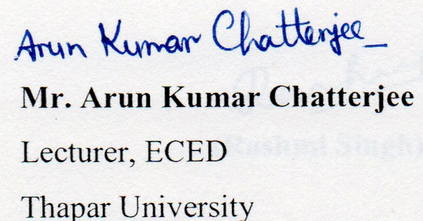
The matter presented in this thesis has not been submitted in any other University/Institute for the award of any degree.



Date: 15/07/09

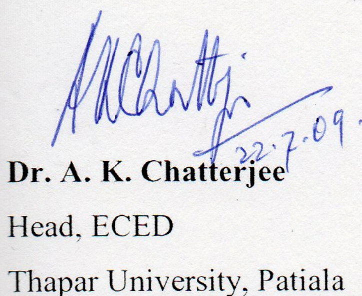
Rashmi Singh
Roll No. 60761016

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

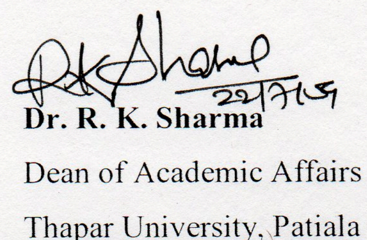


Mr. Arun Kumar Chatterjee
Lecturer, ECED
Thapar University

Counter Signed by:



Dr. A. K. Chatterjee
Head, ECED
Thapar University, Patiala



Dr. R. K. Sharma
Dean of Academic Affairs
Thapar University, Patiala

// ACKNOWLWDGEMENT

A thesis cannot be completed without the help of many people who contributed directly or indirectly through their constructive criticism in the evolution and preparation of this work. A special debt of gratitude is owed to my thesis supervisor, **Mr. Arun Kumar Chatterjee (Lecturer, ECED)** for his gracious efforts and keen pursuit which has remained as a valuable asset for the successful fulfillment of my thesis. His dynamism and diligent enthusiasm have been highly instrumental in keeping my spirits high. His flawless & forth right suggestions blended with an innate intelligent application have crowned my task with success.

I would also like to offer my sincere thanks to all faculty, teaching and non-teaching, of Electronics & Communication Engineering Department (ECED) and staff of central library, TU, Patiala for their assistance.

I am also thankful to the authors whose works I have consulted and quoted in this work. Last, but not the least, very special thanks to God, my parents and all my friends especially Rohit for their constant encouragement and best wishes. Their patience and understanding without which this study would not have been in this present form, is greatly appreciated.

(Rashmi Singh)

ABSTRACT

In this thesis, the analysis and design of Dynamic Random Access Memory (DRAM) cell for low leakage have been done. For the analysis, trench capacitor DRAM cell has been considered. For the design of trench capacitor DRAM cell, 0.18 µm submicron nMOSFET as access transistor and the conventional trench capacitor as storage device have been considered.

Various DRAM cell structures such as three-transistor memory cell, one-transistor one-capacitor memory cell, capacitorless one-transistor floating body cell and two-transistor floating body cell have been explored.

Various leakage mechanisms, such as subthreshold leakage, Gate-induced drain leakage, punchthrough and PN junction reverse bias leakage for an access transistor have been focused. Process-level techniques for leakage reduction, such as light doped drain, retrograde well and halo doping, in the trench capacitor DRAM cell have been discussed and employed.

Process simulation and device simulation of trench capacitor DRAM cell have been carried out using the ATHENA/ATLAS packages of SILVACO. The trench capacitor DRAM cell structure has been generated using ATHENA and then it is used as input to device simulator ATLAS for its electrical characterization.

Leakage reduction from 256 nA to 2.25 nA in DRAM cell has been achieved using the above mentioned process-level techniques.

CONTENTS

ACKNOWLEDGEMENT i

ABSTRACT ii

CONTENTS iii

LIST OF FIGURES v

LIST OF TABLES vii

CHAPTER

1. INTRODUCTION 1

2. LITERATURE SURVEY 2

2.1. Basic Memory Architecture 2

2.1.1. Memory Cell Array 2
2.1.2. The Peripheral Circuitry 4
2.1.3. The Input/Output (I/O) Interface Circuit 4

2.2. DRAM Cell Designs 4

2.2.1. Three-Transistor (3T) Memory Cell 4
2.2.2. One-Transistor One-Capacitor (1T/1C) Memory Cell 5
2.2.2.1. Three Dimensional Cell Structures 7
2.2.2.1.1. Stacked Capacitor Cell 8
2.2.2.1.2. Trench Capacitor Cell 8
2.2.3. Capacitorless One-Transistor Floating Body Cell (1T FBC) 10
2.2.3.1. Z-RAM 12
2.2.4. Two-Transistor Floating Body Cell (2T FBC) 14

3. LEAKAGE MECHANISMS AND REDUCTION TECHNIQUES IN TRENCH CAPCITOR DRAM CELL 16

3.1. DRAM Access Transistor Leakage Mechanisms 17

3.1.1. PN Junction Reverse Bias Current 18
3.1.1.1. Band-to-Band Tunneling Current 18
3.1.2. Subthreshold Leakage 19
3.1.2.1. Drain-Induced Barrier Lowering 21
3.1.2.2. Body Effect 22
3.1.2.3. Narrow Width Effect 22

3.1.2.4. Effect of Channel Length and V_T-Rolloff 24
3.1.2.5. Effect of Temperature 24
3.1.3. Gate Induced Drain Leakage 24
3.1.4. Punchthrough 26
3.1.5. Tunneling into and through Gate Oxide 27
3.1.6. Injection of Hot Carriers from Substrate to Gate Oxide 27

3.2. Process level techniques for leakage reduction in trench capacitor DRAM cell 28

3.2.1. Channel Engineering for Leakage Reduction 28

3.2.1.1. Retrograde Doping 30
3.2.1.2. Halo doping 30

4. PROCESS STEPS FOR FABRICATION OF TRENCH CAPACITOR DRAM CELL 33

5. RESULTS AND CONCLUSION

REFERENCES

LIST OF FIGURES

Figure 1.1: Trends in the memory cell size and memory capacity of DRAMs…… 1
Figure 2.1: Array-structured memory organization………………………………. 3
Figure 2.2: Three-transistor dynamic memory cell……………………………... 5
Figure 2.3: One-transistor dynamic memory cell……………………………….. 6
Figure 2.4: Cross-sectional view of a conventional planar capacitor DRAM cell… 7
Figure 2.5: Cross-sectional view of a stacked capacitor DRAM cell……………… 8
Figure 2.6: Cross-sectional view of a trench capacitor DRAM cell……………….. 9
Figure 2.7: Cross-sectional view of an SOI nMOSFET…………………………… 10
Figure 2.8: The principle of the FBC during write '1' and '0' and read……………. 12
Figure 2.9: Intrinsic bipolar transistor in SOI nMOSFET………………………… 13
Figure 2.10: The basic 2T floating body cell in a DRAM array……………………. 14
Figure 3.1: Leakage paths in a trench capacitor DRAM cell……………………… 17
Figure 3.2: Leakage current mechanisms in deep-submicrometer transistors……. 18
Figure 3.3: N-channel I_D vs. V_G curve showing DIBL, GIDL, weak inversion and pn junction reverse-bias leakage components………………………… 21
Figure 3.4: Condition of the depletion region near the drain-gate overlap region of an MOS transistor when (a) surface is accumulated with low negative gate bias; (b) n^+ region is depleted or inverted with high negative bias 25
Figure 3.5: Graphical representation of different aspects of well engineering……. 29
Figure 3.6: Short channel threshold voltage rolloff for retrograde and superhalo… 31
Figure 4.1: Initial silicon substrate………………………………………………… 34
Figure 4.2: Node dielectric deposited on the sidewalls and bottom of trench…….. 35
Figure 4.3: Recess of first polysilicon fill…………………………………………. 36
Figure 4.4: Formation of oxide collar……………………………………………… 36
Figure 4.5: Recess of second polysilicon fill………………………………………. 37
Figure 4.6: Formation of trench capacitor and shallow trench isolation………….. 37
Figure 4.7: Formation of retrograde well in the substrate………………………… 38
Figure 4.8: Doping profile of substrate after V_{th} adjust implant………………….. 40
Figure 4.9: Boron doping profile after halo implant………………………………. 41
Figure 4.10: Net doping profile after annealing…………………………………….. 42

Figure 4.11: The complete structure of trench capacitor DRAM cell………………. 42

Figure 5.1: Subthreshold characteristics of DRAM cell with LDD (without retrograde well and halo implantation)……………………… 44

Figure 5.2: Subthreshold characteristics of DRAM cell with retrograde well and LDD (without halo implantation)…………………………………….. 44

Figure 5.3: Subthreshold characteristics of DRAM cell with retrograde well, LDD and halo implantation…………………………………………….. 45

LIST OF TABLES

Table 4.1: Physical parameters for the design of 0.18 μm nMOSFET………….... 33

Table 5.1: Leakage current in trench capacitor DRAM cell after employing process-level techniques………………………………………………. 50

CHAPTER 1

INTRODUCTION

Semiconductor memories are the most vital microelectronic component of digital logic system design, such as computers and microprocessor-based applications ranging from satellites to consumer electronics. Therefore, continuous advancements in the fabrication of semiconductor memories including process enhancements and technology developments through the scaling of devices are being made for higher efficiency, higher density and faster speeds.

Dynamic Random Access Memory (DRAM) is the most common and cost efficient random access memory to be used as main memory component in most computers, workstations and many electronic systems. DRAMs provide very high-density storage. The term “dynamic” is used because the leakage phenomenon inherent in DRAM degrades the charge stored in the memory cell with time and makes the periodic refreshment necessary in order to retain the data. The first DRAM was proposed in 1970 with a capacity of 1 Kb. Since then, DRAMs have been the major driving force behind VLSI technology development. The DRAM capacity has increased six times in last three decades - from the 1-Kb level in 1970 to the 1- to 4-Gb level today, as shown in Figure (1.1) [1] – [2].

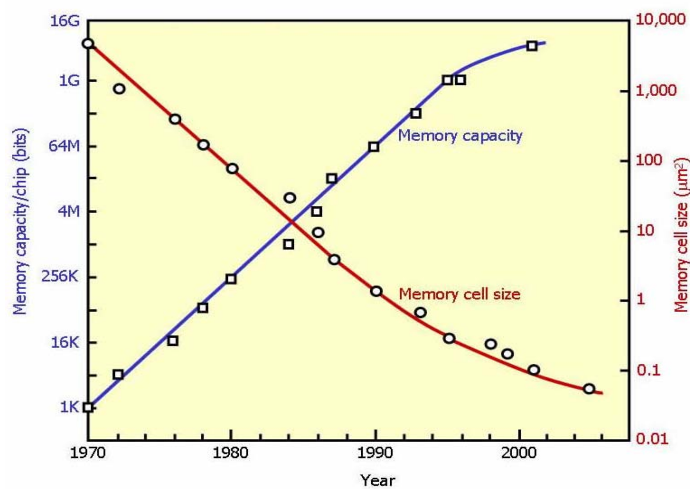


Figure 1.1: Trends in the memory cell size and memory capacity of DRAMs [1] – [2].

CHAPTER 2

LITERATURE SURVEY

The trends in DRAM cell structures since its evolution have been explored and presented in this chapter. This chapter reviews the DRAM cell structures such as: Three-Transistor memory cell, One-Transistor One-Capacitor memory cell, Capacitorless One-Transistor floating body cell and Two-Transistor floating body cell. The characteristics, read/write operations and limitations of all these cell designs have been described.

2.1. Basic Memory Architecture

A memory chip is composed of three blocks: a memory cell array, a peripheral circuit, and an input/output (I/O) interface circuit [3]. A memory cell array comprising a matrix of 2^{X} rows and 2^{Y} columns can store binary information of 2^{X+Y} bits. In random access memory, any cell can be accessed at random with the same speed irrespective of its location by selecting both the corresponding row and column.

2.1.1. Memory Cell Array

The basic array structured memory organization is shown in figure (2.1). In the memory organization, there are memory cell arrays, sense amplifiers and row and column decoders. In the memory, any word can be selected for reading or writing by providing the address word. The address word is partitioned into a column address (A_0 to A_{K-1}) and a row address (A_K to A_{L-1}). The row address enables one row of the memory for R/W, while the column address picks one particular word from the selected row. Multiple words are stored in single row and are selected simultaneously. To route the correct word to the input/output (I/O) terminals, an extra circuitry called the column decoder is needed. The concept is shown in figure (2.1). The horizontal select line that enables a single row of cells is called the word line, while the wire that connects the cells in a single column to the I/O circuitry is called the bit line.

When a memory chip is designed, the memory aspect ratio should be taken into consideration. The distortion in aspect ratio results in a design that can not be implemented. Besides the bizarre shape factor, the resulting design is also extremely

slow. To solve this problem, memory arrays are organized so that the aspect ratio approaches unity.

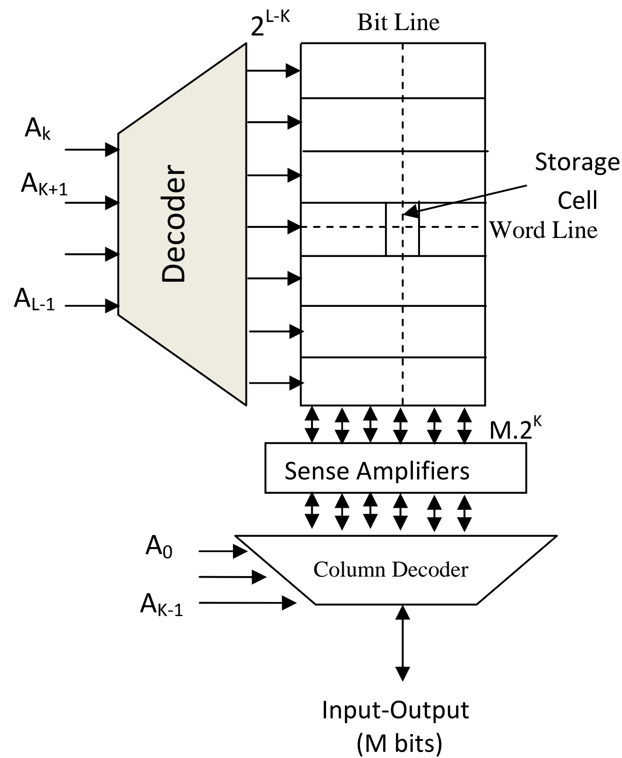


Figure 2.1:- Array-structured memory organization [3].

The area of large memory modules is dominated by the size of the memory core. Thus, it is crucial to keep the size of the basic storage cell as small as possible. The architecture of figure (2.1) works well for memories up to a range of 64 Kbits to 256 Kbits. Larger memories suffer from serious speed degradation as the length, capacitance and the resistance of the word and bit lines become excessively large. In larger memories, one extra dimension is added to the address space. The memory is partitioned into P smaller blocks. The composition of each of the individual blocks is identical to the one shown in the figure (2.1). A word is selected on the basis of row and column addresses that are broadcast to all the blocks. An extra address word called the block address, selects one of the P blocks to be read or written. This approach has a dual advantage.

1. Length of local word and bit lines is kept within bound resulting in faster access time.

2. The block address can be used to activate only the addressed block. Non-active blocks are put in power saving mode resulting in a substantial power saving.

2.1.2. The Peripheral Circuitry

The robustness, performance and power consumption of the memory is affected by the memory peripheral circuitry. This block bridges between the memory array and the I/O interface so that they can communicate with each other. It sends write data to a memory cell in the memory array under the control of the interface circuit, or sends read data from the memory cell to the interface circuit. The peripheral circuitry are address decoders, sense amplifiers, voltage references, drivers/buffers and timing and control circuit.

2.1.3. The Input/Output (I/O) Interface Circuit

This converts external signals, such as addresses, clocks and control signal to the corresponding internal signals that activate the peripheral circuit. It outputs read data from the array as the data output of the chip. Address buffers in the I/O interface circuit generate N and Msets of complementary row and column address signals, respectively, using $(N + M)$ external address signals. Data input and output buffers, a write control buffer, and control clock circuits are also components of the I/O interface circuit.

2.2. DRAM Cell Designs

The DRAM cell array usually covers 50 - 60% of the total chip area. For high density DRAMs, decreasing the cell size is the most efficient way to reduce the DRAM chip size. In early CMOS DRAM storage cell design, three-transistor and four-transistor cells were used in 1-Kb and 4-Kb generations. Later, one-transistor cell, providing smaller cell size and low cost, became the industry standard. In further attempts to scaling of cell size, capacitorless single transistor cell was proposed. Continuous investigations for scaling the device while maintaining its characteristics and meeting the design criteria are being done. This section describes all these cell structures.

2.2.1. Three-Transistor (3T) Memory Cell

The three-transistor cell, shown in figure (2.2) formed the core of the first popular MOS semiconductor memories such as the first 1-Kb memory from Intel.

Read/Write Operation

The cell is written by placing the appropriate data value on BL1 and asserting the write -

word line (WWL). The data is retained as charge on capacitance C_s (node X) once WWL is lowered. To read the data, the read-word line (RWL) is raised. The storage transistor (M_2) is either on or off depending upon the stored value. The BL2 is precharged to either V_{DD} or $V_{DD} - V_T$. The series connection of M_2 and M_3 pulls BL2 low when a '1' is stored. BL2 remains high when a '0' is stored. The inverse value of the stored signal is sensed on the bit line. The stored data gets deteriorated due to leakage through transistors. To refresh the cell the inverse of read data is put on BL1, and WWL is asserted in consecutive order.

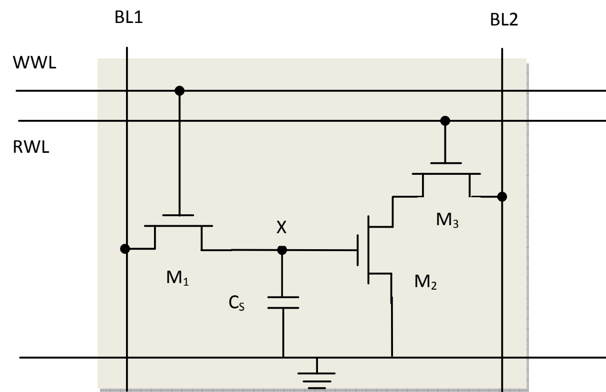


Figure 2.2:- Three-transistor dynamic memory cell [3].

Properties of the 3T cell are as follows [3]:

1. Unlike SRAM cell, there is no constraint on the device ratios. The choice of device sizes is solely based on performance and reliability considerations.
2. Reading the 3T cell contents is nondestructive.
3. No special process steps are needed. The storage capacitance is nothing more than the gate capacitance of the readout device.
4. There is a threshold voltage loss when writing a '1' at the storage node; this reduces the current flowing through M_2 during a read operation and increases the read access time. Bootstrapping the word line voltage is the solution.

2.2.2. One-Transistor One-Capacitor (1T/1C) Memory Cell

The schematic for one-transistor one-capacitor DRAM cell is shown in figure (2.3). It consists of a MOSFET (also referred to as array-access transistor or transfer device) in series with a storage capacitor. The word line (WL) contacts the gate of the access

transistor, and the bit line (BL) contacts the source/drain of the access transistor that is not connected to the storage capacitor. The bit line has a capacitance C_{BL}, including the parasitic load of connected circuits.

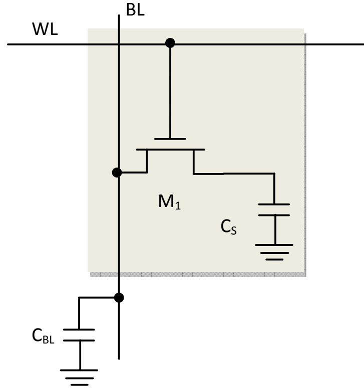


Figure 2.3:- One-transistor dynamic memory cell [3].

Read/Write Operation

To write the data, the data value is placed on bit line BL and the transfer device is turned on by raising the word line. Depending on the data, the cell capacitance is either charged or discharged. Data is stored by turning off the transfer device by lowering the word line, trapping the voltage/charge on the storage capacitor. Before the read operation, the bit line is precharged to a voltage V_{PRE}, midway between the high and low levels. The transfer device is then turned on and then the charge redistribution takes place between the bit line and storage capacitance. This results in a voltage change on the bit line, the direction of which determines the value of the data stored. The magnitude of the swing is given by the expression [3],

$$\Delta V = V_{BL} - V_{PRE} = (V_{BIT} - V_{PRE})\frac{C_S}{C_S + C_{BL}} \quad \text{……(1)}$$

where C_{BL} is the bit line capacitance, V_{BL} is the potential of the bit line after the charge redistribution, and the V_{BIT} the initial voltage over the cell capacitance C_S. As C_S is normally one or two orders of magnitude smaller than C_{BL}, the voltage change is very small, typically around 250 mV for state-of-the-art memories. The ratio C_S=C_S/(C_S + C_{BL}) is called the charge-transfer ratio and ranges between 1% and 10%. Amplification of ΔV to the full voltage swing is necessary if functionality is to be achieved. The cross-

sectional view of conventional planar capacitor DRAM cell is shown in the figure (2.4) given below.

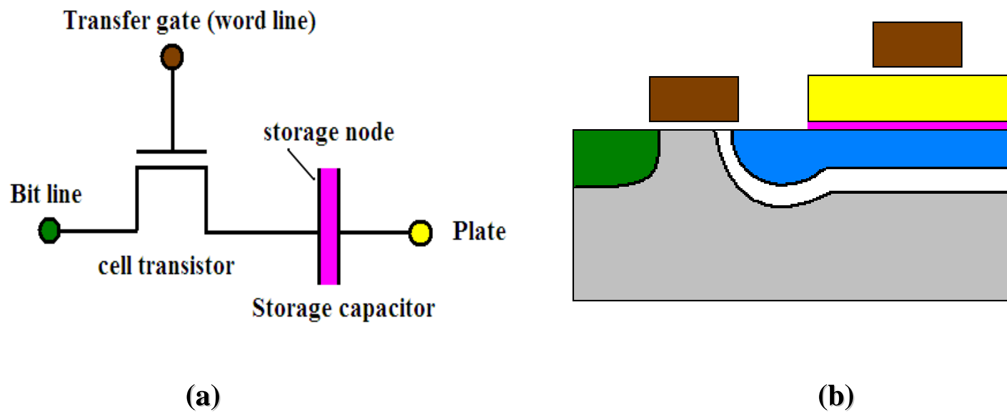


Figure 2.4:- Cross-sectional view of conventional planar capacitor DRAM cell [4].

The characteristics of 1T cell are [3]:-

1. A 1T DRAM requires a sense amplifier for each bit line to be functional. The DRAM memory cells are single ended; it complicates the sense amplifier design.
2. The readout of the 1T DRAM cell is destructive and data refreshing is necessary.
3. It requires an extra capacitance. For reliability, the charge-transfer ratio is kept large, with the minimum value of the capacitance ranging around 30 fF. Fitting that large of a capacitance in as small area is a challenge to DRAM designing.
4. When writing a '1' into the cell, a threshold voltage is lost, which reduces the available charge. This loss can be circumvented by bootstrapping the word lines.

Limitations

Scaling of cell size poses processing complexity due to the capacitors. The need for refreshing circuitry makes the design complex.

2.2.2.1. Three Dimensional Cell Structures

Advanced cell structure design activities focused on the arrangement of these two components (one transistor and one capacitor) by using advanced techniques to meet the design criteria. The major effort has been to use the third dimension in the cell structure.

2.2.2.1.1. Stacked Capacitor Cell (STC Cell)

One approach is to stack the capacitor over the access transistor, figure (2.5) [5]. Advantages of this cell are the use of the space over the transistor for charge storage, low soft error rate, and the configuration is suitable for the use of a high dielectric-constant insulator between two poly layers without the dielectric contacting the bulk silicon.

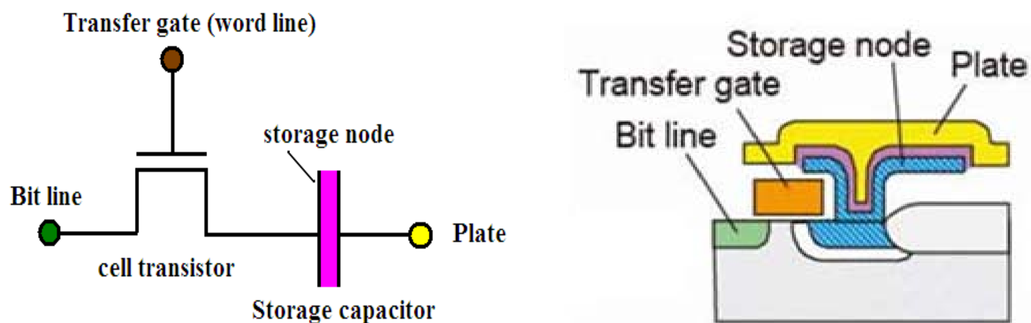


Figure 2.5:- Cross-sectional view of stacked capacitor DRAM cell [4].

The difficulties are how to manage the large surface topography and how to keep high dielectric breakdown between poly layers with scaled insulator thickness. Another difficulty for the cell is that one buried contact without self-alignment is needed for the capacitor connection per cell, thus making the cell area hard to scale. In moving toward higher density DRAMs, the cell storage area will be restricted unless the capacitor storage polysilicon electrode is made thick enough to increase the sidewall areas for charge storage or a low leakage, very high dielectric-constant storage insulator can be used.

2.2.2.1.2. Trench Capacitor Cell

Another approach is to use a trench capacitor cell, Figure (2.6) [5]. Since all four sidewalls and the bottom of the trench can be used for the capacitor electrode, the storage area is greatly enhanced with even reduced planar surface area for the capacitor.

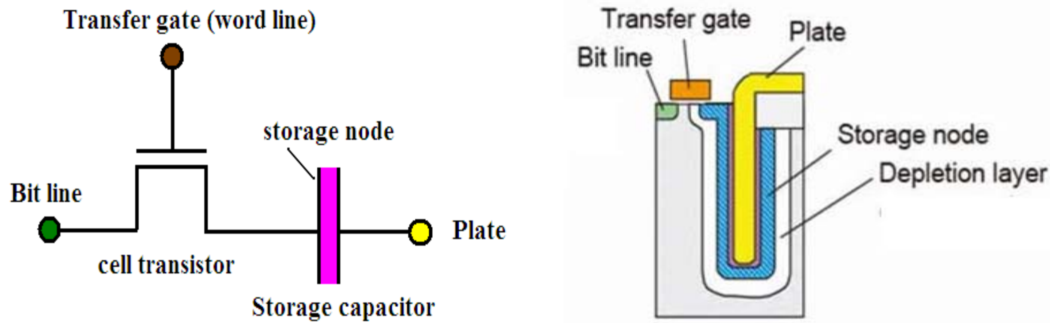


Figure 2.6:- Cross-sectional view of trench capacitor DRAM cell [4].

However, there are some design concerns in this original trench capacitor structure. Because the signal charge is stored at the trench surfaces in the bulk silicon, the expanded storage area results in high susceptibility to minority-carrier leakage disturbance, leading to high leakage and soft error rate. Moreover, unless a deep isolation is used, the minimum distance between trenches and from the trench storage node to the active devices of adjacent cells is limited because of possible junction punchthrough. One solution is to increase the background doping concentration to narrow the depletion region; however, this increases the electric field around the junction, which may cause avalanching. Another solution is to form Hi-C doping profiles at the trench surfaces by using more complicated sidewall doping techniques, though the trench spacing can still be limited by the diffusion depths of doping profiles and depletion regions.

Limitations of three-dimensional cell structures

DRAM is much denser than SRAM but the capacitor poses the problem as the device shrinks. The capacitor can not shrink much because it must stay large enough to store a detectable amount of charge while the transistors continue to scale into the infinitesimal. The growing mismatch between the size of transistors and that of capacitors has led to strange-looking arrangements, such as capacitors built as narrow trenches having a depth many times greater than the chip's transistors. Another configuration has relatively enormous fin-shaped capacitors built above the silicon in the area that usually holds the chip's wiring. In both the cases, the aspect ratio of the cell structure is distorted. Both arrangements are too expensive to put into any logic chip, requiring several extra manufacturing steps[6].

2.2.3. Capacitorless One-Transistor Floating Body Cell (1T FBC)

As CMOS goes nano-scale, the integration of standard 1T/1C DRAMs becomes very challenging. So, alternative memory devices, dense and functional at nano-scale are being investigated and many alternatives have been proposed. In this context, floating-body MOSFETs stand as a very promising architecture.

This floating body memory cell uses a floating body of partially depleted (PD) silicon-on-insulator (SOI) MOSFET as a storage node. Therefore, the 1T DRAM does not need a complicated storage capacitor and has a good process compatibility with logic devices. The schematic cross section of a basic SOI nMOSFET is shown in Figure (2.7).

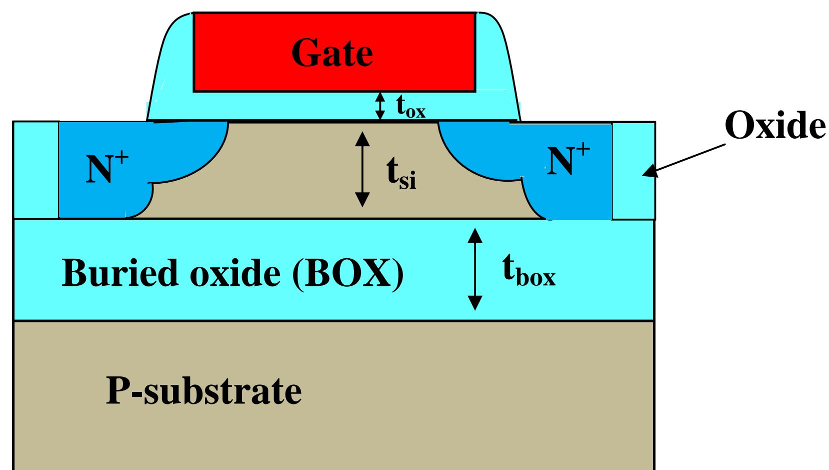


Figure 2.7:- cross sectional view of an SOI nMOSFET [7].

In SOI wafer, a very thin layer of insulating silicon dioxide, buried a few hundred nanometers or less below the surface, cuts the transistor off from the vast bulk of the wafer—which, in turn, limits the amount of charge the transistor that must move to switch on or off. The result is to speed up circuits by as much as 30% [6]. The insulation in SOI wafers blocks a major pathway for the leakage current, thus reducing the power that transistors draw by 30% when they are switching and 50% to 90% when they are not [6]. But SOI wafer are costly than plain silicon wafer of the same size. As bulk silicon devices are now facing a number of critical limitations, SOI devices become extremely attractive because of the advantages in high speed, low power dissipation and extended scalability.

The term "floating body effect" comes from the fact that the insulation layer in an SOI wafer electrically separates the body of the transistors from the rest of the silicon, letting

its voltage to float. The primary advantages of SOI DRAMs are the superior soft error rate and static data-retention time, which promise for higher integration density than bulk-Si DRAMs. The most fundamental device-design issue is the choice between a fully depleted device versus a partially depleted device.

Read/Write Operation

In the floating body cell, the floating body is used as a storage node of electric charge. An nMOSFET formed on PD-SOI were chosen as a candidate to realize this concept. For this DRAM cell, a logic state is defined by creating an excess or a shortage of the majority carriers (ΔQ) inside the body of the transistor. This causes a shift in the body potential ΔV_b below or above its equilibrium given by $\Delta V_b = \Delta Q / C_{si}$, where C_{si} is the silicon body capacitance. As a result, there is a shift in the threshold voltage ΔV_T, which is directly proportional to ΔV_b. The state is read by sensing the drain current, which has two different levels corresponding to the ΔV_T shift, and this distinguishes the two states [8].

The operation principle of the FBC is also explained in figure (2.8). The cell senses whether the majority carriers (holes) accumulate in the floating-body as the threshold voltage changes. For write operation, the source of the FBC is biased to 0 V, the drain is connected to a bit line, and the gate is connected to a word line. To write data, the nMOSFET is operated in saturation leading to impact ionization which injects holes into the body. When excess holes exist in the floating body and V_T lowers, the cell state can be regarded as '1'. This operation is shown in figure (2.8(a)). To write data '0', the p-n junction between the body and the drain is forward-biased, ejecting the stored holes from the body and thereby increasing the V_T. The cell state can be regarded as '0'. This operation is shown in figure (2.8(b)). To read the data, the nMOSFET is operated in a linear ohmic region so as to keep the data state '0', figure (2.8(c)). From being violated by impact ionization, it leads to a drain current difference due to the body effect depending on the number of holes in the body. This difference can be read and sensed. The drain current difference ΔI_{ds}, the current for the data state '1', I_1, minus the current for the data state '0', I_0, is a function of the WL bias. By performing a read operation in the linear current region, the 1T DRAM cell achieves the nondestructive read operation during a refresh interval. However, the FBC is volatile and needs to be refreshed, because holes are generated in the body of the data state '0' through the PN junction reverse-bias leakage

between the body and the source/drain and they need to be bailed out to maintain the difference in the number of holes between the data '1' and the data '0' [7].

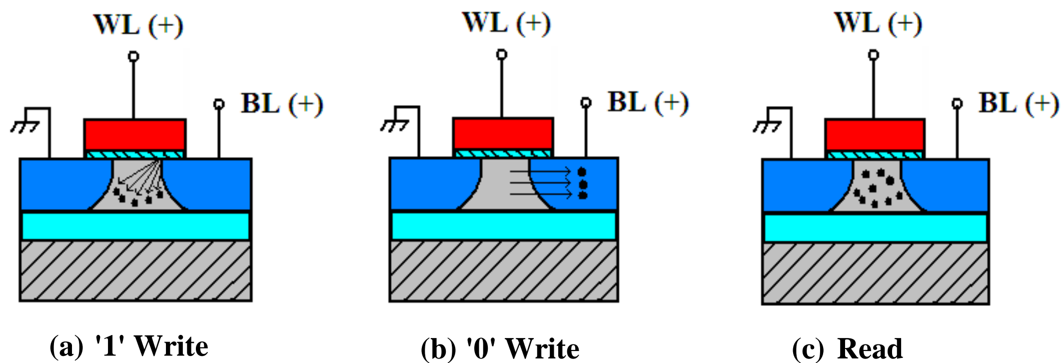


Figure 2.8:- The principle of the FBC during write '1' and '0' and read [7].

Several designs of this cell based on both partially depleted (PD) and fully depleted (FD) SOI technologies have been proposed.

2.2.3.1. Z-RAM

Innovative Silicon, Inc. (ISi) – a memory technology company founded in 2002 –claims to provide a revolutionary new type of floating body memory called Z-RAM [9]. Although functionally similar to other floating body memories, Z-RAM improves upon them by providing high signal margin and data retention time, faster read and write speeds while simultaneously reducing write power.

Read/Write of operation

Previous floating body memories have used the MOSFET to pass current and create charge in the body using impact ionization. Although memory functionality can be demonstrated, the amount of charge created is insufficient to create a robust and manufacturable memory device.

Z-RAM utilizes the bipolar transistor intrinsic in the SOI MOSFET to create charge (Figure (2.9)). This allows a much larger charge to be created and stored due to increased capacitance of the memory cell. In an nMOSFET, the n^+ source, p-type body and the n^+ drain form the emitter, base and collector respectively of an NPN bipolar transistor. The body of the MOSFET is the base of the bipolar transistor and is used as a storage node.

To write a '1' into a Z-RAM memory cell, the intrinsic bipolar transistor is triggered, causing current to flow throughout the transistor body. This differs from MOS behavior where current flows only at the interface. Charge collects at the interface due to the slight bias at the gate. The impact ionization effect used to create an excess of majority carriers in the floating body is more efficient in this bipolar bit cell structure, charging the body quickly, resulting in a very rapid write time. To write a '0' to the cell, the holes are drawn out through the source electrode, leaving excess negative charge. Z-RAM memory is read using a similar mechanism which senses the bipolar current through the transistor.

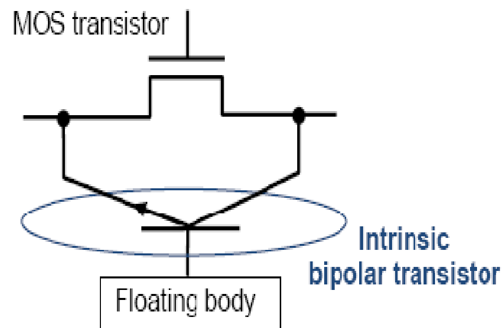


Figure 2.9:- Intrinsic bipolar transistor in SOI nMOSFET [9].

The signal margin of Z-RAM cell is superior to other floating body memories due to the large current gain available from the bipolar device. By Innovative Silicon's estimates, replacing the conventional memory with Z-RAM would shrink a chip. That would boost the number of chips per wafer and cut the final cost of the chip almost in half [6].

Limitations of 1T floating body cell design

Although the removal of the external capacitor reduces area overhead and the process complexities, the scaling of this 1T DRAM has its own specific challenges [10].

- The signal margin(ΔI_{ds}) is restricted because the attainable ΔV_T is limited. Wide devices and several parallel fins are needed to increase the current to get acceptable ΔI_{ds}, and this undermines the memory density actually achievable.
- It requires the bias-induced accumulation, which complicates the cell/chip design, undermines reliability, and sacrifice layout area.
- It relies on current based sensing of the stored data, which is less desirable than conventional voltage based sensing because of varying ΔI_{ds}, and hence, more complex sensing circuitry and added power consumption.

2.2.4. Two-Transistor Floating Body Cell (2T FBC)

The major disadvantage of restricted signal margin in 1T FBC can be overcome by a novel two-transistor (2T) FBC that can yield much better signal margin and density, while offering other significant advantages over the 1T cell ([10],[11]). The 2T FBC concept is conceived from an insightful understanding of the basic 1T FBC operation.

Read/Write Operation

A memory array based on 2T FBC is shown in Figure (2.10). The cell comprises transistors T1 and T2 with the body (B1) of T1 connected to the gate (G2) of T2; in essence, B1 is a floating gate on T2. The write/erase operations are done by charging/discharging the floating body of T1, as in the 1T FBC. But, the stored data are read via T2 with $\Delta V_{GS2} = \Delta V_{BS} > \Delta V_T$, which implies about a $2 \times (1/2r)$ density increase for same current-signal margin. Two bit lines are needed: one (BL1) connecting all drain nodes of the T1 transistors in a column of the DRAM array for programming the cells, and the other (BL2) connecting all drain nodes of the T2 transistors in the column to sense the data. Preferred voltage sensing at the (precharged) drain node of T2 can be used, which exploits the amplification of V_{BS} by T2 for more margin and higher density. Then, the 2T FBC has to be designed such that T2 will be turned on and off by V_{BS} corresponding to the charged and discharged T1 body. Since T2 inverts, stored '0' and '1' correspond to the TI body being charged and discharged, respectively.

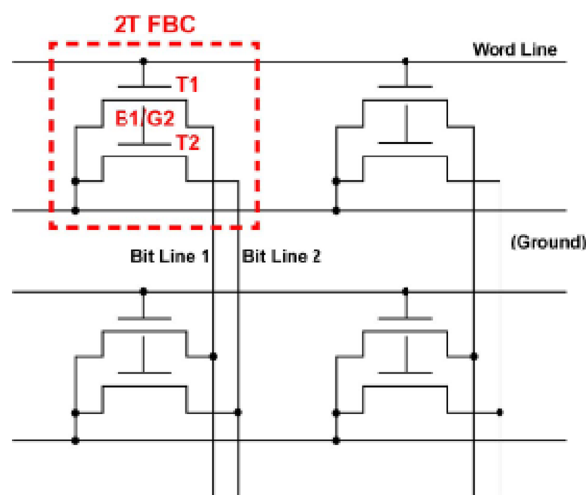


Figure 2.10:- The basic 2T floating body cell [10].

Appropriate WL pulsing for 'read' is needed to raise the B1 voltage to a level sufficiently above or below V_T of T2 for stored '0' or '1', respectively. This operation is confirmed by the predicted transient drain voltage (BL2) of T2, which needs to be precharged before reading. Indeed, with the T1 body charged (stored '0'); BL2 drops quickly to 0 V, as V_{B1} turns on T2. With the T1 body discharged (stored '1'), BL2 remains at its precharged value as V_{B1} remains well below V_T. The efficient reads of both '0' and '1' are done with reasonable WL and BL1 voltage pulsing. The 2T concept is compatible with any SOI technology, like the 1T FBC.

CHAPTER 3

LEAKAGE MECHANISMS AND REDUCTION TECHNIQUES IN TRENCH CAPACITOR DRAM CELL

In DRAM, every memory cell experiences leakage current which consumes part of the stored charge. As the DRAM cell size shrinks, the leakage increases. The duration till the stored signal can be read out successfully is called data retention time. The product refresh time is determined by the minimum retention time. Therefore, to maintain the desired data retention time, the leakage current must be kept within the acceptable limit. In this chapter, various leakage currents responsible for data degradation in trench capacitor DRAM cell are discussed.

DRAM data retention time can be expressed as [12]:

$$T_{ret} = \frac{C_S \cdot \frac{1}{2} V_{CC} - \Delta V_{BL} \cdot (C_S + C_{BL})}{I_{Leak}} \quad \text{......(2)}$$

where T_{ret} is the retention time and C_S , C_{BL} , I_{Leak} , V_{CC} and ΔV_{BL} are the storage capacitance, bitline capacitance, total leakage current, bitline operation voltage, and sense voltage, respectively. From the above equation, we can see that the retention time depends strongly on the storage capacitance, bitline capacitance and cell leakage. In order to retain the data for longer time, it is required to reduce the leakage as much as possible. With DRAM scaling. the leakage must be kept as low as 1 fA per cell for each technology generation to maintain the required retention time [12]. So, the leakage reduction of a DRAM cell is very important design issue.

There are several leakage current components, shown in figure (3.1), that degrade the stored signal in a trench capacitor DRAM cell [12], [13]:

(a). Gate-induced drain leakage (GIDL);
(b). Junction leakage from storage node;
(c). Subthreshold leakage of the cell transistor;
(d). Parasitic vertical transistor leakage;
(e). Isolation leakage between the neighboring cells;
(f). Capacitor node dielectric leakage current;

(g). Gate leakage

(h). Passing wordline leakage

The sum of all leakages determines the total leakage current and therefore limits the maximum data retention time of the specific cell.

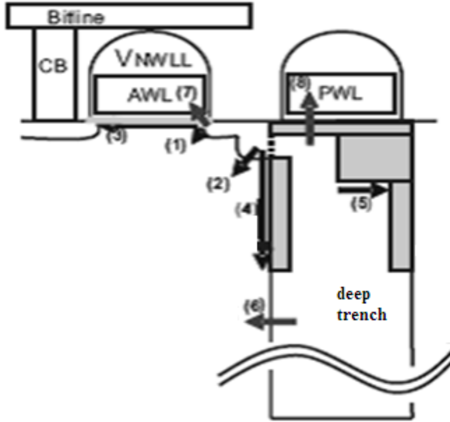


Figure 3.1: Leakage paths in a trench capacitor DRAM cell [13].

High leakage current in deep-submicrometer regimes is also becoming a significant contributor to power dissipation of CMOS circuits. Consequently, the identification, modeling and minimization of different leakage components of individual cells is very important aspect of cell designing for the improvement of data retention time and for estimation and reduction of leakage power, especially for low-power applications. In this thesis, DRAM access transistor's intrinsic leakage mechanisms specifically subthreshold leakage, GIDL and punchthrough have been focused.

3.1. DRAM Access Transistor Leakage Mechanisms

Transistor off-state current (I_{OFF}) is the drain current when the gate voltage is zero. I_{OFF} is influenced by the threshold voltage, channel physical dimensions, channel/surface doping profile, drain/source junction depth, gate oxide thickness and V_{DD}. In long-channel devices, I_{OFF} is dominated by leakage from the drain-well and well-substrate reverse biased pn junctions. Short-channel transistors require lower power supply levels to reduce their internal electric fields and power consumption. This forces a reduction in the threshold voltage that causes a substantially large increase in I_{OFF}. This increase is due to the weak inversion state leakage and is a function of V_T.

In figure (3.2), six short-channel leakage mechanisms are illustrated. I_1 is the reverse-bias pn junction leakage; I_2 is the subthreshold leakage; I_3 is the GIDL; I_4 is the channel punchthrough current; I_5 is the oxide tunneling current and I_6 is the gate current due to hot-carrier injection.

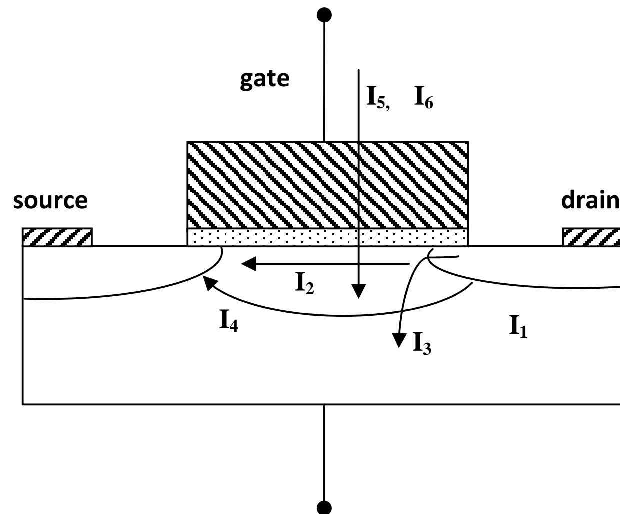


Figure 3.2: Leakage current mechanisms of deep-submicrometer transistors [14].

3.1.1. PN Junction Reverse Bias Current (I_1)

Drain and source to well junctions are typically reverse biased, causing pn junction leakage current. A reverse-bias pn junction leakage has two main components: one is minority carrier diffusion/drift near the edge of the depletion region; the other is due to electron-hole pair generation in the depletion region of the reverse-biased junction [14]. For an MOS transistor, additional leakage can occur between the drain and well junction from gated diode device action (overlap of the gate to the drain-well pn junctions) or carrier generation in drain to well depletion regions with influence of the gate on these current components. PN junction reverse-bias leakage is a function of junction area and doping concentration. If both n and p regions are heavily doped band-to-band tunneling (BTBT) dominates the pn junction leakage.

3.1.1.1. Band-to-Band Tunneling Current

High electric field (10^6 V/cm) across the reverse-biased pn junction causes significant current to flow through the junction due to tunneling of electrons from the valence band of the p region to the conduction band of the n region. For the tunneling to occur, the total

voltage drop across the junction has to be more than the band gap. The tunneling current density is given by [14],

$$J_{b-b} = A\frac{EV_{app}}{E_g^{1/2}}\,exp\left(-B\,\frac{E_g^{3/2}}{E}\right) \quad \text{......(3)}$$

where,

$$A = \frac{\sqrt{2m^*}\,q^3}{4\pi^3\hbar^2} \quad \text{......(4)}$$

and

$$B = \frac{4\,\sqrt{2m^*}}{3q\hbar} \quad \text{......(5)}$$

where m^* is effective mass of electron; E_g is the energy-band gap; V_{app} is the applied reverse bias; E is the electric field at the junction; q is the electronic charge; and $\hbar$ is $1/2\pi$ times Planck's constant. Assuming a step junction, the electric field at the junction is given by [14],

$$E = \sqrt{\frac{2q\,N_aN_d(V_{app}+V_{bi})}{\varepsilon_{si}\,(N_a+N_d)}} \quad \text{......(6)}$$

where N_a and N_d are the doping in the p and n side, respectively; ε_{si} is permittivity of silicon; and V_{bi} is the built in voltage across the junction. In scaled devices, high doping concentrations and abrupt doping profiles cause significant BTBT current through the drain-well junction.

3.1.2. Subthreshold Leakage (I_2)

Subthreshold or weak inversion conduction current between source and drain in an MOS transistor occurs when gate voltage is below threshold voltage, V_{th}. In the weak inversion, the minority carrier concentration is small, but not zero. Let us consider that the source of the n-channel MOSFET is grounded, $V_g < V_{th}$, and the drain to source voltage $|V_{ds}| \geq 0.1$ V. For such weak inversion condition, V_{ds} drops almost entirely across the reverse-biased substrate-drain pn junction. As a result, the variation of the electrostatic potential ϕ_s at the semiconductor surface along the channel is small. The component of the electric field vector E, along the channel, is also small. With both the number of mobile carriers and the longitudinal electric field small, the drift component of the subthreshold drain-to-source current is negligible. Therefore, unlike the strong inversion region in which the drift current dominates, the subthreshold conduction is dominated by the diffusion current. The carriers move by diffusion along the surface. Weak inversion typically

dominates modern device off-state leakage due to the low V_{th}. The weak inversion current can be expressed based on the following [14]:

$$I_{ds} = \mu_0 C_{ox} \frac{W}{L} (m-1)(\upsilon_T)^2 \times e^{(V_g - V_{th})/m\upsilon_T} \times \left(1 - e^{-\upsilon_{DS}/\upsilon_T}\right) \quad ......(7)$$

where

$$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{\frac{\varepsilon_{si}}{W_{dm}}}{\frac{\varepsilon_{ox}}{t_{ox}}} = 1 + \frac{3t_{ox}}{W_{dm}} \quad ......(8)$$

where V_{th} is the threshold voltage, and $\upsilon_T = KT/q$ is the thermal voltage. C_{ox} is the gate oxide capacitance; μ_0 is the zero bias mobility; and m is the subthreshold swing coefficient (also called body effect coefficient). W_{dm} is the maximum depletion layer width, and t_{ox} is the gate oxide thickness. C_{dm} is the capacitance of the depletion layer.

In long-channel devices, the subthreshold current is independent of the drain voltage for V_{DS} larger than a few υ_T. On the other hand, the dependence on the gate voltage is exponential. The inverse of the slope of the $\log_{10}(I_{ds})$ versus V_{gs} characteristic is called the subthreshold slope (S_t) [14] and is given by

$$S_t = \left(\frac{d(\log_{10} I_{ds})}{dV_{gs}}\right)^{-1} = 2.3\frac{mkT}{q} = 2.3\frac{kT}{q}\left(1 + \frac{C_{dm}}{C_{ox}}\right) \quad ......(9)$$

Subthreshold slope indicates how effectively the transistor can be turned off (rate of decrease of I_{OFF}) when V_{gs} is decreased below V_{th}. As device dimensions and the supply voltage are scaled down to enhance performance, power efficiency, and reliability, subthreshold characteristics may limit the scalability of the supply voltage. The parameter S_t is measured in millivolts per decade of the drain current. For the limiting case of $t_{ox} \rightarrow 0$ and at room temperature, $S_t \approx 60$ mV/decade. Typical S_t values for a bulk CMOS process can range from 70 to 120 mV/decade. A low value for subthreshold slope is desirable. It can be seen from the preceding expression (9) that S_t can be made smaller by reducing t_{ox} or a lower substrate doping concentration. Changes in lower temperature or a substrate bias also modifies S_t. The factors affecting the subthreshold current are described in the following subsections.

3.1.2.1. Drain-Induced Barrier Lowering

In long-channel devices, the source and drain are separated far enough that their depletion regions have no effect on the potential or field pattern in most part of the device. Hence, for such devices, the threshold voltage is virtually independent of the channel length and drain bias. In a short-channel device, however, the source and drain depletion width in the vertical direction and the source drain potential have a strong effect on the band bending over a significant portion of the device. Therefore, the threshold voltage, and consequently the subthreshold current of short-channel devices, vary with the drain bias. This effect is referred to as drain induced barrier lowering (DIBL). One way to describe it is to consider the energy barrier at the surface between the source and drain. Under off conditions, this potential barrier prevents electrons from flowing to the drain. For a long-channel device, the barrier height is mainly controlled by the gate voltage and is not sensitive to V_{DS}. The barrier of a short-channel device reduces when the depletion regions of the drain and the source interact with each other near the channel surface and when a high drain voltage is applied, it further lowers the barrier height, resulting in further decrease of the threshold voltage and increase of subthreshold current. The source then injects carriers into the channel surface (independent of gate voltage). The surface DIBL typically occurs before the deep bulk punchthrough. DIBL does not change the subthreshold slope, but does lower V_{th}. Higher surface and channel doping and shallow source/drain junction depths reduce the DIBL effect on the subthreshold leakage current.

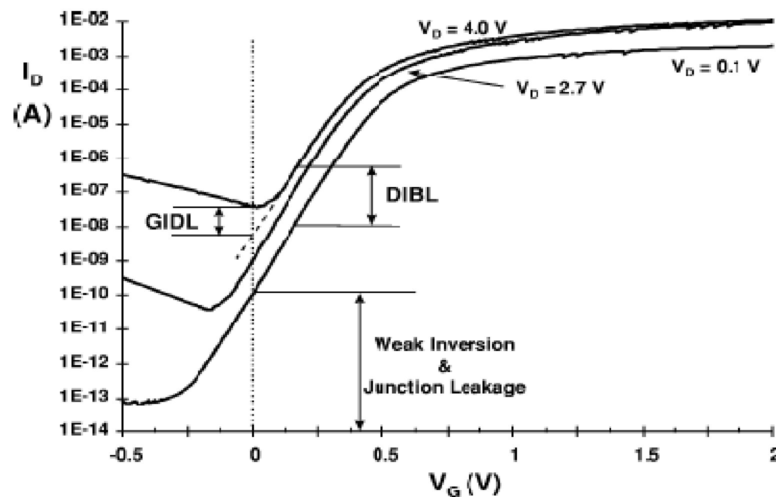


Figure 3.3: N-channel I_D vs. V_G curve showing DIBL, GIDL, weak inversion and pn junction reverse-bias leakage components [14].

Figure 3.3 illustrates the DIBL effect as it moves the $I_D - V_G$ curve up and to the left as the drain voltage increases. DIBL can be measured at constant V_G as the change in I_D for a change in V_D.

3.1.2.2. Body Effect

Reverse biasing well-to-source junction of a MOS transistor widens the bulk depletion region and increases the threshold voltage. The effect of body bias can be considered in the threshold voltage equation [14],

$$V_{th} = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B + V_{bs})}}{C_{ox}} \quad \text{......(10)}$$

where V_{fb} is the flat-band voltage; N_a is the doping density in the substrate; and $\psi_B = (KT/q)\ln(N_a/n_i)$ is the difference between the Fermi potential and the intrinsic potential in the substrate.

The subthreshold leakage of an MOS device including weak inversion, DIBL, and body effect, can be modeled as [14],

$$I_{subthreshold} = A \times e^{1/m\upsilon_T (V_G - V_S - V_{th0} - \gamma' \times V_S + \eta V_{DS})} \times \left(1 - e^{-\upsilon_{DS}/\upsilon_T}\right) \quad \text{......(11)}$$

where

$$A = \mu_0 C'_{ox} \frac{W}{L_{eff}} (\upsilon_T)^2 e^{1.8} e^{-\Delta V_{th}/\eta \upsilon_T} \quad \text{......(12)}$$

V_{th0} is the zero bias threshold voltage. The body effect for small values of source to bulk voltages is linear and is represented by the term $\gamma' V_S$ in (11), where γ' is the linearized body effect coefficient, η is the DIBL coefficient, C_{ox} is the gate oxide capacitance, μ_0 is the zero bias mobility, and m is the subthreshold swing coefficient of the transistor. ΔV_{th} is a term introduced to account for transistor-to-transistor leakage variations.

3.1.2.3. Narrow Width Effect

The decrease in gate width modulates the threshold voltage of a transistor, and thereby modulates the subthreshold leakage. In the LOCOS gate MOSFET, the existence of the fringing field causes the gate-induced depletion region to spread outside the defined channel width and under the isolations. This results in an increase of the total depletion

charge in the bulk region above its expected value. The threshold voltage of MOS can be defined using depletion approximation as [14],

$$V_{th} = V_{fb} + \phi_s + \frac{Q_B}{C_{ox}} \quad \text{......(13)}$$

where V_{fb} is the flat-band voltage; ϕ_s is the surface potential; C_{ox} is the capacitance across the oxide; and Q_B is the depletion charge in the bulk. Due to narrow-width effect, Q_B increases by ΔQ_B. This effect becomes more substantial as the channel width decreases, and the depletion region underneath the fringing field is comparable to the classical depletion formed by the vertical field. This results in increase of threshold voltage due to narrow-channel effect. This narrow-width effect can be modeled as an increase in V_{th} by the amount V_{NCE} given by [14],

$$V_{NCE} = \frac{\pi q N_{sub} x_{d,max}^2}{2 C_{ox} W_{eff}} = 3\pi \frac{t_{ox}}{W_{eff}} \phi_s \quad \text{......(14)}$$

where N_{sub} is the substrate doping; $x_{d,max}$ is the maximum vertical depletion width; C_{ox} is the capacitance across the oxide; W_{eff} is the effective width; t_{ox} is the oxide thickness; and ϕ_s is the surface potential. The second way that narrow-width modulates the threshold voltage is due to the fact that the channel doping is higher along the width dimension in LOCOS gates. Due to the channel stop, dopants encroach under the gate. Hence, a higher voltage is needed to completely invert the channel.

A more complex effect is seen in trench isolation devices, known as inverse-narrow-width effect. In the case of trench isolation devices, depletion layer cannot spread under the oxide isolation. Hence, the total depletion charge in the bulk does not increase ($\Delta Q_B \approx 0$), thereby eliminating the increase in the threshold voltage. On the other hand, due to the two-dimensional (2D) field-induced edge fringing effect at the gate edge, formation of an inversion layer at the edges occurs at a lower voltage than the voltage required at the center. Moreover, the overall gate capacitance (C_T) now includes the sidewall capacitance (C_F) due to overlap of the gate with the isolation oxide. This increases the overall gate capacitance. Overall gate capacitance is therefore given by $C_T = C_{ox}W + 2C_F$, which is greater than C_{ox} given in eqn.(14). Hence, the overall V_{th} reduces.

3.1.2.4. Effect of Channel Length and V_T-rolloff

Threshold voltage of MOSFET decreases as the channel length is reduced. This reduction of threshold voltage with reduction of channel length is known as V_{th} rolloff. The principal reason behind this effect is the presence of 2D field patterns in short-channel devices instead of one-dimensional (1D) field patterns in long-channel devices. This 2D field pattern originates from the proximity of source and drain regions [14]. There are depletion regions surrounding the source and drain junctions. In long channel devices, since the source and drain are far apart, their depletion regions do not have much effect on the potential profile or field pattern in most parts of the channel. However, in the case of short-channel devices, source-to-drain distance is comparable to the depletion width in the vertical direction. As a result, source drain depletion width has a more pronounced effect on potential profiles and field patterns. The source and drain depletion regions now penetrate more into the channel length, resulting in part of the channel being already depleted. Thus, gate voltage has to invert less bulk charge to turn a transistor on. In other words, for the same gate voltage, there is more band bending in the Si–SiO_2 interface in a short-channel device as compared with a long-channel one. Consequently, the threshold voltage is lower for a short-channel device. The effect of the source drain depletion region is more severe at a high drain bias. High drain bias results in more depletion charge in the channel from the drain and source, resulting in further decrease of the threshold voltage, and hence, larger subthreshold current.

3.1.2.5. Effect of Temperature

Temperature dependence of the subthreshold leakage current is important, since digital very large scale integration (VLSI) circuits usually operate at elevated temperatures due to the power dissipation (heat generation) of the circuit. Two parameters increase the subthreshold leakage as temperature is raised:

1) S_t linearly increases with temperature; and

2) the threshold voltage decreases.

3.1.3. Gate Induced Drain Leakage (I_3)

GIDL is due to high field effect in the drain junction of an MOS transistor. When the gate is biased to form an accumulation layer at the silicon surface, the silicon surface under the

gate has almost same potential as the p-type substrate. Due to presence of accumulated holes at the surface, the surface behaves like a p region more heavily doped than the substrate. This causes the depletion layer at the surface to be much narrower than elsewhere as shown in Figure 3.4(a). The narrowing of the depletion layer at or near the surface causes field crowding or an increase in the local electric field, thereby enhancing the high field effects near that region. When the negative gate bias is large (i.e., gate at zero or negative and drain at V_{DD}), the n^+ drain region under the gate can be depleted and even inverted as shown in Figure 3.4(b). This causes more field crowding and peak field increase, resulting in a dramatic increase of high field effects such as avalanche multiplication and BTBT.

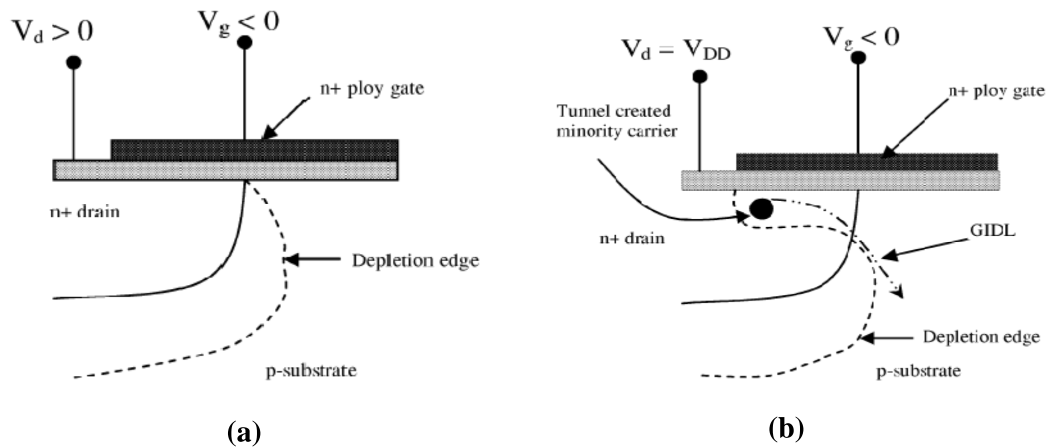


Figure 3.4: Condition of the depletion region near the drain-gate overlap region of an MOS transistor when (a) surface is accumulated with low negative gate bias; (b) n^+ region is depleted or inverted with high negative bias.

The possibility of tunneling via near-surface traps also increases. As a result of all these effects, minority carriers are emitted in the drain region underneath the gate. Since the substrate is at a lower potential for minority carriers, the minority carriers that have been accumulated or formed at the drain depletion region underneath the gate are swept laterally to the substrate, completing a path for the GIDL. Thinner oxide thickness and higher V_{DD} (higher potential between gate and drain) enhance the electric field and therefore increase GIDL. The impact of drain and well doping on GIDL is rather complicated. At low drain doping, the electric field is not high enough to cause tunneling. At very high drain doping, the depletion width—and, therefore, the tunneling volume—are limited, causing less GIDL. Hence, GIDL is worse for moderate drain doping, where

both the electric field and depletion width (tunneling volume) are considerable. Very high and abrupt drain doping is preferred for minimizing GIDL, as it provides lower series resistance required for high transistor drive currents.

3.1.4. Punchthrough (I_4)

In short-channel devices, due to the proximity of the drain and the source, the depletion regions at the drain-substrate and source-substrate junctions extend into the channel. As the channel length is reduced, if the doping is kept constant, the separation between the depletion region boundaries decreases. An increase in the reverse bias across the junctions (with increase in V_{ds}) also pushes the junctions nearer to each other. When the combination of channel length and reverse bias leads to the merging of the depletion regions, punchthrough is said to have occurred. In submicrometer MOSFETs, a V_{th} adjust implant is used to have a higher doping at the surface than that in the bulk. This causes a greater expansion of the depletion region below the surface (due to smaller doping there) as compared to the surface. Thus, the punchthrough occurs below the surface. An increase in the drain voltage beyond the value required to establish the punchthrough lowers the potential barrier for the majority carriers in the source. Thus, more of these carriers cross the energy barrier and enter into the substrate, and the drain collects some of them. The net effect is an increase in the subthreshold current. Furthermore, punchthrough degrades the subthreshold slope. The device parameter commonly used to characterize the punchthrough is the punchthrough voltage V_{PT}, which estimates the value of V_{ds} for which the punchthrough occurs at $V_{gs} = 0$. It is roughly estimated as the value of the V_{ds} for which the sum of the widths of the drain and source depletion regions is equal to effective channel length [14],

$$V_{BT} \propto N_B\left(L - W_j\right)^3 \qquad \text{......(15)}$$

where N_B is the doping concentration at the bulk; L is the channel length; and W_j is the junction width.

The most suitable method for controlling the punchthrough is to use additional implants. A layer of higher doping at a depth equal to that of the bottom of the junction depletion regions is one possible solution. Another approach could be to form a halo implant at the leading edges of the drain and source junctions.

3.1.5. Tunneling into and through Gate Oxide (I_5)

Reduction of gate oxide thickness results in an increase in the field across the oxide. The high electric field coupled with low oxide thickness results in tunneling of electrons from substrate to gate and also from gate to substrate through the gate oxide, resulting in the gate oxide tunneling current.

When a positive bias is applied to the gate, due to the small oxide thickness, which results in a small width of the potential barrier, the electrons at the strongly inverted surface can tunnel into or through the SiO_2 layer and hence give rise to the gate current. On the other hand, if a negative gate bias is applied, electrons from the n^+ polysilicon can tunnel into or through the oxide layer and give rise to the gate current.

The mechanism of tunneling between substrate and gate polysilicon can be primarily divided into two parts, namely:

(1) Fowler–Nordheim (FN) tunneling

(2) direct tunneling.

In the case of FN tunneling, electrons tunnel through a triangular potential barrier, whereas in the case of direct tunneling, electrons tunnel through a trapezoidal potential barrier. The tunneling probability of an electron depends on the thickness of the barrier, the barrier height, and the structure of the barrier. Therefore, the tunneling probabilities of a single electron in FN tunneling and direct tunneling are different, resulting in different tunneling currents.

3.1.6. Injection of Hot Carriers from Substrate to Gate Oxide (I_6)

In a short-channel transistor, due to high electric field near the Si–SiO_2 interface, electrons or holes can gain sufficient energy from the electric field to cross the interface potential barrier and enter into the oxide layer. This effect is known as hot-carrier injection. The injection from Si to SiO_2 is more likely for electrons than holes, as electrons have a lower effective mass than that of holes, and the barrier height for holes (4.5 eV) is more than that for electrons (3.1 eV) [14].

3.2. Process Level Techniques For Leakage Reduction In Trench Capacitor Dram Cell

The reduction in leakage current has to be achieved using both process- and circuit-level techniques. At the process level, leakage reduction in transistors can be achieved by controlling the dimensions (length, oxide thickness, junction depth, etc.) and doping profile in transistors. At the circuit level, threshold voltage and leakage current of transistors can be effectively controlled by controlling the voltages of different device terminals [drain, source, gate, and body (substrate)]. In this chapter, we will consider major process techniques for leakage control and reduction.

3.2.1. Channel Engineering for Leakage Reduction

Based on constant field scaling, the short channel effects can be kept under control by scaling down the vertical dimensions, for example, gate insulator thickness, junction depth, along with the horizontal dimensions, while also proportionally decreasing the applied voltages. The substrate doping concentration should increase to decrease the depletion width proportionally. The principle of constant field scaling lies in scaling the device voltages and the device dimensions (both horizontal and vertical) by the same factor, $K\ (> 1)$, such that the electric field remains unchanged. Constant electric field assures the reliability of the scaled device in terms of hot-carrier injection.

To minimize short channel effects, a sufficiently large aspect ratio (AR) of the device is required [14]. AR is defined as,

$$AR = \frac{dimension_{lateral}}{dimension_{vertical}} \quad \text{……(16)}$$

For a MOSFET, AR can be expressed as,

$$AR = \frac{L}{\left[t_{ox}\left(\frac{\varepsilon_{si}}{\varepsilon_{ox}}\right)\right]^{1/3} W_{dm}^{1/3} X_j^{1/3}} \quad \text{……(17)}$$

where ε_{si} and ε_{ox} are silicon and oxide permittivities; and L, t_{ox}, W_{dm} and X_j are channel length, gate oxide thickness, depletion depth, and junction depth, respectively. From eqn.(17), we can see that reducing t_{ox}, W_{dm} and X_j will reduce the short channel effects of a MOSFET.

A key parameter is the maximum gate depletion width, W_{dm}, within which mobile carriers (holes in the case of nMOSFETs) are swept away by the applied gate field. For uniformly doped case

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}KT\ln\left(\frac{N_a}{n_i}\right)}{q^2N_a}} \qquad \text{……(18)}$$

where n_i is the intrinsic carrier concentration.

In addition to gate oxide thickness and junction scaling, another technique to improve short-channel characteristics is well engineering. By changing the doping profile in the channel region, the distribution of the electric field and potential contours can be changed. The goal is to optimize the channel profile to minimize the OFF-state leakage while maximizing the linear and saturated drive currents of MOSFET. Supersteep retrograde wells and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the OFF-state leakage current.

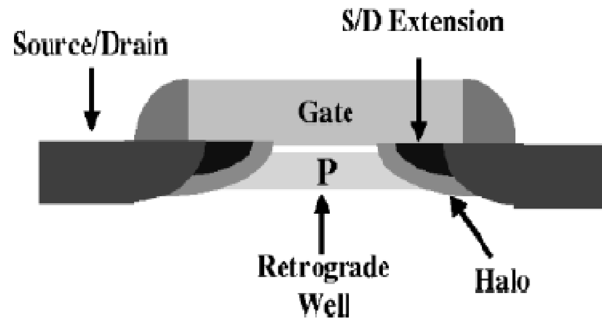


Figure 3.5: Graphical representation of different aspects of well engineering [14].

Figure 3.5 is a schematic representation of the transistor regions that are affected by the different types of well engineering. Retrograde well engineering changes the 1D characteristics of the well profile by creating a retrograde profile toward the Si–SiO_2 surface. The halo profile creates a localized 2D dopant distribution near the S/D extension regions. The use of these two techniques to increase the device performance, while keeping leakage to a tolerable limit, is discussed in the following subsections.

3.2.1.1. Retrograde Doping

To maintain acceptable OFF-state leakage with continually decreasing channel lengths, both the oxide thickness and the gate-controlled depletion width in silicon,

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a}} \quad ......(19)$$

must be reduced in proportion to the channel length to offset the degradation in short channel effects for extremely small devices. This requires an increase in the channel-doping concentration (N_a). This leads to a higher threshold voltage for a uniformly doped channel, according to the following [14]:

$$V_{th} = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}} \quad ......(20)$$

However, if the threshold voltage is not scaled, the device performance for low supply voltages will degrade due to the large reduction in gate drive. To reduce the gate-controlled depletion width while fulfilling the V_{th} reduction trend, retrograde doping can be used.

Retrograde channel doping is a vertically nonuniform, low-high channel doping. It is used to improve the short channel effects and to increase surface channel mobility by creating a low surface channel concentration followed by a highly doped subsurface region. The low surface concentration increases surface channel mobility by minimizing channel impurity scattering while the highly doped subsurface region acts as a barrier against punchthrough. For the same gate depletion width (W_{dm}), the surface electric field and the total depletion charge of an extreme retrograde channel is one-half that of a uniformly doped channel. This reduces the threshold voltage and improves mobility. Retrograde channel doping allows the threshold voltage (V_{th}) to be decoupled from the gate-controlled depletion width (W_{dm}). However, the body effect coefficient and the subthreshold slope are still coupled to the gate depletion width (W_{dm}).

3.2.1.2. Halo Doping

The most important device change over the past decade to control subthreshold current was the addition of large angle implants called “halos”. Halo doping or nonuniform

channel profile in a lateral direction was introduced below 0.25-μm technology node to provide another way to control the dependence of threshold voltage on channel length [14]. The halo implant is a high angle implant of well type dopant species introduced into the device after transistor gate patterning (for example, p-type dopant implanted into the p-well of nMOSFET). Typically the same lithography step as the source/drain extension implant is used. Because of the high halo implant angle, multiple implants with rotations are needed to ensure uniform doping on all sides of channel.

For n-channel MOSFETs, more highly p-type doped regions are introduced near the two ends of the channel as shown in figure (3.5). Under the edges of the gate an at the end of the channel, point defects are injected during sidewall oxidation. These point defects gather doping impurities from the substrate, thereby increasing the doping concentration near the source and drain end of the channel. More highly doped p-type substrate near the edges of the channel reduces the charge-sharing effects from the source and drain fields, thus reducing the width of the depletion region in the drain-substrate and source-substrate regions. As the channel length is reduced, these highly doped regions consume a larger fraction of the total channel.

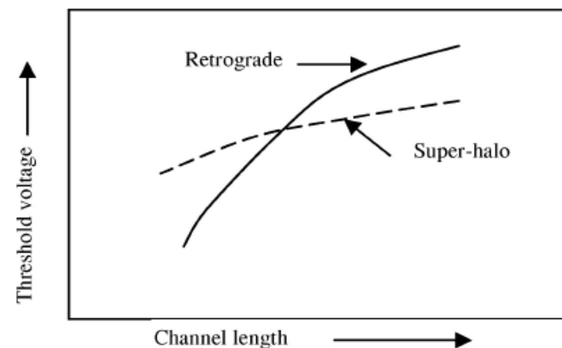


Figure 3.6: Short channel threshold voltage roll-off for retrograde and superhalo [14].

Reduction of charge-sharing effects reduces the threshold voltage degradation due to channel length reduction. Thus, threshold voltage dependence on channel length becomes more flat as shown in figure 3.6. Hence, the off-current becomes less sensitive to channel length variation. The reduction in drain and source junction depletion region width also reduces the barrier lowering in the channel, thus reducing DIBL. Since the channel edges are more heavily doped and junction depletion widths are smaller, the distance between

source and drain depletion regions is larger. This reduces the punchthrough possibility. The higher doping near the channel edges causes larger BTBT and higher GIDL. The BTBT currents in the high-field region near the drain ultimately limit the halo doping level.

CHAPTER 4

PROCESS STEPS FOR FABRICATION OF TRENCH CAPACITOR DRAM CELL

This chapter describes the process steps for the fabrication of trench capacitor DRAM cell. The 2D process simulation has been done using ATHENA package of SILVACO. The first step in modeling any simulation structure is specifying its initial grid. The correct specification of a grid is critical in process simulation. The number of nodes in the grid N_p has a direct influence on simulation accuracy and time. A finer grid should exist in those areas of the simulation structure where ion implantation will occur, where p-n junction will be formed, or where optical illumination will change photoactive component concentration [15].

For the DRAM cell structure, deep-submicron 0.18 µm nMOSFET as the access transistor and the conventional trench capacitor as the storage device have been considered for the analysis. The physical parameters for the nMOSFET has been specified according to the International Technology Roadmap for Semiconductors (ITRS) [16]. Table (4.1) illustrates values of the physical parameters of 0.18 µm nMOSFET taken into consideration for the design.

Table 4.1:- Physical parameters for the design of 0.18 µm nMOSFET [16].

Physical Parameter	Value
Technology Node	0.18 µm Technology
DRAM Bits/Chip	1 Gbit
Minimum Supply Voltage (volts)	1.5 – 1.8
Minimum Logic Vdd (volts)	1.5 – 1.8
Polysilicon Deposit (nm)	200
Gate Length (nm)	0.18
Gate Oxide Thickness (nm) (equivalent)	3 – 4
Spacer Deposition (nm)	100 nm (oxide)
Channel Dopant Concentration	2×10^{18} cm^{-3}
Initial Substrate doping concentration	B = 1×10^{15} cm^{-3}

Source/Drain implant Doping Concentration	As = 5 x 10^{15} cm^{-2}, E = 60 keV
V_{th} Adjust Implant Doping Concentration	B = 1.5 x 10^{13} cm^{-2}, E = 45 keV
Halo Implant Doping Concentration	B = 3 x 10^{13} cm^{-2}, E = 15 keV, 30° (implantation angle)
Retrograde Well Doping Concentration, Energy	B = 3 x 10^{13} cm^{-2}, E = 200 keV
LDD Implant Doping Concentration	As = 1 x 10^{15} cm^{-2}, E = 30 keV
Final Rapid Thermal Anneal (RTA)	1000°C/1 sec

B => Boron, As => Arsenic, E => Energy

As per the first step in the process flow, the grid has been specified and then the cleaned silicon substrate is lightly doped with boron with concentration of $10^{15}/cm^3$. This is shown in figure (4.1).

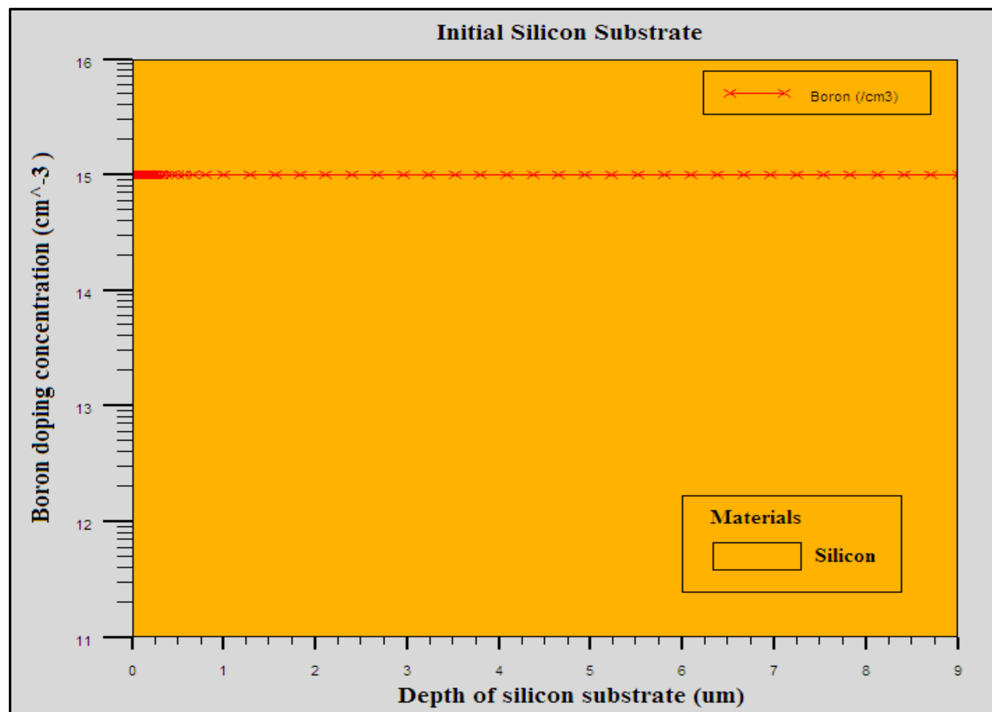


Figure 4.1:- Initial silicon substrate.

Then the substrate surface is typically covered by a pad insulator such as silicon nitride. Silicon nitride is then etched to make window for the trench formation. Then the silicon substrate is etched up to depth of 8 μm using reactive ion etching [17]. A node dielectric (silicon dioxide is chosen in this case) is then deposited on the sidewalls and bottom of the trench and on the top of the pad insulator. This is shown in figure (4.2).

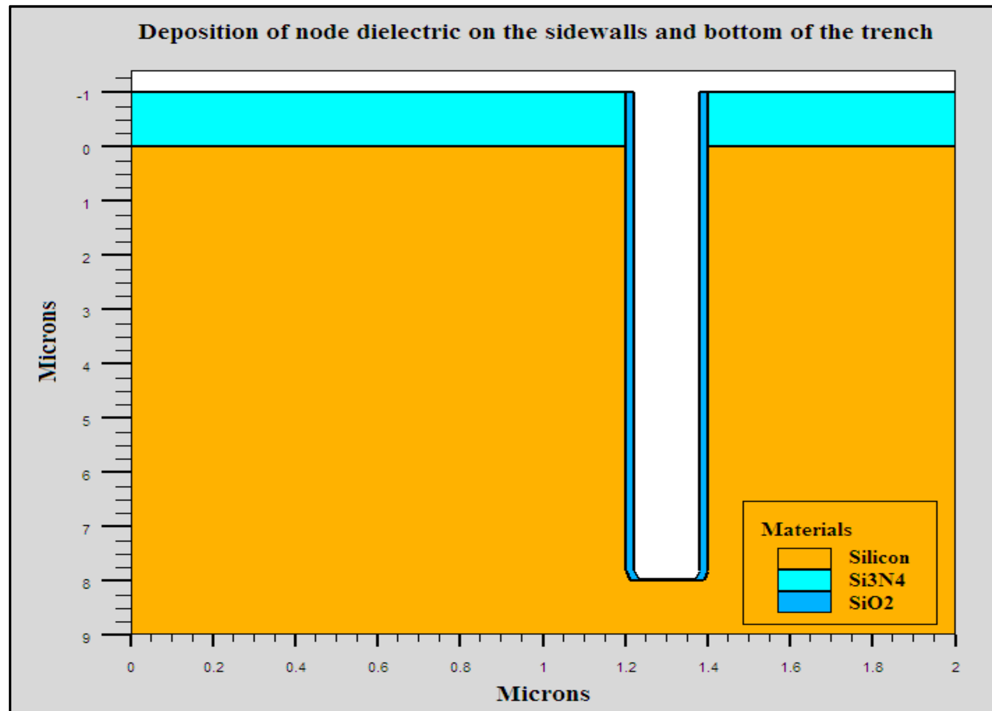


Figure 4.2:- Node dielectric deposited on the sidewalls and bottom of trench.

A layer of polysilicon is deposited on this dielectric, thereby covering the top surface of the pad insulator and covering the trench. The polysilicon is etched so that it is recessed in the trench. The node dielectric is then removed from the top surface and upper sidewalls of the trench. The recessed polysilicon the forms a node at the bottom of the trench. This is shown in figure (4.3).

Another dielectric layer (typically oxide) is deposited on the top surface, the trench sidewalls and the top surface of node polysilicon: this layer is etched so as to leave a collar in the interior of the trench on the upper sidewalls in the trench. This is shown in figure (4.4).

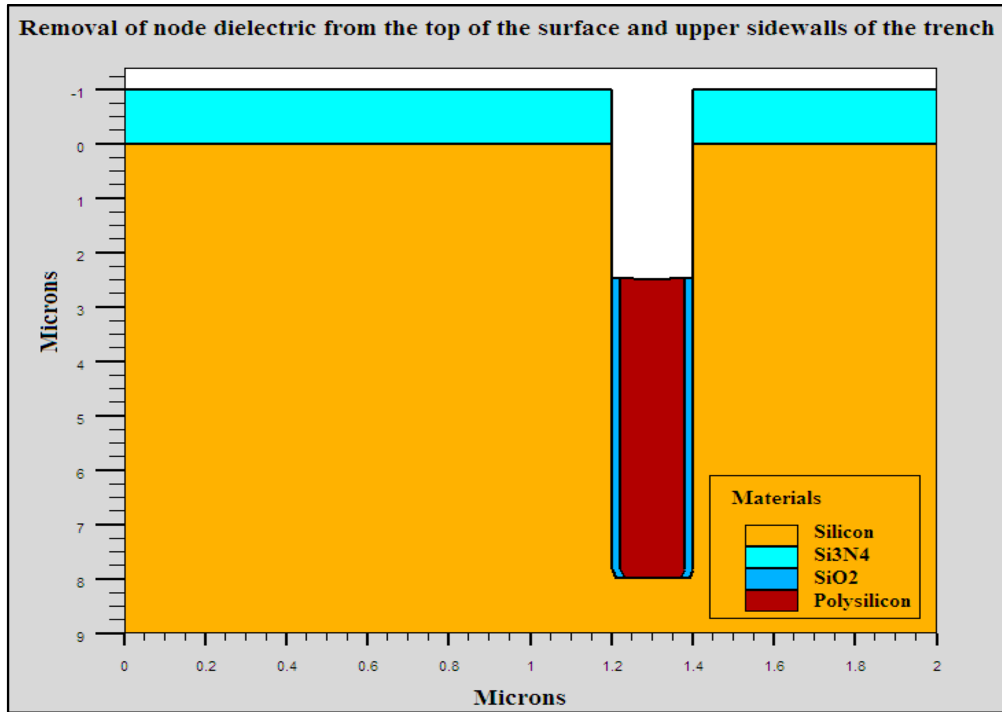


Figure 4.3:- Recess of first polysilicon fill.

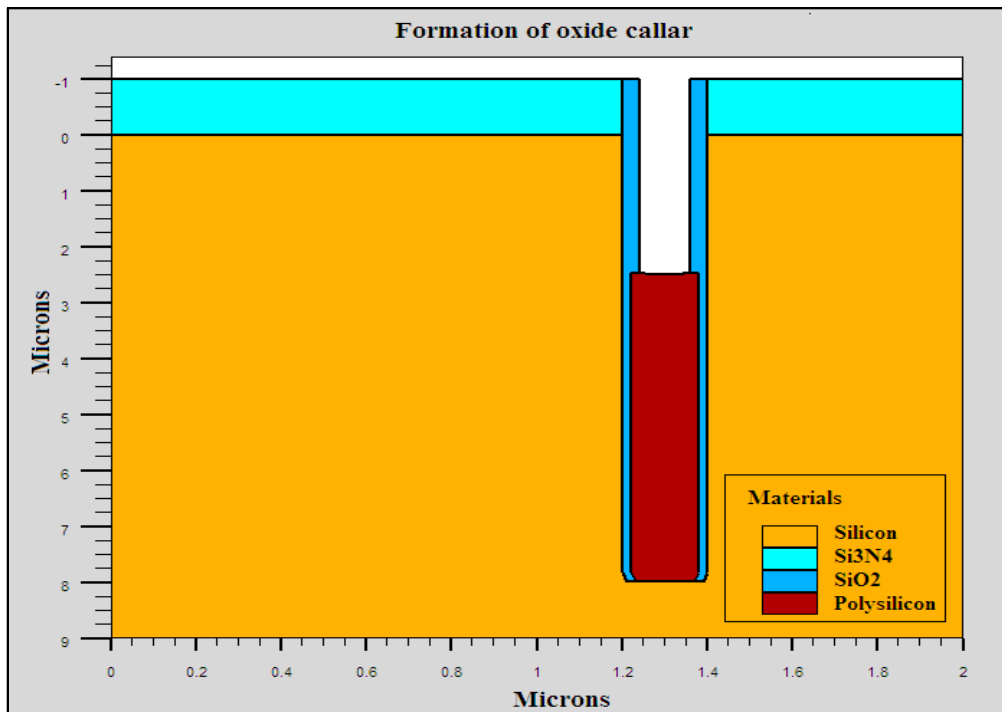


Figure 4.4:- Formation of oxide collar.

A second polysilicon deposition is performed to fill the trench and cover the surface. This polysilicon is then polished (typically by chemical mechanical polishing) so that it is planar with the surface. Polysilicon is subsequently etched so that it is recessed in the trench. The collar is then etched so that top portion of the trench sidewall is again exposed. This is shown in figure (4.5).

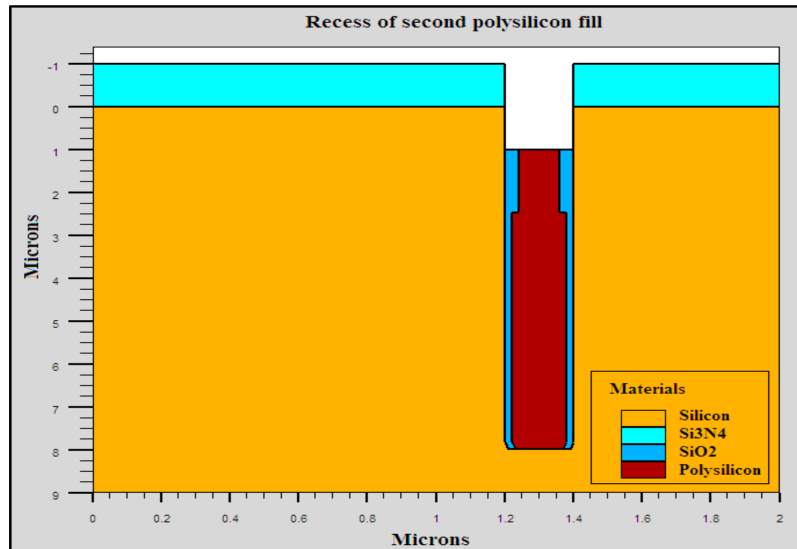


Figure 4.5:- Recess of second polysilicon fill.

A third polysilicon deposition is performed, followed by planarization and another etch process so that polysilicon is recessed below the substrate surface. The trench capacitor is subsequently covered at its top surface by the shallow trench isolation (STI) which also overlaps a portion of trench capacitor as shown in figure (4.6).

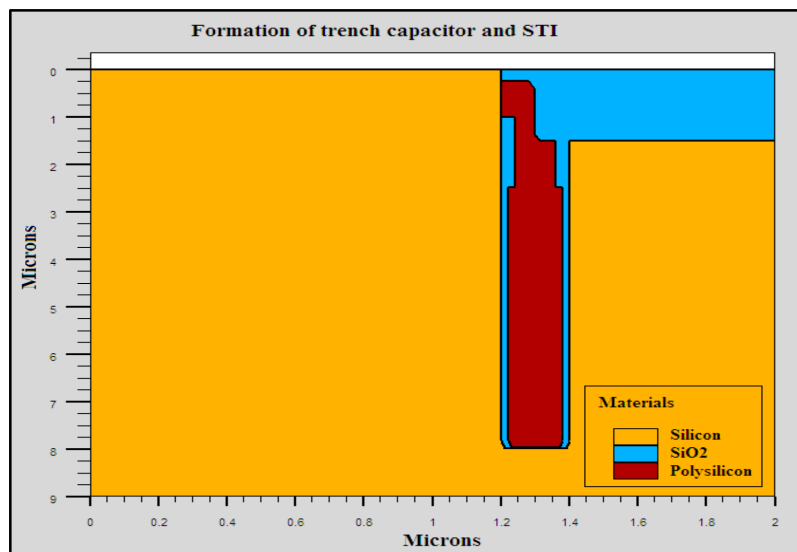
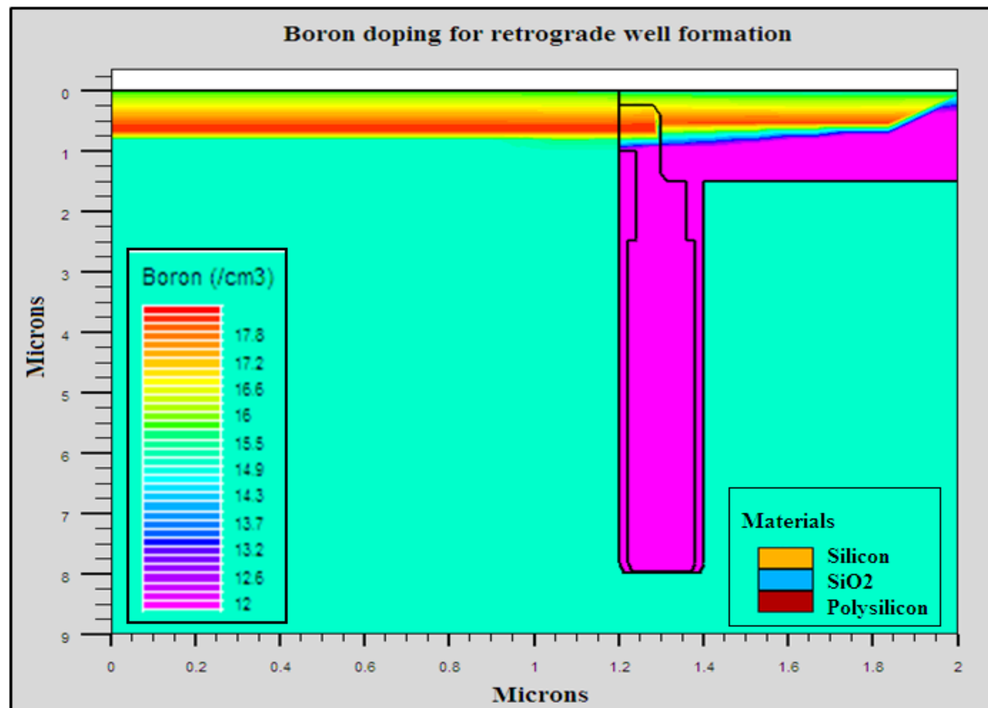
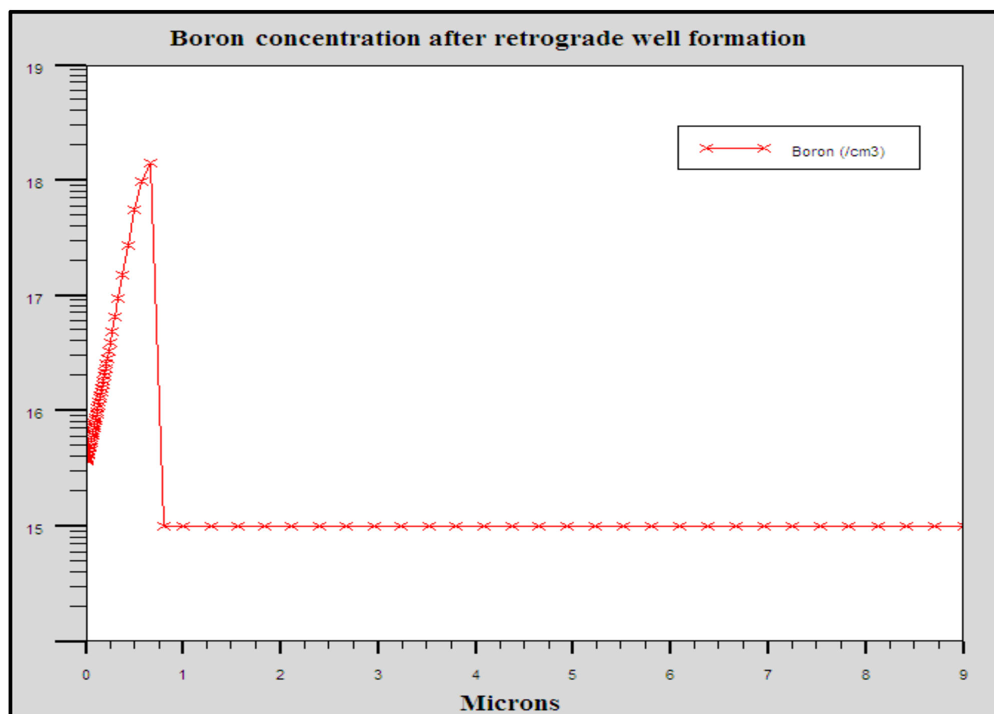


Figure 4.6:- Formation of trench capacitor and shallow trench isolation.

An nMOSFET is to be formed adjacent to the trench capacitor. Figure (4.7(a)) shows the doping profile of substrate after retrograde well formation and figure (4.7(b)) shows the concentration of Boron along of the depth of the substrate.



(a)



(b)

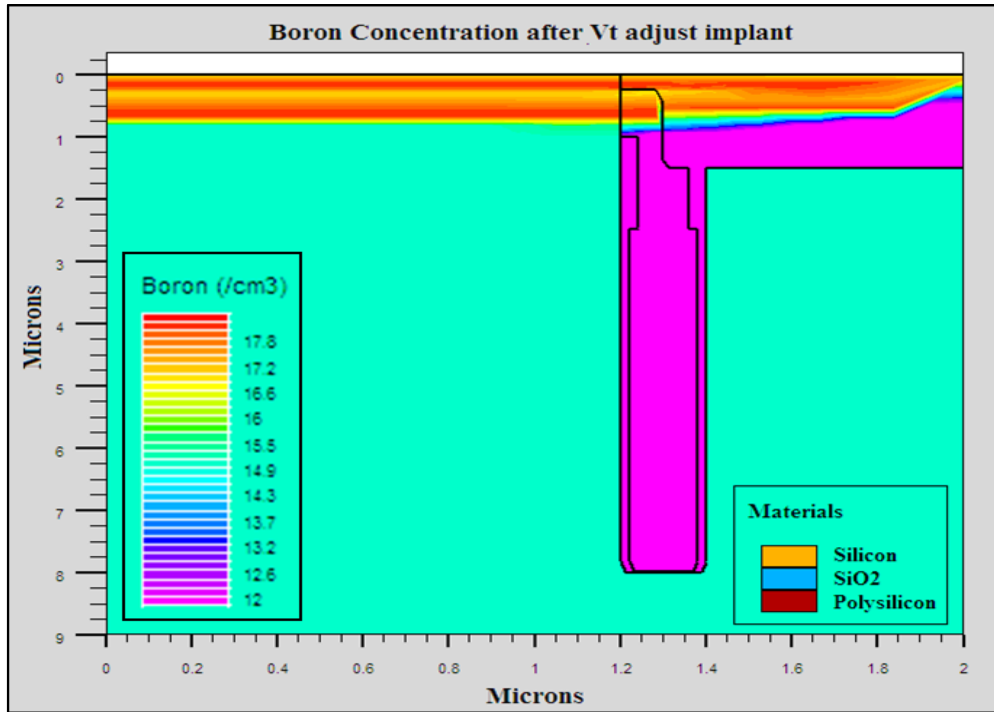
Figure 4.7:- Formation of retrograde well in the substrate.

However, scaling-down has resulted in the development of short-channel effects. Therefore, it has been attempted to extinguish the short-channel effect by providing the channel region of MOSFET with a semiconductor region having the same conductivity type as that of the substrate (retrograde-well). The p-well is formed using a boron implantation with an energy of E = 200 keV and a dose of D=3 x $10^{13}/cm^2$. By applying such a moderate energy implant without drive-in diffusion, a retrograde well profile is achieved which suppresses parasitic bipolar devices (latch-up immunity). Figure (4.7(b)) shows that the concentration of boron has increased a few nanometers below the silicon surface. Thus a low-to-high doping channel profile is observed. The low concentration at channel region improves surface mobility.

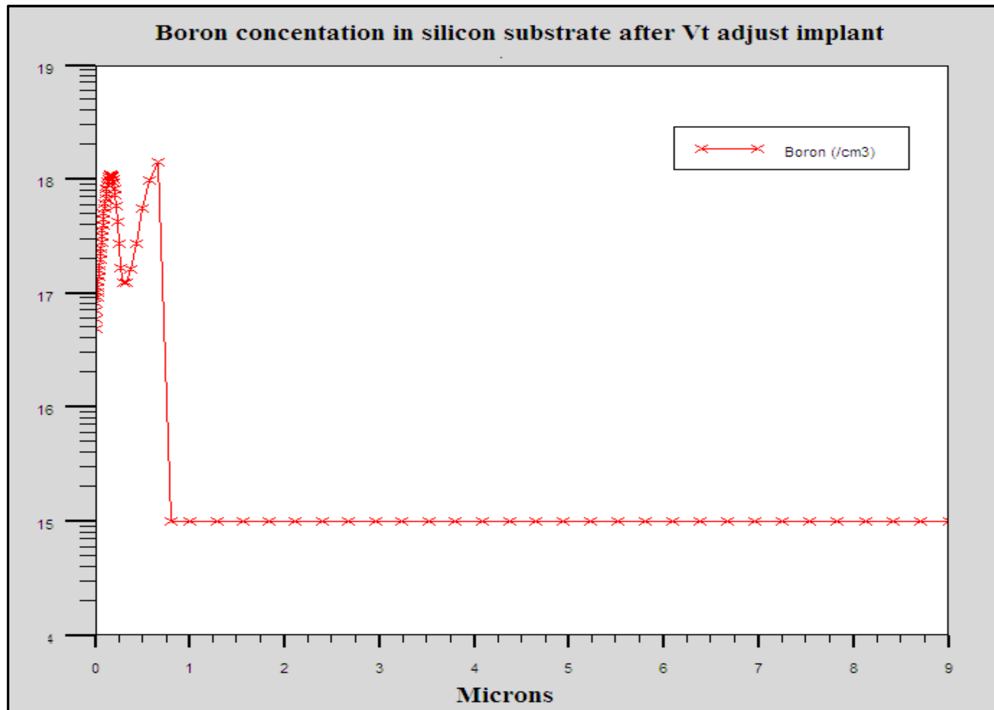
Gate oxide of thickness 3.5 nm is grown over the substrate. The gate oxide thickness has been approximately linearly scaled with channel length, with the ratio of gate length L_G to the thickness t_{ox} being maintained at approximately 45 [18].

A low-energy ion implantation (boron implant of about 1.5 x $10^{13}/cm^2$ with energy E = 45 keV) is done to slightly dope the well region below the gate electrode to adjust the V_{th} of the device. Primarily this implant, in combination with the previous p-well implant and the gate oxide capacitance, determines the final threshold voltage value. The figure (4.8(a)) shows the doping profile after this V_{th} adjust implant and figure (4.8(b)) shows the boron concentration along the depth of the substrate.

A heavily n+ doped polysilicon layer of 200 nm thickness is deposited. Now after the lithography step the gate contact with length 0.18 μm is formed. Now the step of polysilicon doping and the formation of lightly doped drain is done. Implant dose and energy needed to be selected carefully and controlled to produce a desired graded drain junction. Arsenic implant with 1 x $10^{15}/cm^2$ dose at 30 keV was used. In comparison to conventional devices, light doped drain in modern devices reduces the electric field at the oxide semiconductor interface. The Light doped drain reduces the GIDL and hot electron effects.



(a)



(b)

Figure 4.8:- Doping profile of substrate after V_{th} adjust implant.

Halo implant is done with opposite dosage to the lightly doped source and drain implant (LDD). At this state, a boron implantation of 3 x $10^{13}/cm^2$ with 30° tilt angle and with energy E = 15 keV is performed. The boron doping profile after halo implant is shown in figure (4.9). The halo implant is deeper both vertically and laterally than the LDD. This effectively suppresses the high electrical field regions causing punchthrough. They are also called punchthrough stoppers because they improve punchthrough resistance.

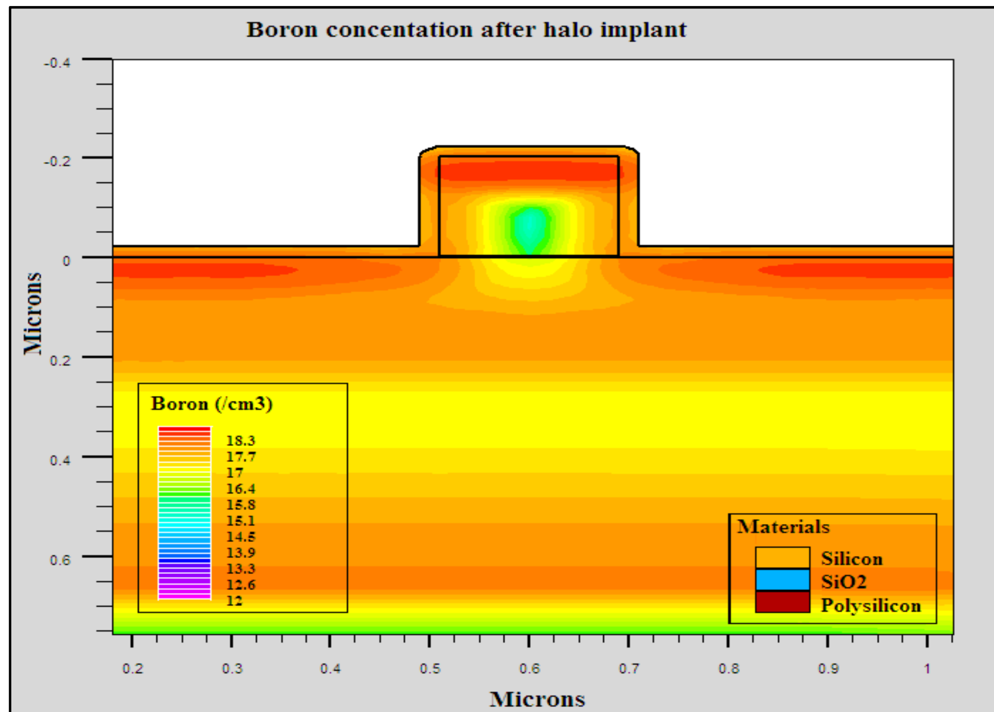


Figure 4.9:- Boron doping profile after halo implant.

The thin oxide layer is removed and about 100 nm oxide is deposited by low pressure chemical vapor deposition (LPCVD) and anisotropically etched so that oxide spacers at the polysilicon edges are created. A high dose (5 x $10^{15}cm^{-2}$) of arsenic is implanted with 60 keV to build the low resistance of source and drain regions. To activate the implanted dopants without too much diffusional redistribution, a rapid thermal annealing (RTA) step is necessary; e.g. 1 seconds at about 1000°C for nMOSFET. Figure (4.10) shows the net doping profile after annealing. Finally, the nMOSFET is contacted using aluminum. A buried strap region (formed by diffusion of dopants from polysilicon) connects the drain with the polysilicon. The collar serves to prevent the charge leakage from the capacitor at the buried strap junction. The complete structure of trench capacitor DRAM cell is shown in figure (4.11).

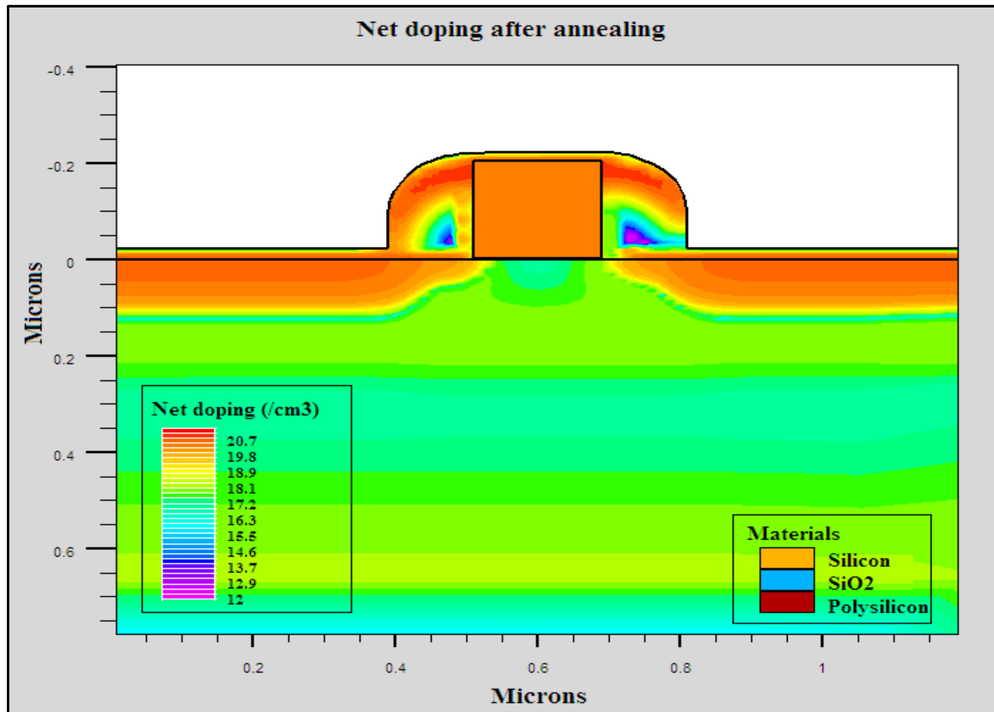


Figure 4.10:- Net doping profile after annealing.

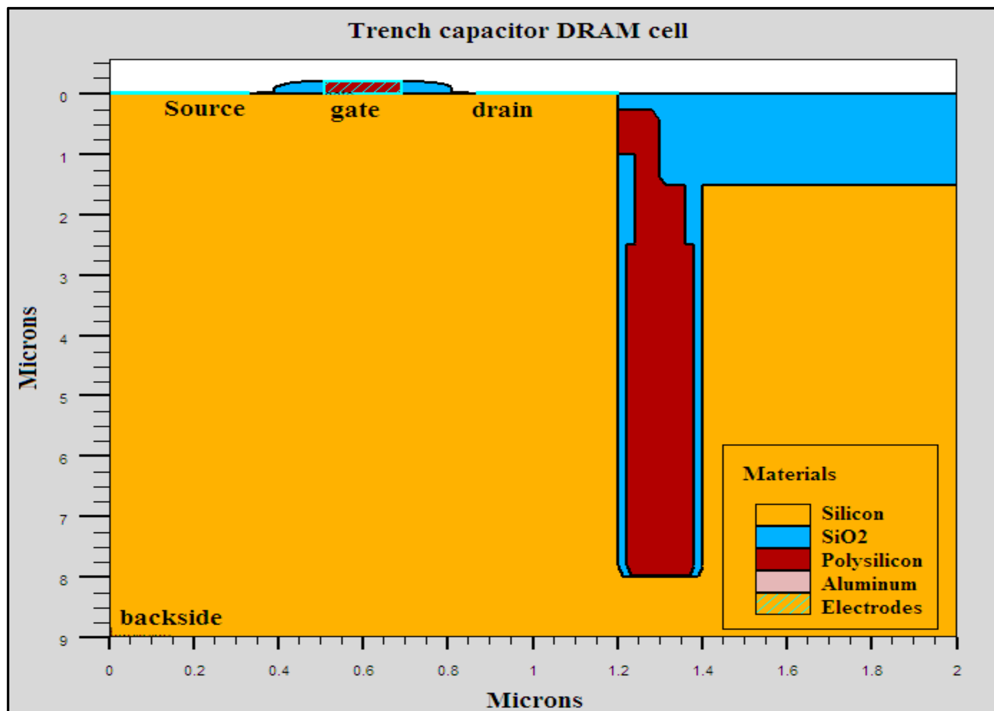


Figure 4.11:- The complete structure of trench capacitor DRAM cell.

CHAPTER 5

RESULTS AND CONCLUSION

5.1. Results and discussion

Process and device simulation of trench capacitor DRAM cell has been carried out using the ATHENA/ATLAS packages of SILVACO. The trench capacitor DRAM cell structure has been generated using ATHENA and then it is used as input to device simulator ATLAS for its electrical characterization [19]. With device simulation, the effect of process parameters and its variation on the electrical characteristics of DRAM cell is studied.

Process-level techniques such as LDD implantation, retrograde well and halo doping in DRAM cell in 0.18 µm technology has been employed for leakage reduction. Leakage reduction from 256 nA to 2.25 nA has been achieved.

Figure 5.1 below shows the subthreshold characteristics of a DRAM cell in which nMOSFET is fabricated with LDD implantation done and without halo doping and retrograde well. At $V_g = 0$ V, off-state leakage $I_{ds} = 256$ nA has been observed.

Figure 5.2 below shows the subthreshold characteristics of a DRAM cell in which nMOSFET is fabricated with LDD implantation, retrograde well and without halo implantation. At $V_g = 0$ V, off-state leakage $I_{ds} = 238$ nA has been observed. Leakage reduction has been achieved.

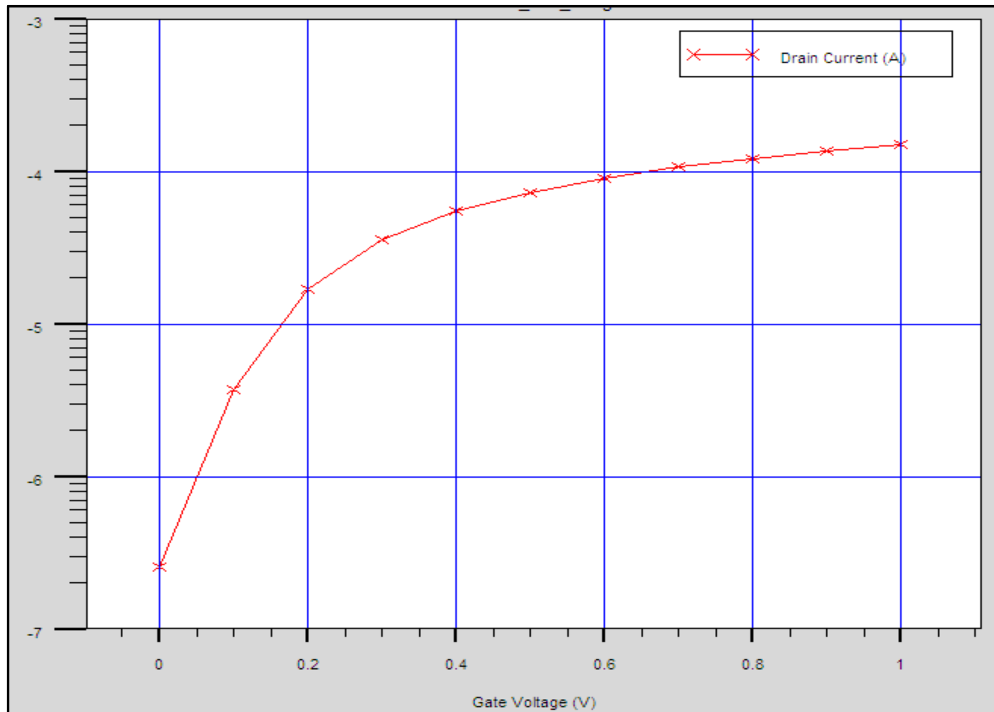


Figure 5.1:- Subthreshold characteristics of DRAM cell with LDD (without retrograde well and halo implantation)

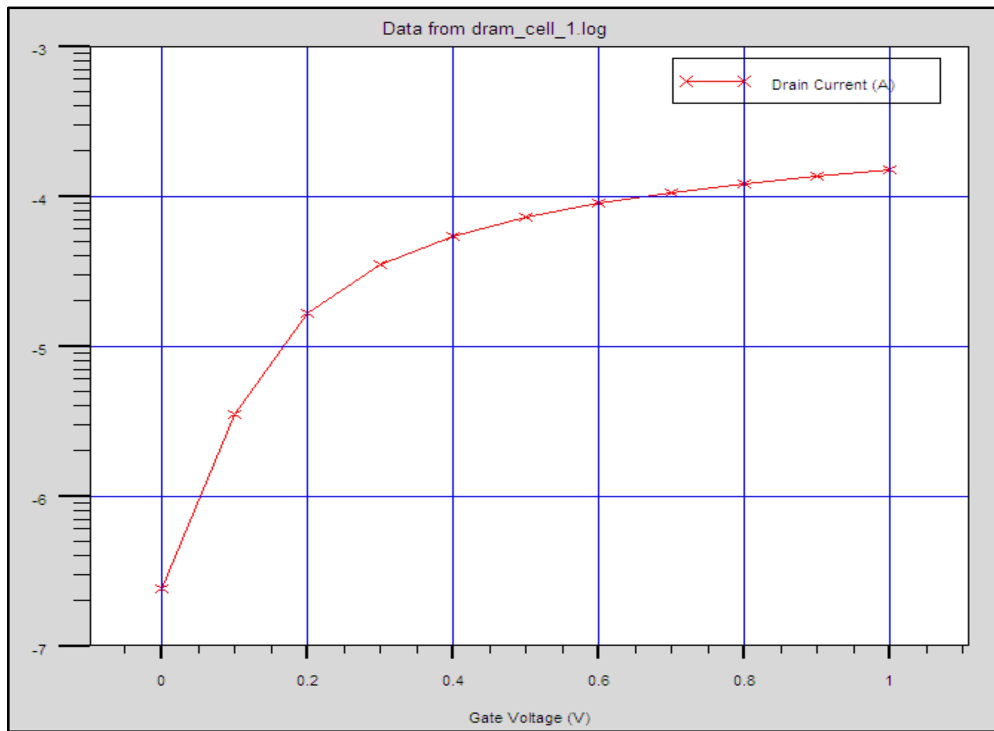


Figure 5.2:- Subthreshold characteristics of DRAM cell with retrograde well and LDD (without halo implantation)

Figure 5.3 shows the subthreshold characteristics of a DRAM cell in which nMOSFET is fabricated with retrograde well, LDD and halo implantation done. At $V_g = 0$ V, off-state leakage $I_{ds} = 2.25$ nA has been observed. We can see that the leakage is drastically reduced with halo implantation.

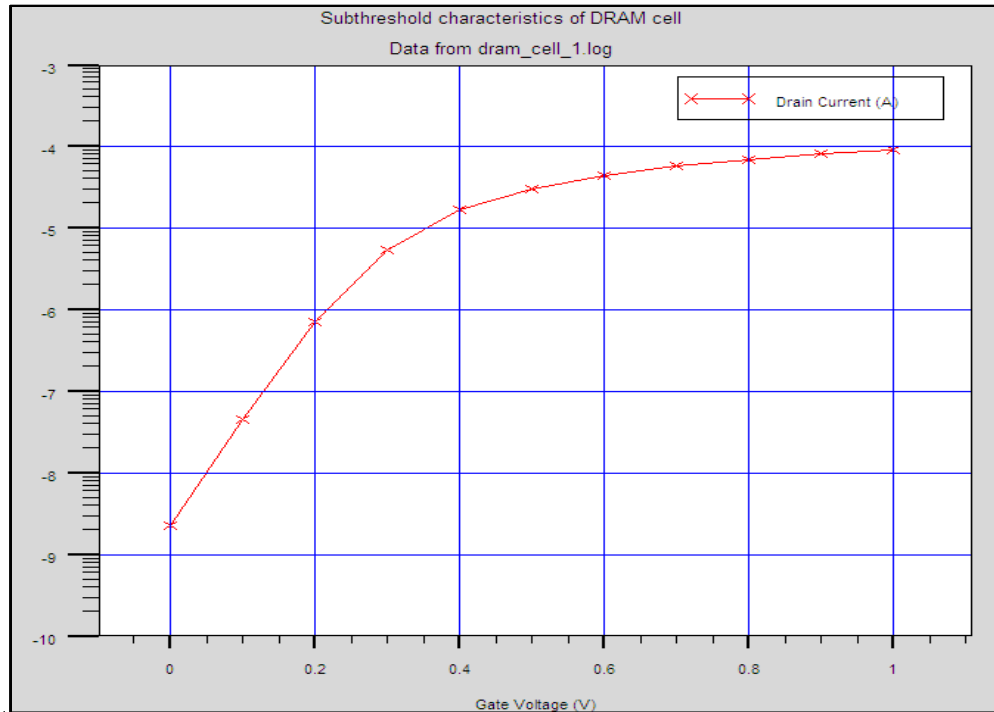


Figure 5.3:- Subthreshold characteristics of DRAM cell with retrograde well, LDD and halo implantation.

All the results obtained have been summarized is the following table (5.1).

Table 5.1:- Leakage current in trench capacitor DRAM cell after employing process-level techniques.

Process techniques employed	Leakage observed
Light doped drain (LDD)	256 nA
LDD, retrograde well	238 nA
LDD, retrograde well, halo implant	2.25 nA

5.2. Conclusion and Future Scope

Leakage current analysis of trench capacitor dynamic random access memory cell has been done. For the analysis, 0.18 µm submicron nMOSFET as access transistor and the conventional trench capacitor as storage device have been designed.

For this a low leakage access transistor has been modeled and simulated using ATHENA/ATLAS packages of SILVACO. It has been found that process techniques such as light doped drain (LDD) implantation, retrograde well doping and halo doping in DRAM cell in 0.18 µm technology reduce the leakage current significantly. A leakage current of 256 nA is found in LDD DRAM while by using LDD, retrograde well and halo doping simultaneously in DRAM cell we found that the leakage current is reduced to 2.25 nA. Thus these leakage reduction techniques reduce the power consumption of DRAM cell efficiently.

As we are approaching the physical limits of MOSFET scaling and the demand for higher density, efficiency and performance DRAMs is continuously increasing, research on other alternatives to conventional device and process is accelerating. Some of the techniques under investigation are Silicon On Insulator (SOI), strained silicon, high-k gate dielectric and gate stack engineering, and non-planar MOSFET.

This thesis work features the use of bulk technology for the fabrication of DRAM cell. The benefits of SOI over bulk technology can be investigated as future work. The DRAM using the floating body cell can be shown to have a potential to be used as a high-speed memory and as an alternative to bulk-Si DRAM because SOI devices are have the advantages in high speed, low power dissipation and extended scalability. The research on SOI DRAM can be done as future work.

REFERENCES

[1]. Kiyoo Itoh, "The History of DRAM Circuit Designs – At the Forefront of DRAM Development", IEEE SSCS News, pp. 27 – 31, 2008.

[2]. Kiyoo Itoh, M. Horiguchi, H. Tanaka, "Ultra-Low Voltage Nano-Scale Memories", Springer, Aug. 2007.

[3]. Jan M. Rabaey, Anantha Chandrakasan, Borivoje Nikolic, "Digital Integrated Circuits", Prentice Hall.

[4]. H. Sunami, "The Role of the Trench Capacitor in DRAM Innovation", Technical Literature, Solid State Circuits, January, 2008.

[5]. Nicky C. C. Lu, "Advanced Cell Structures for Dynamic RAMs", IEEE Circuits and Devices Magazine, pp. 27 – 36, January, 1989.

[6]. Samual K. Moore, "Masters of Memory", IEEE Spectrum, January, 2007.

[7]. Jyi-Tsaong Lin, Mike Chang, "A New 1T DRAM with Enhanced Floating Body Effect", Proceedings of the 2006 IEEE International Workshop on Memory Technology, Design and Testing, 2006.

[8]. Nauman Z. Butt, Muhammad A. Alam, "Scaling Limits of Double-Gate and Surround-Gate Z-RAM Cells", IEEE Transactions on Electron Devices, Vol. 54, No. 9, September 2007.

[9]. Serguei Okhonin, "In Search of a Better DRAM", Innovative Silicon, Inc.

[10]. Zhichao Lu, Jerry G. Fossum, Vishal P. Trivedi, "A Novel Two-Transistor Floating-Body/Gate Cell for Low Power Nanoscale Embedded DRAM", *IEEE Transactions on Electronics Devices,* Vol. 55, No. 6, pp. 1511-1518, June 2008.

[11]. J.G. Fossum et al., "A Novel Two-Transistor Floating-Body Memory Cell", *IEEE International SOI Conference Proceedings*, 2007, pp. 105-106.

[12]. Minchen Chang et. al., "Impact of Gate-Induced Drain leakage on Retention Time Distribution of 256 Mbit DRAM With Negative Wordline Bias", IEEE Transactions on Electron Devices, Vol. 50, pp. 1036 – 1041, 4 April, 2003.

[13]. A. Weber et. al., "Data Retention Analysis on Individual Cells of 256 Mb DRAM in 110 nm Technology ", Proceedings of ESSDERC, pp. 185 – 188, Grenoble, France, 2005.

[14]. Kaushik Roy et. al., “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits”, Proceedings of the IEEE, Vol. 91, No. 2, pp. 305 – 327, 2003.

[15]. Silvaco, “ATHENA User’s Manual, 2D Process Simulation Software”, Silvaco, August, 2004.

[16]. The International Technology Roadmap for Semiconductors: 2001 Edition.

[17]. K.P. Muller et. al., “Trench Storage Node Technology for Gigabit DRAM Generations”, IEDM, pp. 507 – 509, 1996.

[18]. Stanley Wolf, “Silicon Processing for the VLSI Era, Vol. 4 – Deep-Submicron Process Technology”, LATTICE Press, 2002.

[19]. Silvaco, “ATLAS User’s Manual, Device Simulation Software”, Silvaco, September, 2004.