1993

A single-supply, low-voltage, programmable SONOS memory array for high density EEPROM and semiconductor disk applications

Harikaran Sathianathan
Lehigh University

AUTHOR:

Sathianathan, Harikaran

TITLE:

A Single-Supply, Low-Voltage, Programmable SONOS Memory Array; For High Density EEPROM and Semiconductor Disk Applications

DATE: May 30, 1993

A SINGLE-SUPPLY, LOW-VOLTAGE, PROGRAMMABLE SONOS MEMORY ARRAY FOR HIGH DENSITY EEPROM AND SEMICONDUCTOR DISK APPLICATIONS

by
Harikaran Sathianathan

A Thesis
Presented to the Graduate Committee
of Lehigh University
in Candidacy for the Degree of
Master of Science
in
Electrical Engineering

Lehigh University
May 21, 1993

This thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science.

May 20, 1993
(date)

Marvin H. White
Thesis Advisor

Alastair McAulay
Chairman of Department

Acknowledgements

This thesis is dedicated, with love and affection, to my parents, Dr. and Mrs. Sathianathan, and to my brothers, Raghavan, Ramesh and Kumar for their continuing support, encouragement and inspiration.

I am deeply indebted to my advisor, Dr. Marvin H. White, for overseeing my research and flavoring it with the depth and breadth of his expertise. It has been both a privilege and an honor to work with Amit Banerjee, Ron Paulsen and Bill Wagner as I have benefitted from our camaraderie and collaboration on issues ranging from the trivial to the technical. Malcolm Chen deserves recognition for his fraternal guidance during the initial stages of my research and his continuing interest in the current developments. I am grateful to Dr. Yin Hu, and especially to Peggy French, for their availability, patience and eagerness in discussing SONOS theory and fabrication. I would like to thank Zhigang Ma and Dr. Susan Gong for sharing with me their unique perspectives on some fundamental device issues. I am also obliged to Dr. Richard Siergiej and Dr. Sukyoon Yoon for setting a standard of academic excellence and professionalism for me to follow. I am beholden to Paul Orphanos for structuring our computing network and Linda Dreisbach for her concern and for structuring my graduate school paperwork. Ray Filozof and Floyd Miller are thanked for their their contagiously friendly demeanor and their willingness to help. I must not forget to acknowledge Dr. M.K. Hatalis, Greg Sarcona, Fuyu Lin, Apostolos Voutsas, Ji-Ho Kung, Alex Mattachini, Su-Heng Lin and Aida Cabassa for providing a cordial and entertaining, yet productive work atmosphere during those long nocturnal hours at the lab. Finally, I would like to mention the "summer students" Matt Martin, Neil Cohen, Lou Bellanca and Chris Kulp for reminding me of myself many eons ago.

Contents

Acknowledgements iii

List of Figures vi

List of Symbols ix

Abstract 1

1 Introduction 2

1.1 A Perspective on Electronic Memories 2

1.2 A Comparison of Contemporary EEPROMs 5

1.3 Scope of this Thesis 14

2 The SONOS Memory Device 16

2.1 Basic Structure and Operation 16

2.2 Testing 20

2.2.1 Erase/Write 20

2.2.2 Retention 25

2.2.3 Endurance 27

2.3 Memory Window Design Considerations 28

3 The Memory Array Design 31

3.1 Approach 31

3.2 Basic Operation 32

3.3 Input/Output Signals 36

3.4 The Addressing/Decoding Circuitry 37

3.4.1 Row Decoder 37

3.4.2 Column Decoder 40

3.4.3 Sense Amplifier 42
3.5 Array Layout 47
3.6 Circuit Simulations 48

4 Conclusions **61**
4.1 Conclusions and Recommendations for Future Work 61

References **64**

A CMOS/SONOS Fabrication Sequence **67**

B Sample SPICE Code **70**
B.1 Sample SPICE code 70

Vitae **75**

List of Figures

1.1 A tree structure depiction of electronic memories 3

1.2 An n-channel NAND FLOTOX cell architecture. [1] 6

1.3 A Split-Gate cell architecture.[2] 8

1.4 A cross-section of an ETOX cell. [3] 9

1.5 A cross section of a Triple Poly cell. [3] 10

2.1 The cross section of a SONOS transistor. 17

2.2 The energy band diagram and equivalent circuit of the SONOS device during Erase and Write modes 18

2.3 The n-channel SONOS Erase/Write curves for two different tunnel oxide thicknesses are presented. Note that thinner devices are able to program faster because for a given programming voltage, they will have a higher electric field across the tunnel oxide. Both devices had a 49$\mathring{A}$ nitride and a 40$\mathring{A}$ blocking oxide and were tested with $\pm$5V. [4] 21

2.4 An Erase/Write curve on an n-channel SONOS device for 9V programming. In this measurement the device was actually programmed to saturation with a 1ms pulse. Contrast this with the 5V programming on a similar device in Fig.2.3 where 11s pulses were not sufficient to saturate the device because the smaller electric field across the tunnel oxide resulted in slower programming.[5] 24

2.5 The retention characteristics of a SONOS device ($x_{ot} = 18\mathring{A}, x_N = 49\mathring{A}, x_{ob} = 40\mathring{A}$) are presented for 5V programming.[5] 26

2.6 The threshold voltage of an n-channel SONOS device (x_{ot}=20$\mathring{A}$, x_N= 73$\mathring{A}$, x_{ob}=54$\mathring{A}$is presented versus number of Erase/Write cycles. The excessive cycling increases charge trapping at the Si-SiO_2 interface and prevents complete discharging, thus, increasing the threshold voltage. [6] 27

2.7 A hypothetical optimum dynamic memory window under 5V programming. 30

3.1 The SONOS memory array and its operating conditions 33

3.2 The cross section of a memory cell during its operational modes . . . 34

3.3 Circuit schematic for memory array 38

3.4 Row decoder circuit 39

3.5 Column Decoder circuit 41

3.6 The transimpedance amplifier sensing scheme for analog output voltage. 43

3.7 The cross-coupled inverter/flip-flop sense amplifier 44

3.8 Row decoder test pattern 49

3.9 Column decoder test pattern. 50

3.10 Sense amplifier test pattern 51

3.11 Memory array layout using the Mentor Graphics GDT software and Lehigh-SFC technology rules. 52

3.12 A timing diagram which demonstrates the flip flop sense amplifier's ability to convert the read current into a digital compatible voltage. Output A is inverted to form Output B because the former toggles to 0V when the memory device is in the high conduction state, which we have previously defined to be the logic 1 state. 55

3.13 A timing diagram which shows the read current charging the input capacitor of the flip-flop sense amplifier to a digital voltage. Because SPICE requires all the device parameters to be predefined prior to simulation, the memory threshold is preset to high conduction state independent of the programming modes. Thus, the sense amp input always charges to 5V during the Read cycle. 56

3.14 A timing diagram showing the Bit Line outputs. When the column address signal, A1, strobes high, BL1 is selected and undergoes the appropriate Write, Erase and Read voltages while depicting the different Inhibit voltages when A1 strobes low. The converse is true for BL2. 57
3.15 Timing diagram for the Select Lines. When A0 strobes high, row 1 is selected and X1 activates the series PMOS select transistors only when a memory cell in row 1 is read, else it inhibits their conduction by maintaining 5V on the memory gates. X2 behaves similarly when A0 strobes low. 58
3.16 Timing diagram for the Word Lines. When A0 is high, row 1 is selected and WL1 depicts the voltages for the Erase, Read and Write operations, respectively, while WL2 inhibits row 2. The strobing of A1 during this time determines which particular cell of row 1 is addressed. When A0 is low, row 2 is selected while row 1 is inhibited, hence WL2 reflects the applied Erase, Read and Write voltages whereas WL1 shows the different inhibit voltages. 59

List of Symbols

abstol	absolute current error tolerance
A	some constant
B	modified Fowler Nordheim constant
chgtol	minimum charge tolerance
C_1	input capacitance (F)
C_2	output capacitance (F)
C_{eff}	effective gate capacitance (F/cm^2)
C_{OX}	gate oxide capacitance (F/cm^2)
E_{OT}	electric field across the tunnel oxide (V/cm)
$E_{OT}(0)$	initial electric field across the tunnel oxide (V/cm)
$E_{OT}(t)$	time dependent tunnel oxide field (V/cm)
E_T	attenuation constant (V/cm)
g_{DS}	drain to source conductance (S)
gmin	minimum conductance
h	Planck's constant (J-sec)
$\hbar$	Planck's constant divided by 2π (J-sec)
I_D	drain current (A)
$I_{D_{m4}}$	transistor m4 drain current (A)
I_{DS}	drain to source current (A)
I_{Read}	read current (A)
J_{OT}	tunnel oxide current density (A/cm^2)
L	channel length (μm)
m^*_{OX}	effective mass in the oxide (kg)
N_B	bulk doping density (cm$-$3)
pivtol	minimum value of pivotal matrix element

q	electronic charge (C)
Q_F	fixed oxide charge (C)
Q_N	total charge in the nitride (C)
$Q_N(0)$	initial charge in the nitride (C)
r_e	erase decay rate (V/log(sec))
r_w	write decay rate (V/log(sec))
reltol	relative error tolerance per iteration
R_f	feedback resistance (Ω)
t	time (sec)
t_{RD}	read delay time (sec)
t_{read}	read time (sec)
trtol	transient error tolerance
V_{DD}	supply voltage (V)
V_{DS}	drain to source voltage (V)
V_{GS}	gate to source voltage (V)
V_{in}	input voltage (V)
$V_{in}(t)$	time dependent input voltage (V)
V_{mw}	memory window (V)
V_{out}	output voltage (V)
V_P	programming voltage (V)
V_r	read voltage (V)
V_{TD}	transistor m4 threshold voltage (V)
V_{TH}	threshold voltage (V)
$V_{TH}(Erase)$	erase threshold shift (V)
$V_{TH}(Write)$	write threshold shift (V)
V_{Tm}	memory transistor threshold voltage (V)
$V_{Tm(min)}$	minimum memory transistor threshold voltage (V)
$V_{turn-on}$	turn on voltage (V)
W	channel width (μm)
$\overline{x}$	charge centroid (cm)

x_{eff}	effective dielectric thickness (cm)
x_N	nitride thickness (cm)
x_{ob}	blocking oxide thickness
x_{ot}	tunnel oxide thickness
β	current gain (A/V^2)
β_m	memory transistor current gain (A/V^2)
ΔV_{TH}	threshold voltage shift (V)
ϵ_{ob}	permitivitty of SiO_2 (F/cm)
ϵ_{ox}	permitivitty of SiO_2 (F/cm)
ϵ_{Si}	permitivitty of Si (F/cm)
ϕ_1	SiO_2 energy barrier height (eV)
ϕ_2	Si_3N_4 energy barrier height (eV)
ϕ_B	bulk potential (B)
ϕ_{GS}	gate to source potential difference (V)
ϕ_S	surface potential (V)
τ	tunneling time constant (sec)

Abstract

To achieve the muti-gigabit density required for semiconductor disk implementation, the existing EEPROM memory cells must be scaled down in size as well as voltage. Floating gate EEPROMs require high on-chip programming voltages (~15V), which, over repeated Erase and Write operations, can lead to single bit failures. SONOS memories, however, can exhibit data retention and excellent endurance to many Erase/Write cycles (~10^6 cycles) while requiring a small cell area and low programming voltage (~5-10V).

Previous work on discrete SONOS memory transistors has been extended to the integrated circuit level by designing a small SONOS memory array to serve as a test vehicle for high density EEPROMs and NVRAMs. This array is novel in that it is powered by a single 5V supply which achieves the bipolar voltages necessary for memory programming and erasure via switching circuitry in the addressing decoders. The memory elements of the array are chosen to be p-channel SONOS devices in order to minimize the destructive effects of electron injection into the gate as well as to provide memory well isolation in our n-well CMOS fabrication process.

The test memory array consists of a 2 x 2 memory cell structure that effectively mimics the operation of larger EEPROMs. Each of the cells can be accessed individually, to undergo the Write and Read operations, while the Erase operation affects an entire sector. The logic of the initial design was verified with the SPICE circuit simulation software and then a mask layout was performed with the Mentor Graphics GDT software package. A post-layout equivalent circuit netlist was extracted and exercised through a rigorous series of timing simulations utillizing SPICE. With the completion of the design and verification stage, the circuit is ready to be fabricated in-house using our custom n-well CMOS/SONOS processing sequence.

Chapter 1

Introduction

1.1 A Perspective on Electronic Memories

The ideal electronic memory should offer fast random access to data, long-term data retention, preservation of data integrity (over repetitive write/read cycles), high density, low cost per bit, low power consumption, a single supply low programming voltage, and a high degree of radiation tolerance (for military and space applications). The difficulty in creating a compact, light-weight memory that can optimize all these criteria has led to the development of application specific memories. Generally, these can be grouped into short term (or volatile) memories and long term (or nonvolatile) memories.

Volatile memories provide fast data retrieval and storage at the expense of data retention. Semiconductor memories such as Dynamic Random Access Memories (DRAMs) and Static Random Access Memories (SRAMs) dominate the volatile memory market. DRAMs combine high speed memory access and programming with high density but are volatile and the data must be refreshed periodically. SRAMs provide faster memory access and do not need to be refreshed, but they require the presence of a constant power supply in order to maintain the integrity of the data. To improve the data retention some efforts have been aimed towards utilizing battery back-up or a nonvolatile memory (NVRAM) to preserve the data in case of power failure.

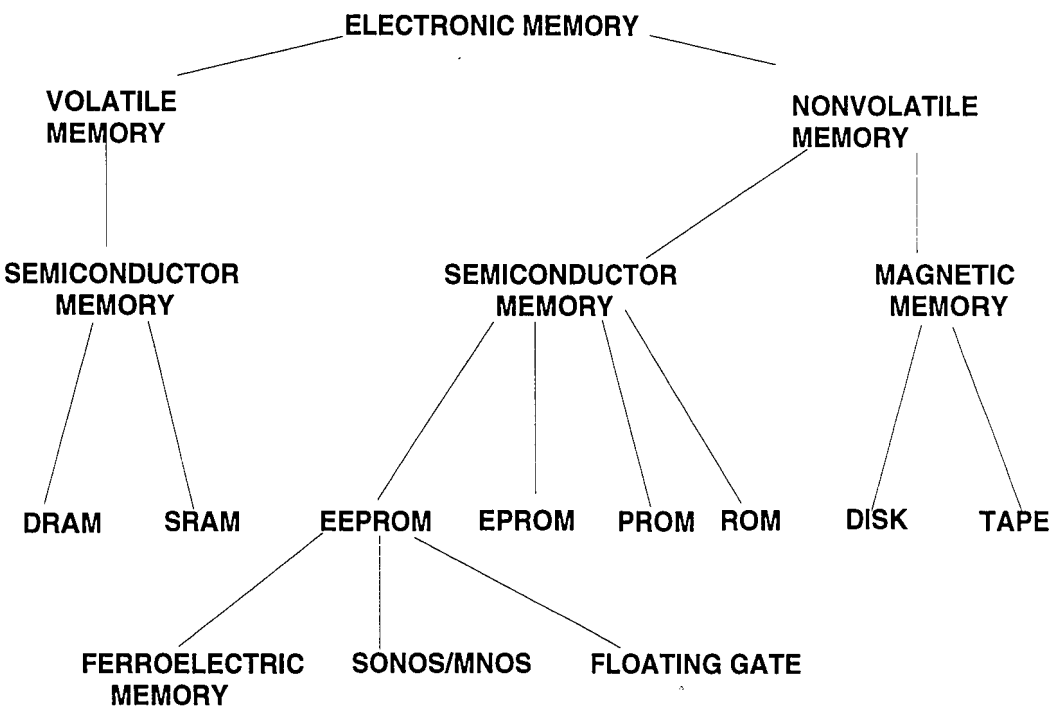


Figure 1.1: A tree structure depiction of electronic memories

Nonvolatile memories are considerably slower than DRAMs or SRAMs, but they are able to retain the data for long periods of time without the presence of a power supply. Unlike the fast access memory market, the nonvolatile memory market contains several technologies vying for dominance. Historically, the leading technology has been magnetic core, specifically in the form of magnetic disks and tape. Though this approach provides excellent data retention and integrity for current applications, it has certain shortcomings which may hinder their dominance in the near future. As computers evolve into more portable forms such as notebook or even hand-held computers, it becomes vitally important to develop light weight, low power, reliable disk drives. Magnetic disk drives comprise of mechanical components which are heavier, less reliable and require more power than solid state components. The most significant constraint on magnetic technology is the inherent latency in storing or retrieving the first byte of data. Either process requires a lag time in which the disk drive's head moves to the appropriate track and waits until the spindle spins the appropriate sector of the track into position before any information can be accessed. The effect of this lag will become far more pronounced as computer data processing and transfer speeds increase. The difficulties involved in modifying magnetic memories to alleviate this bottleneck, combined with the inherent random data access of the competing semiconductor memories, may impede the decision to choose a magnetic based technology as the nonvolatile data storage in laptop computers.

Semiconductor Read Only Memories (ROMs) present the logical alternative to magnetic disks because they are reliable solid state memories which dissipate little power and can randomly access data with less lag time than magnetic memories. The simplest ROMs are ROM and Programmable ROM (PROM), which allow data to be written once only. Electrically Programmable ROMs (EPROMs) improve over PROMs by allowing repetitive data alterations. However, while the Write operation may be performed electrically, the Erase operation requires exposure to ultraviolet radiation, necessitating the physical removal of the chip from the computer. Electrically Erasable Programmable ROMs (EEPROMs) can be both programmed and

erased electrically, making them the most versatile of all the ROM devices. Despite their versatility, EEPROMs have not been able to capture the nonvolatile semiconductor memory market because EPROMs are cheaper and generally have a smaller cell size. However, improvements in EEPROMs such as the recent developments on the oxide-nitride-oxide (ONO) based memory devices and the introduction of Flash EEPROMs, will soon thrust EEPROMs into prominence [3, 7].

1.2 A Comparison of Contemporary EEPROMs

In order to supplant magnetic disks, there is ongoing research to make solid state disks viable. EEPROMs are the most likely candidates to evolve into semiconductor disks, but the question that now remains is, "which particular type of EEPROM will be adopted?" Generally, EEPROMs can be divided into floating gate MOSFETs and floating trap MOSFETs. Floating gate devices store digital data as a function of the "on" (logic 1) and "off" (logic 0) states of the transistor by the presence or absence of charge on a conducting layer of polysilicon embedded within the gate dielectric. Charge storage in floating trap devices occurs in the actual gate dielectric itself due to its trap rich composition.

Of the many floating gate structures, only the Flash memories will be discussed because they promise to yield the high densities to achieve semiconductor disk implementation. Flash EEPROMs utilize the same basic concept as regular floating gate devices, but the cell topology is such that the need for select transistors is minimized or even eliminated by employing sector or bulk erasure of the memory. There are essentially four different approaches to Flash EEPROMs:

The NAND Floating Gate Tunnel Oxide or NAND FLOTOX [1] (Fig.1.2) is a popular EEPROM structure which consists of a thin tunnel oxide above the drain, a floating polysilicon gate, an inter-poly oxide, followed by the control gate. Prior

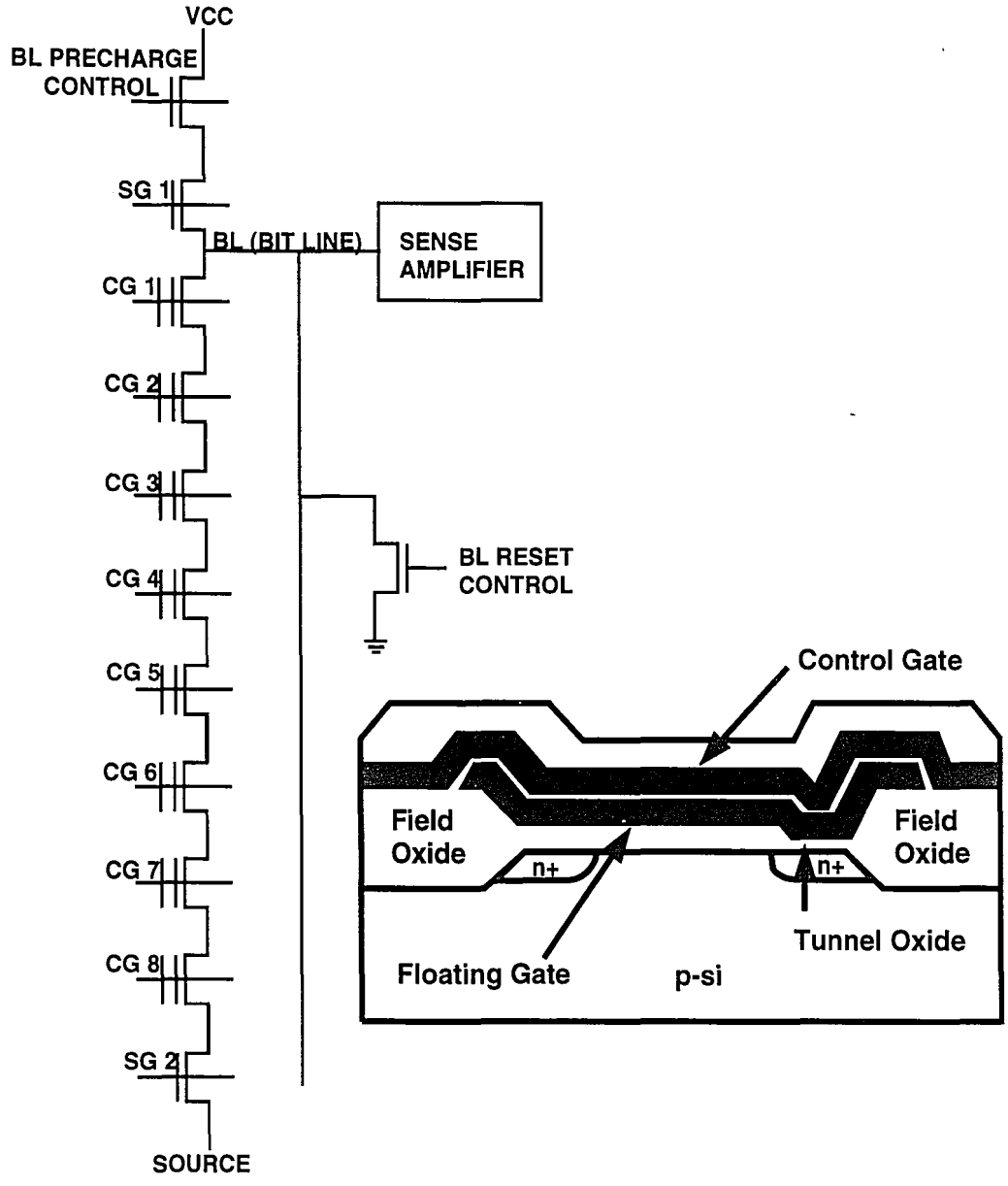


Figure 1.2: An n-channel NAND FLOTOX cell architecture. [1]

to the development of Flash memory, the FLOTOX cells were connected in parallel to the Bit Line (NOR structure), requiring one series address (or select) transistor per memory cell. To enhance the density, the NAND FLOTOX stores each byte of data in eight serially connected FLOTOX transistors which are sandwiched between two select transistors. The first address transistor is connected to the Bit Line and selects the data to be read, while the second prevents current from propagating during programming. Overall, the NAND cell architecture effectively reduces the memory cell size from one select transistor per bit to 1/4 selector transistor per bit. All the n-channel NAND FLOTOX cells are simultaneously set to the "off" state by applying 17V [1] to all of the control gates while grounding the Bit Lines, resulting in negative charge accumulation on the floating gates, turning the devices off (logic 0). Programming the data is done one bit at a time. The control gates of the unselected transistors and the Bit Line are set at 22V, while the control gate of the unselected transistor is set to 0V. Thus, the unselected transistors merely transfer the bit line voltage to the selected transistor, programming the logic state. The Read operation is performed by precharging the Bit Line, grounding the desired control gate and applying 5V to the control gates of the unselected cells. If the selected device is in depletion mode, then it generates a current that passes through the unselected transistors to a sense amplifier to convert the data into a digital output.

The Split-Gate [2, 8, 9, 7] is a Flash structure (Fig.1.3) whose cell area is minimized by incorporating the select (or address) transistor with the memory transistor by overlapping their gates. Thus, the floating gate only controls half the channel and will not turn on the device unless an appropriate bias is put on the control gate as well. Chip erasure for an n-channel split-gate cell occurs when -11V is applied to the Word Lines, 5V to the Bit Lines and the drains are floated, allowing discharge of the floating gate via Fowler-Nordheim tunneling [2]. Single bit programming (Write) requires 18V on the particular Word Line and 0V on the particular Bit Line to facilitate electron tunneling to the floating gate. The other memory cells must be inhibited during this process. As with the FLOTOX, data is read by detecting

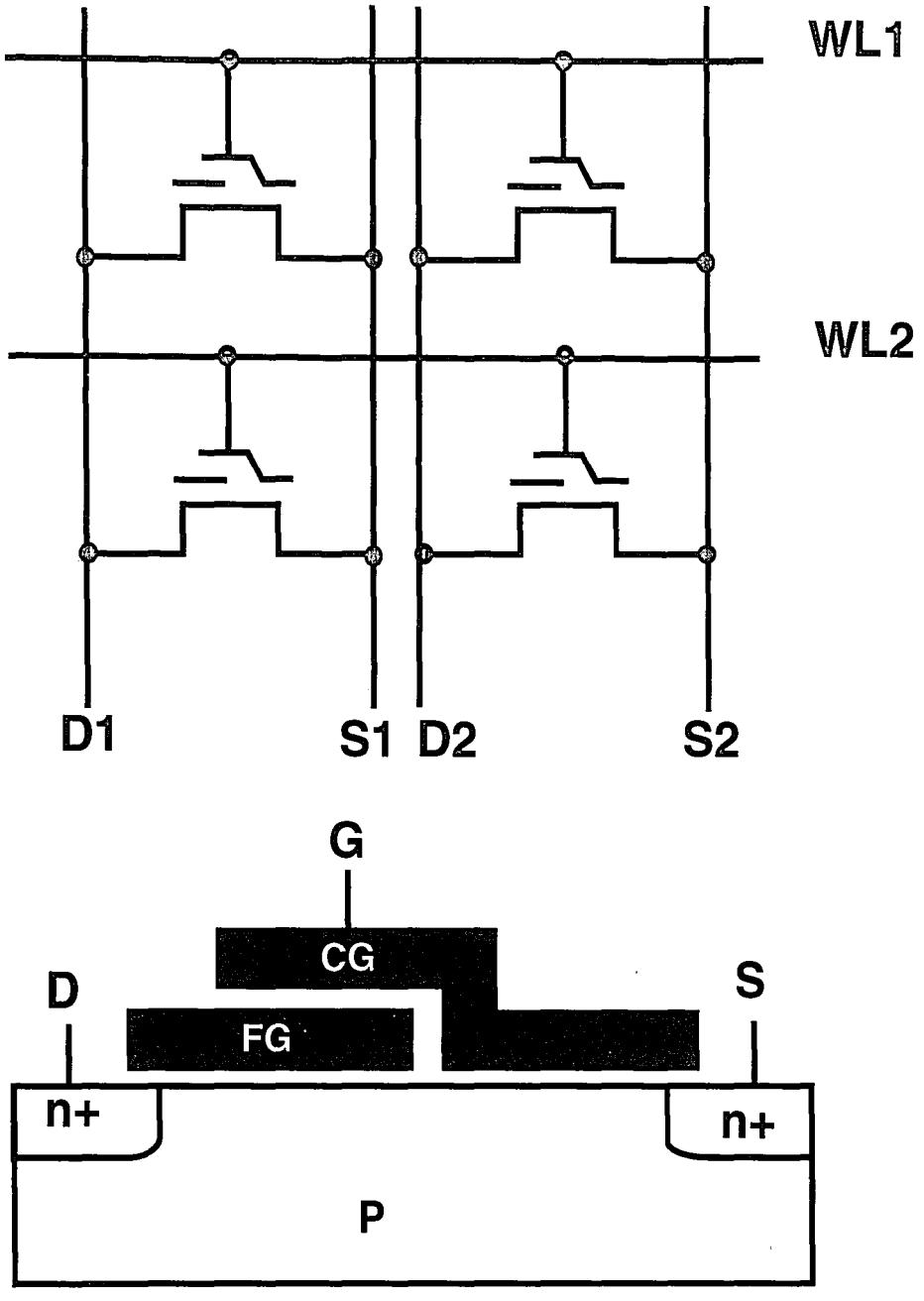


Figure 1.3: A Split-Gate cell architecture.[2]

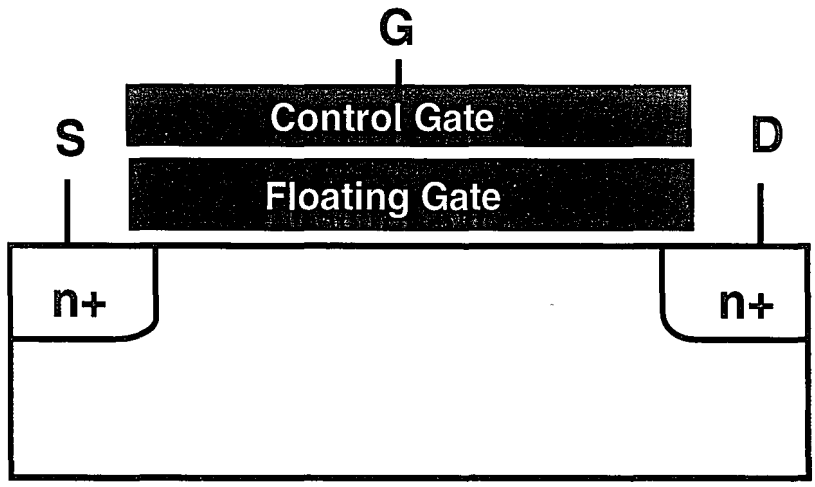


Figure 1.4: A cross-section of an ETOX cell. [3]

the amount of current generated by the selected device.

The EPROM Tunnel Oxide or ETOX [10, 3, 7] is, perhaps, the most widely used Flash EEPROM (Fig.1.4). It programs by channel hot electron injection into the floating gate and erases by discharging via Fowler-Nordheim tunneling to the source. Bulk erasure for n-channel ETOX occurs by applying a high voltage to all the sources while grounding all the select lines [11]. Writing a cell requires grounding the Bit Line and the source while the Word Line receives 13V, enabling electrons to tunnel to the floating gate and program the device. During the Read operation, a bias is applied to make the device source a current (depending upon its "on" or "off" state) without resulting in erroneous programming or erasure.

The Triple Poly [12, 13, 3, 7] structure (Fig.1.5) employs an erase gate in addition to the contol and floating gates, increasing the cell size and making it more unpopular in the market. Data programming occurs via hot channel hot electron injection when

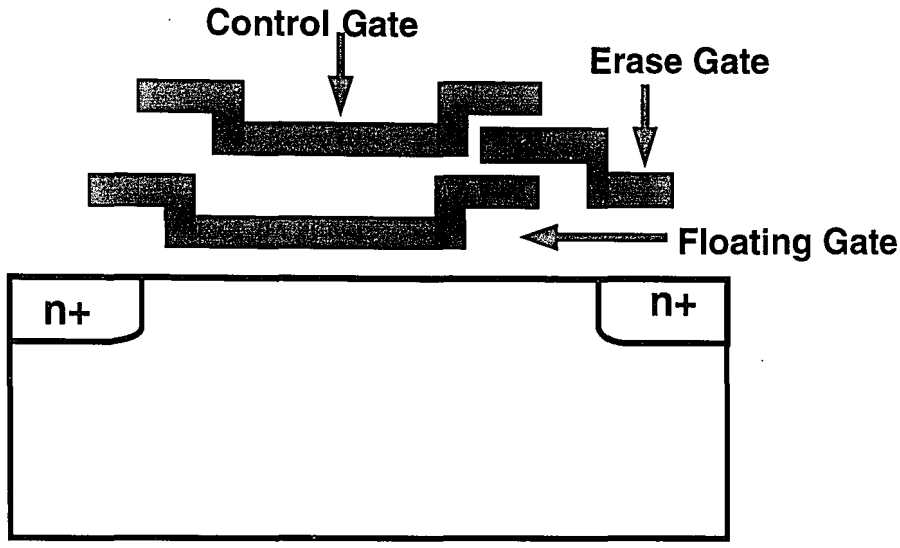


Figure 1.5: A cross section of a Triple Poly cell. [3]

20V is applied to the control gate and 16V to the drain. To erase the data 30V is applied to the erase gate [12], forcing the floating gate to discharge to the erase gate.

While these four Flash memories have comparable erase/write times and exhibit excellent retention, the most likely candidates for semiconductor disks seem to be the NAND FLOTOX and ETOX. The high density provided by the ETOX, the low programming voltage of the NAND FLOTOX combined with the relatively simple fabrication sequence for both structures lend to their viability as semiconductor disks.

The relative ease of fabrication and long term retention of the floating gate Flash memories are offset by several serious disadvantages, namely, the high programming voltage and the poor endurance to erase/write cycles. Despite the claim of 5V off-chip compatibility, the input voltage must be charge pumped up to levels necessary to program the device. The repetitive pulsing of the high programming voltages on the device terminals during the Write/Read/Erase cycles may disturb or even damage the regular MOSFETs in the peripheral circuitry. In addition, the resultant

high electric fields also limit the endurance of the floating gate device itself to less than the desired one million cycles. Excessive cycling can result in charge unintentionally trapped in the oxide (trap-up) or breakdown in the gate and tunnel oxides, resulting in single-bit failures, one of the most serious reliability problems of floating gates. [3, 7, 14, 15]

Since most manufacturers are more concerned with the excellent retention of the floating gate approach rather than its questionable endurance, it remains the most popular EEPROM structure. However, as device scaling is pushed deeper into the submicron regime to achieve higher densities, the high programming voltage related endurance problems will become more pronounced and make floating gate semiconductor disks more unreliable. Hence, there is sufficient reason to look at other technologies, such as the floating trap devices, to produce the nonvolatile memories of the future.

Within the floating trap family of devices, there are two approaches to high performance EEPROMs. Of these, the ferroelectric devices are still in their early stage of development. Data is stored by polarizing the dipoles within the novel PZT gate dielectric. [7] In principle, these memories should provide the fastest data access and program times since they do not require charge to be physically transferred to and from the gate dielectric. However, the basic ferroelectric device, unless used in conjunction with a MOSFET, is a a destructive readout, rather than a Nondestructive Readout (NDR) as in the floating gate and floating trap devices. Thus, each readout cycle requires a subsequent refresh cycle to maintain the integrity of the data. Moreover, there are still some processing problems that must be overcome before ferroelectric memories become a serious competitor in the market.

Perhaps the ideal technology for the future of nonvolatile memory is the silicon nitride floating trap memory. Silicon nitride memories operate under a similar principle to the floating gate devices, namely, the charge stored within the gate dielectric

of the MOSFET determines the logic value of the data. The primary difference between the two, however, is the floating gate stores charge on a conducting layer while the nitride memories store it in discrete traps within the silicon nitride layer of the gate dielectric. This approach renders the device relatively invulnerable to pinhole defects in the oxide since only those traps located spatially near the defect will discharge. Thus, the nitride memories such as Metal-Nitride-Oxide-Silicon (MNOS) and polySilicon-Oxide-Nitride-Oxide-Silicon (SONOS) are inherently more reliable than the floating gates because they display a gradual deterioration in operating features with cycling rather than an abrupt failure of the cell demonstrated by floating gate EEPROMs [15]. In addition, they require smaller programming voltages. Of the two nitride memories, SONOS provides the smaller programming voltage and cell size, because the addition of a "blocking" oxide permits the nitride to be scaled while maintaining the same internal electric fields in the dielectrics.

While the development of MNOS and SONOS memories was actively pursued in the 1970s, they failed to seize the commercial memory market. This had less to do with the performance of the devices than the issues of fabrication. The processing of most floating gate structures (e.g. FLOTOX) is a straight forward variation of the regular MOSFET process sequence while the dielectric growth of MNOS and SONOS devices is more complex and requires careful control of the various layers. Controlling the growth and quality of the ultra-thin tunneling oxide and the silicon nitride layer proved too much of a gamble for many industries to invest, and thus, MNOS/SONOS was abandoned by many U.S. commercial memory manufacturers in favor of the floating gate. However, it is important to note that the U.S. aerospace and defense industries still pursue SONOS research because the radiation hardness of SONOS nonvolatile memories is ideal for military and space applications.

Nevertheless, the 1980s proved a fruitful decade for MNOS/SONOS research as a handful of institutions (e.g. Hitachi, Citizen, Lehigh University, Westinghouse, Sandia and Hughes) made remarkable advances in improving the device. In the late 1970s, Chen [16] gave birth to the SONOS transistor by introducing a blocking

oxide to the MNOS structure to scale the device and reduce gate leakage. The early SONOS devices required 25V programming, but Agarwal [17] was able to develop a scaled version that required 10V. Then in 1983, Suzuki [18] developed the first low voltage SONOS device with 6V programming. However, it had only a 0.6V memory window between the Erase and Write states which lead to poor retention of data. Chao [19] was able to improve the static memory window to 6.5V with 8V programming, while Libsch and Roy [20], his successors, demonstrated a 5V programmable SONOS device with 3.8V static memory window, though the work was mainly done on capacitors. Finally, French [4] and Hu [21] were able to extend these results to transistors and improve the retention and dielectric scaling.

SONOS research in the last decade has not been limited to the single transistor level. Two major corporations, Hitachi and Citizen, manufacture and sell MNOS and SONOS EEPROMs. Hitachi produces EEPROMs using an MNOS/SONOS technology based upon the scaling guidelines put forth by Minami and Kamigaki [22] while Citizen has recently produced a 1Mbit SONOS EEPROM [23] that employs a split-gate cell architecture similar to that used in Flash EEPROMs. The circuit operates using a 9V differential voltage generated by a 5V supply and a -4V supply. Single bit programming occurs when 5V is applied to the Word Line and -4V to the Chip Select and Bit Line, resulting in Fowler-Nordheim tunneling of electrons from the channel to the nitride. In order to optimize the Read current, the memory window of the SONOS devices was designed such that the Write time is 100us and the Erase time is 10ms. However, since the chip is page (128 bytes) written and block (512 bytes) erased, the effective Write time becomes 0.1μs per bit and the effective Erase time becomes 2.4μs per bit. The memory is block erased when all the bit lines are set to 5V and the word line to -4V. To read a particular cell, the address gate turns the series transistor on and forces a current proportional to the memory transistor's logic state to the sense amplifier.

1.3 Scope of this Thesis

Over the last decade, researchers at Lehigh University's Sherman Fairchild Center for Solid State Studies have made significant contributions to the state of the art of SONOS nonvolatile memory devices. By experimenting with thinner dielectrics, novel dielectric compositions and buried channel structures, we have been able to optimize device endurance, improve retention and reduce the programming voltage to TTL levels (5V). We believe we can duplicate these results onto the integrated memory circuit level by developing a small solid state SONOS EEPROM test circuit.

This memory array will be novel in three ways. **First**, it will be powered by a true 5V supply without the need to charge pump to higher programming voltages. **Second**, the power supply will be unipolar, forcing the circuit to switch the gate and bulk voltage polarities in order to achieve the negative programming voltage. **Finally**, the memory array elements will be p-channel SONOS, not the traditional n-channel devices so commonly adopted by industry. This serves the twofold purpose of isolating the memory elements from the address transistors by placing them in a separate n-well (for our n-well technology) as well as allowing us the opportunity of studying the unique properties of p-channel devices in memory arrays, such as improved Erase/Write characteristics since electron injection to the gate will be minimized.

In the next chapter, the discrete SONOS transistor will be discussed with particular emphasis on their implementation in practical EEPROMs. The operating modes will be explained explicitly and an analytical treatment of the physics involved in writing the device will be reviewed. Some experimental results on discrete n-channel SONOS devices will be presented to demonstrate the adaptability of our existing memory testing station to test integrated memory arrays.

In the third chapter, the actual memory array will be discussed. First, there will

be a treatment of the cell topology and operation. Next, the individual addressing circuit components will be analyzed with special emphasis on the sense amplifiers. Finally, the particular characteristics of the memory circuit is analyzed with SPICE simulation software and the memory circuit is created with a CAD mask layout employing Mentor Graphics GDT.

In the fourth chapter, the conclusions of this study are summarized and the current status of this project is presented. In addition, suggestions for related research issues that merit further investigation are listed.

Two sets of appendices are presented in the final portion of this thesis. The first lists the steps of the custom n-well CMOS/SONOS processing sequence that will be employed in our laboratory to actually fabricate the circuit while the second appendix contains a sample SPICE code based on a netlist gnerated from the layout.

Chapter 2

The SONOS Memory Device

2.1 Basic Structure and Operation

A SONOS memory device may be visualized as a MOSFET with a modified gate dielectric that is ideal for charge trapping. The ability to trap charge enables data storage to be expressed as a function of the conductivity of the transistor under an applied bias. This is achieved by shifting the intrinsic threshold voltage of the device to one polarity when the device is written and to the opposite polarity when the device is erased, defining the logic "0" and "1" states. The threshold voltage of a PMOS transistor can be expressed as:

$$V_{TH} = \phi_{GS} - 2|\phi_B| - \frac{\sqrt{4\epsilon_{Si} q N_B |\phi_B|}}{C_{OX}} - \frac{Q_F}{C_{OX}} \tag{2.1}$$

where ϕ_{GS} denotes the gate to source potential difference, $|\phi_B|$ is the bulk potential, N_B is the bulk doping, ϵ_{Si} is the dielectric permitivitty of silicon, Q_F is the fixed charge stored in the gate oxide (as an unintended byproduct of the fabrication), and C_{OX} is the capacitance of the gate oxide.

Note, the fixed charge (usually positive) stored within the gate dielectric comprises a significant term in this expression. The SONOS structure permits additional charge storage into (or removal from) the gate dielectric when the device terminals are appropriately biased. This is facilitated by the novel oxide-nitride-oxide (ONO) dielectric sandwich structure which encourages charge tunneling into and subsequent trapping within the nitride region when a sufficiently large gate to bulk bias

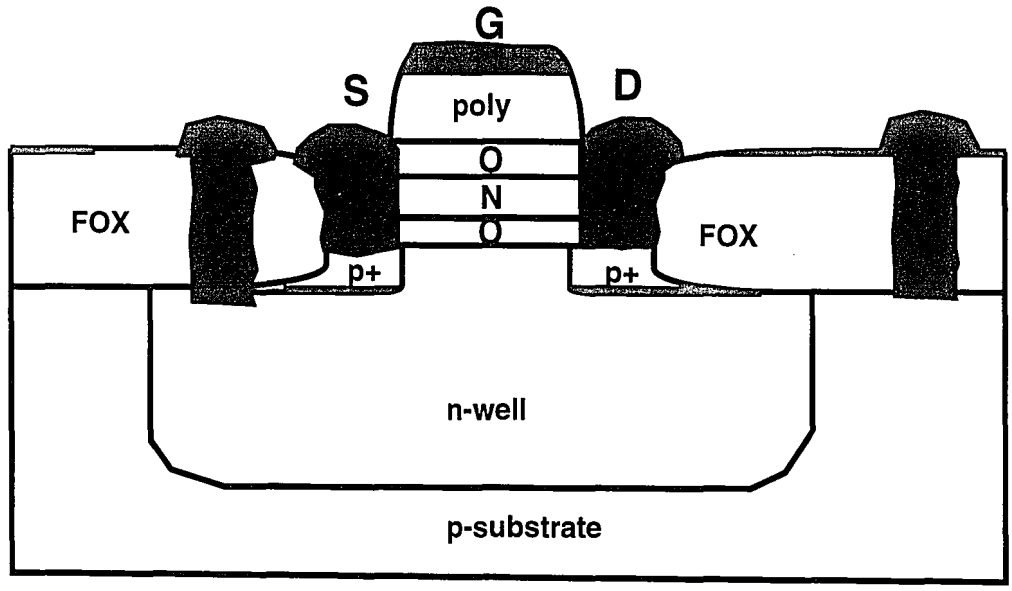


Figure 2.1: The cross section of a SONOS transistor.

is applied. To reflect the incorporation of the nitride charge, the threshold voltage of a p-channel SONOS transistor can be expressed as:

$$V_{TH} = \phi_{GS} - 2|\phi_B| - \frac{\sqrt{4\epsilon_{Si} q N_B |\phi_B|}}{C_{eff}} - Q_N\left(\frac{x_{OB}}{\epsilon_{OB}} + \frac{x_N - \overline{x}}{\epsilon_N}\right) \qquad (2.2)$$

where $\overline{x}$ is the charge centroid in the nitride, x_N is the nitride thickness, ϵ_N is the nitride permitivitty, x_{ob} is the blocking oxide thickness, ϵ_{ob} is the oxide dielectric permitivitty, and Q_N is the total charge in the nitride, which may be positive or negative, depending upon the previous programming operation.

The **Write** operation for a SONOS memory device is defined by the necessary threshold voltage shift that results in the low conduction state of the transistor. This requires biasing the device into inversion and allowing the carriers in the channel to tunnel into the nitride and shift the threshold voltage to the enhancement mode or normally-off state. For a p-channel SONOS device, the **Write** operation results in an inversion layer of holes in the channel with subsequent hole tunneling into the

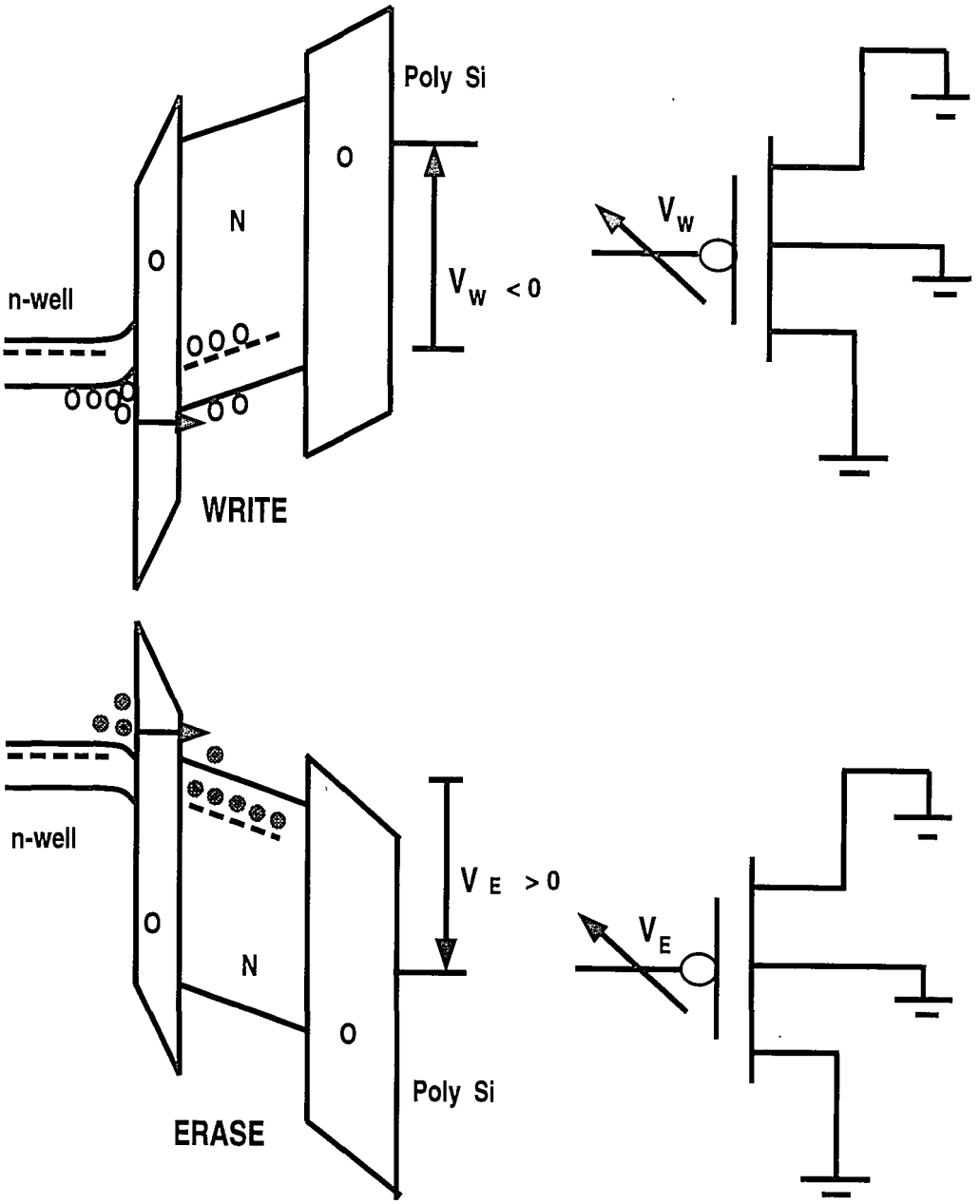


Figure 2.2: The energy band diagram and equivalent circuit of the SONOS device during Erase and Write modes

nitride, shifting the threshold voltage in a negative direction. The **Erase** operation is defined by biasing the device into accumulation and allowing majority carrier tunneling from the substrate to set the high conduction state of the transistor. For a p-channel device, the electron tunneling shifts the threshold voltage in a positive direction. If this shift is sufficiently large, the device may be in the depletion mode or normally-on state.

During the **Read** operation a constant current ($\sim 10\mu$A) is sourced through the discrete SONOS device with the gate and drain grounded, forcing the source to gate voltage to settle at a value close to the threshold voltage. This so-called "turn-on" voltage includes the threshold voltage value but also incorporates the effect of the current level. Examining the expression for the Read current, we note that

$$I_{Read} = \frac{\beta}{2}(V_{GS} - V_{TH})^2 \tag{2.3}$$

where

$$\beta = \mu_p(\frac{W}{L})C_{OX} \tag{2.4}$$

Since V_{GS} is also $V_{turn-on}$, the equation can be rewritten as:

$$V_{turn-on} = V_{TH} + \sqrt{\frac{2I_{DS}}{\beta}} \tag{2.5}$$

where the second term in the above equation represents a slight deviation from the threshold voltage ($\sim$ 0.1V for most of the tested SONOS devices). Another concern regarding this readout scheme is that the turn-on voltage may induce erroneous programming. However, if the Read pulse is sufficiently smaller than the programming pulses the read disturbance may be considered minimal.

2.2 Testing

Our strength in testing and characterizing SONOS transistors lies in their treatment as electronic memories, and not merely as regular MOSFETs. Conventional transistor testing techniques such as current-voltage (I-V) and capacitance-voltage (C-V) provide information about certain device characteristics, but do not indicate their performance in actual EEPROMs. In order to test the memory properties of SONOS we have a custom built software controlled Erase/Write circuit that generates voltage pulses to the device terminals corresponding to the **Read**, **Write**, **Erase** and **Idle** modes of a memory array. By varying the pulse amplitude, period, and total number of cycles, we can extract data for the memory window, retention and reliability. The timing delays of the applied patterns are similar to the timing delays of the programming signals in an integrated memory array. Thus, by observing the transient response of the device we are able to reasonably predict its performance on the circuit level.

The three primary tests that are performed on SONOS devices are: **Erase/Write**, **Retention** and **Endurance**.

2.2.1 Erase/Write

Erase/Write measurements depict the maximum threshold voltage memory window, and cross-over time of the SONOS device. The memory window represents the difference in threshold voltage shift when the device is erased and then written while the cross-over time is a figure of merit used to indicate the programming speed by the intersection of the Erase and Write curves.

For a typical Erase/Write measurement the device is initialized into the Erase

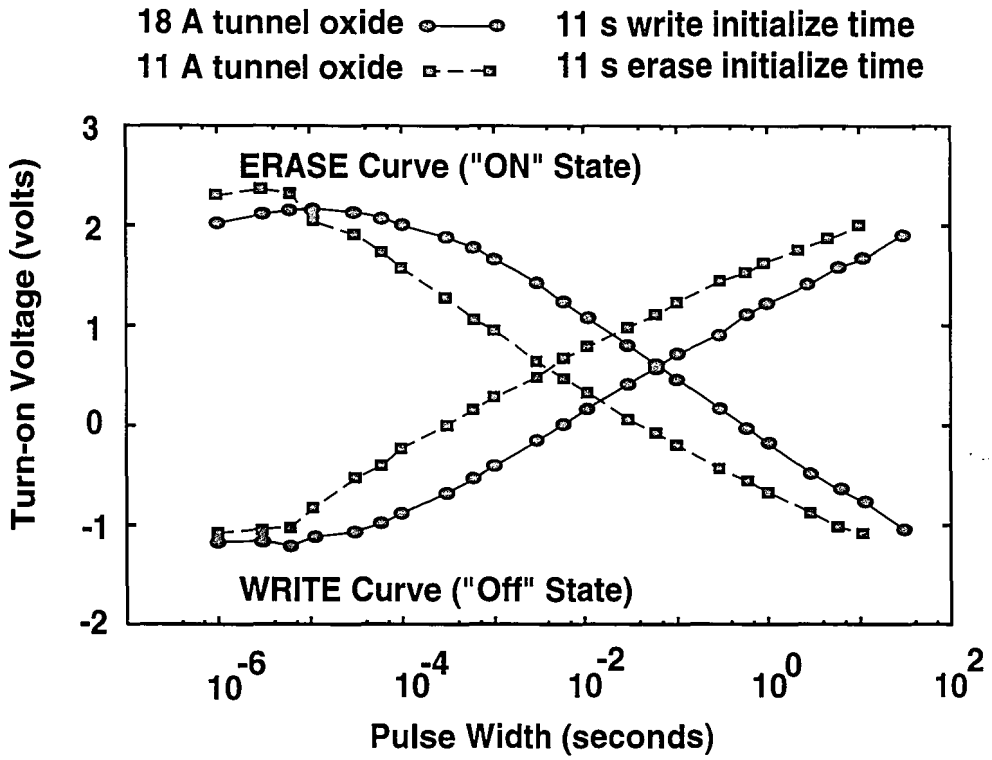


Figure 2.3: The n-channel SONOS Erase/Write curves for two different tunnel oxide thicknesses are presented. Note that thinner devices are able to program faster because for a given programming voltage, they will have a higher electric field across the tunnel oxide. Both devices had a 49$\mathring{A}$ nitride and a 40$\mathring{A}$ blocking oxide and were tested with $\pm$5V. [4]

(Write) state, and then Write (Erase) pulses of increasing magnitude are applied to the gate in order to shift the threshold voltage to the Write (Erase) state. After each Write (Erase) pulse is administered, the device is again reprogrammed to its inital state. Once the device is preset to one state, e.g. Erase, the quantity of charge in the nitride proportionately diminishes the effective electric field generated by the Write pulse across the tunnel oxide. Similarly, when the device is reset to the original state, the resultant threshold shift will depend upon the amount it had previously been written. Thus, each device operation displays a dependence upon its programming history. This effect can be incorporated into the mathematical expression for the threshold voltage shift [4].

The following Erase/Write analysis attempts to model the programming behavior depicted in typical Erase/Write curves, e.g. Fig.2.3. For a first order analysis, we make several approximations:

1. There is negligble current through the blocking oxide.
2. There is no contribution from the Si-SiO_2 interface traps to the charge injection.
3. The nitride traps are treated as a sheet charge located at a distance $\overline{X}$ (also referred to as the charge centroid) from the tunnel oxide-nitride interface.
4. The relative dielectric constants of the blocking oxide and the tunnel oxide are the same.

By Gauss's Law, the programming voltage can be expressed as:

$$V_P = (x_{OT} + x_{OB} + \frac{\epsilon_{OX}}{\epsilon_N} x_N) E_{OT} + \phi_{GS} + \phi_S + (\frac{x_{OB}}{\epsilon_{OX}} + \frac{x_N - \overline{x}}{\epsilon_N}) Q_N \qquad (2.6)$$

where the initial electric field is

$$E_{OT}(0) = \frac{V_P - (\frac{x_{OB}}{\epsilon_{OX}} + \frac{x_N - \overline{x}}{\epsilon_N}) Q_N(0) - \phi_{GS} - \phi_S}{x_{eff}} \qquad (2.7)$$

We assume no time dependence on the programming voltage (i.e. it is constant) so differentiating Eqn.(2.6) with respect to time and rearranging, we obtain

$$(\frac{x_{OB}}{\epsilon_{OX}} + \frac{x_N - \overline{x}}{\epsilon_N})\frac{dQ_N}{dt} = -(x_{OT} + x_{OB} + \frac{\epsilon_{OX}}{\epsilon_N}x_N)\frac{dE_{OT}}{dt} \tag{2.8}$$

Note, $\frac{dQ_N}{dt}$ represents the net injection current into the nitride. Normally, this would include the injection and leakage tunneling currents from the gate and bulk. However, we assume the dominant component to be Modified-Fowler Nordheim tunneling from the semiconductor bands to the nitride bands, which can be represented by J_{OT}. Eqn.(2.8) is rewritten to emphasize the time dependence of the electric field.

$$\frac{dE_{OT}}{dt} = -AJ_{OT}[|E_{OT}(t)|] \tag{2.9}$$

where

$$A = \frac{\frac{x_{OB}}{\epsilon_{OX}} + \frac{x_N - \overline{x}}{\epsilon_N}}{x_{OT} + x_{OB} + \frac{\epsilon_{OX}}{\epsilon_N}x_N} \tag{2.10}$$

Expressing J_{OT}, the tunnel oxide injection current, in terms of an abbreviated binomial expansion of the Modified Fowler-Nordheim tunneling mechanism, yields,

$$J_{MFN} = BE_{OT}^2 e^{(-E_T/|E_{OT}|)} \tag{2.11}$$

where

$$B = \frac{q^2}{16\pi^2 h\phi_1} \tag{2.12}$$

and

$$E_T = \frac{4(\phi_1 - \phi_2)\sqrt{2m_{OX}^* q\phi_1}}{3\hbar} \tag{2.13}$$

The potentials ϕ_1 and ϕ_2 denote the oxide and nitride quantum mechanical barrier heights, respectively and m_{OX}^* represents the effective mass in the oxide. Eqn.(2.9) is solved by separation of variables and the time-dependent electric field becomes

$$E_{OT}(t) = \frac{E_{OT}(0)}{1 + \frac{|E_{OT}(0)|}{E_T} ln(1 + \frac{t}{\tau})} \tag{2.14}$$

where

$$\tau^{-1} = ABE_T e^{-E_T/|E_{OT}(0)|} \tag{2.15}$$

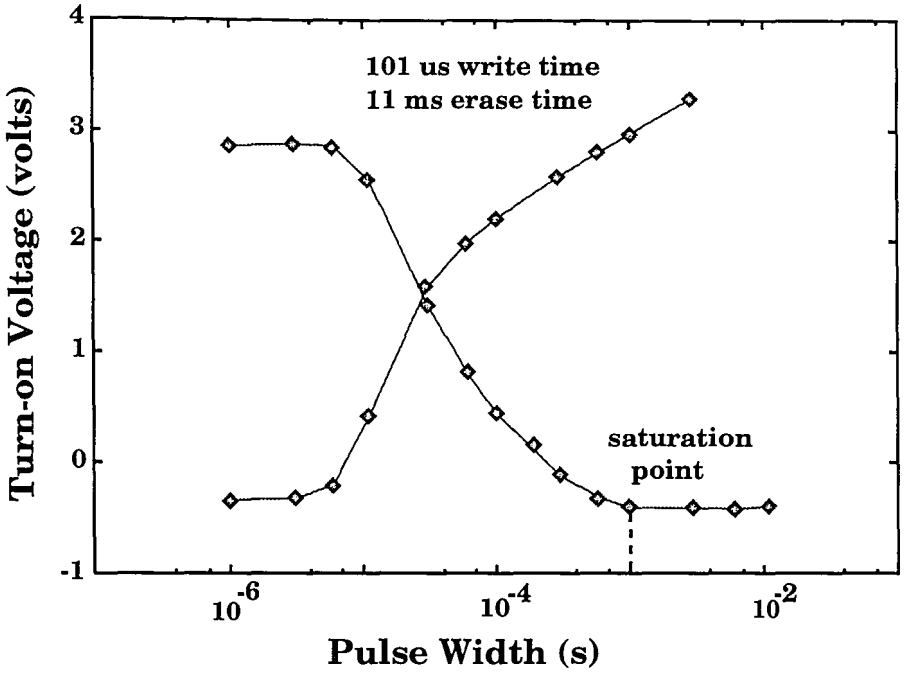


Figure 2.4: An Erase/Write curve on an n-channel SONOS device for 9V programming. In this measurement the device was actually programmed to saturation with a 1ms pulse. Contrast this with the 5V programming on a similar device in Fig.2.3 where 11s pulses were not sufficient to saturate the device because the smaller electric field across the tunnel oxide resulted in slower programming.[5]

Combining Eqns.(2.2),(2.6) and (2.14), we find the expression for the threshold voltage shift. expression as a function of the initial tunnel oxide electric field prior to programming.

$$\Delta V_{TH} = \pm \frac{x_{eff} \frac{E_{OT}^2(0)}{E_T} ln(1 + t/\tau)}{1 + \frac{|E_{OT}(0)|}{E_T} ln(1 + t/\tau)} \tag{2.16}$$

As Eqn.(2.16) shows, the threshold voltage shift is dependent upon the initial tunnel oxide electric field prior to programming. $E_{OT}(0)$, in turn, is dependent upon the initial charge is the nitride, $Q_N(0)$, as related by Eqn.(2.7). Thus, each programming operation on a SONOS device is dependent upon its programming history.

If the initial/reset state is always certain, then the threshold voltage shift caused

by different programming pulses can be accurately determined. However, our memory testing station [24, 6] permits the device to be read only after it is programmed but not after it is reset. Hence, we depend upon the assumption that the reset threshold voltage does not deviate significantly each time the device is reinitialized. However, since the Erase/Write measurement maintains a constant reset pulse while varying the widths of the programming pulses, the device may not necessarily be reinitialized to the same threshold voltage as the programming pulse widths become larger. Often, this problem is avoided by initially saturating the nitride to one state and reprogramming to that saturated state after each programming cycle. With the reset state fixed, the effect of each programming pulse can be independently determined.

2.2.2 Retention

Retention measurements indicate the efficiency of data storage over time. They are performed by forcing the device into the **Write (Erase)** state and then reading the data after a delay to see if significant charge has leaked out of the nitride. The data is refreshed after each Read phase and then the read delay time is increased logarithmically in order to project the threshold voltage memory window over a ten year period. Fig.2.5 shows the retention characteristics of an n-channel SONOS device fabricated in our processing lab. Note, the decay rate for holes (which provides the Erase state for an n-channel device) is greater than that for electrons. If this memory window were to be implemented in an actual SONOS memory array with the read voltage set at ground, the retention would be limited to approximately one second when the Erase curve would intersect the zero volt axis. After this point, the Erase and Write states would become indistinguishable and the data integrity would no longer be preserved. Ideally, in this example, the retention is maximized

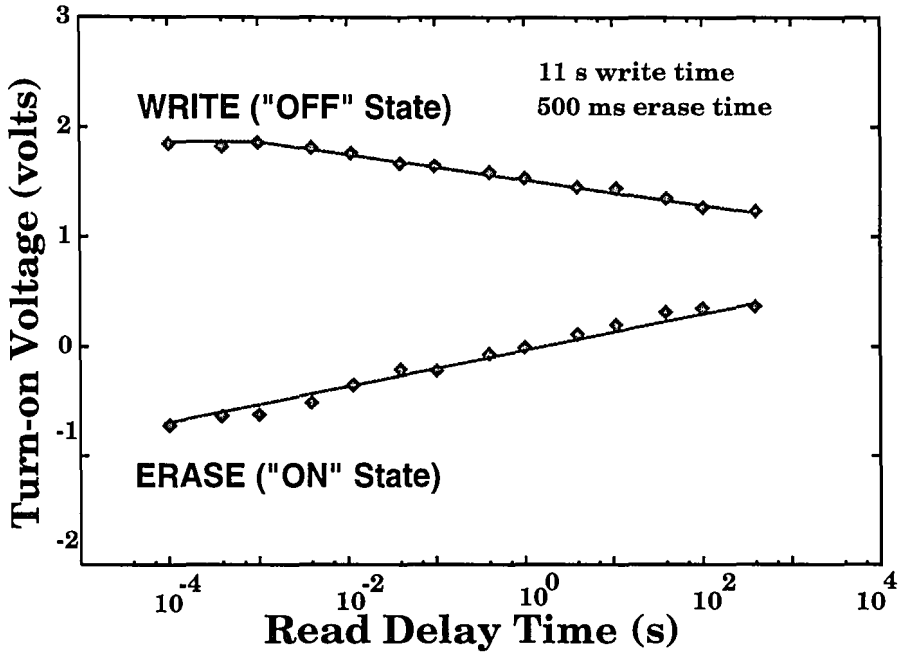


Figure 2.5: The retention characteristics of a SONOS device ($x_{ot} = 18\mathring{A}, x_N = 49\mathring{A}, x_{ob} = 40\mathring{A}$) are presented for 5V programming.[5]

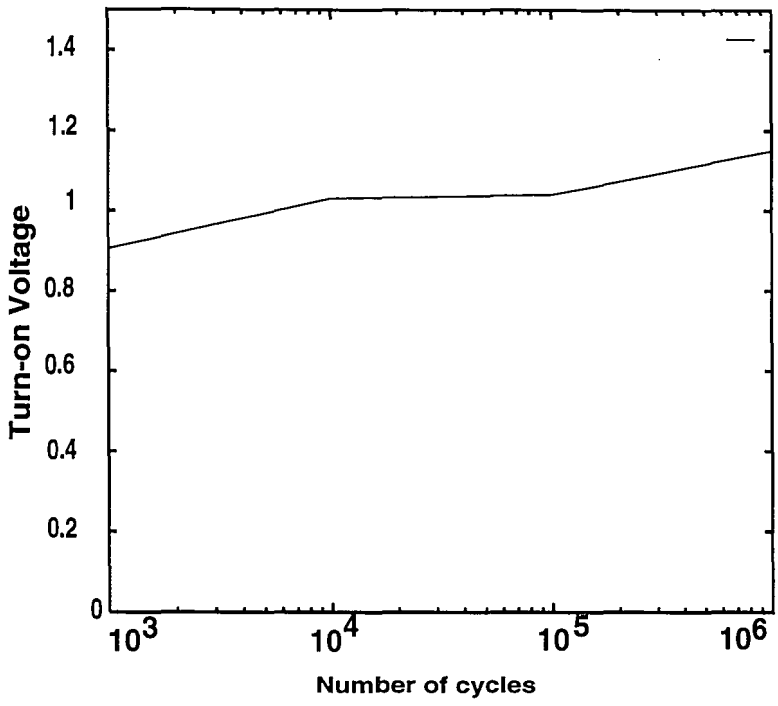


Figure 2.6: The threshold voltage of an n-channel SONOS device (x_{ot}=20$\mathring{A}$, x_N= 73$\mathring{A}$, x_{ob}=54$\mathring{A}$is presented versus number of Erase/Write cycles. The excessive cycling increases charge trapping at the Si-SiO_2 interface and prevents complete discharging, thus, increasing the threshold voltage. [6]

with the memory window designed such that both the Write and Erase decay curves intersect the 0V axis at the same point.

2.2.3 Endurance

Endurance measurements demonstrate the reliability of the memory over repeated **Write/Read/Erase** cycles. As the device is repetitively programmed and erased, the electric field across the dielectric is constantly reversed. This resultant stress builds up interface states which impede charge injection into the nitride and

enhance charge leakage from the nitride, thus reducing the memory window. In addition, the oxide, itself, can build up charge, and upon further stress, it may eventually break down. We have optimized the endurance of our SONOS devices by growing high quality (near defect-free) oxides to minimize the buildup of interface states [25].

For most EEPROM applications, endurance is not much of an issue since the data is not frequently altered. However, in order to develop semiconductor disks, it is imperative that the base nonvolatile memory technology be able produce devices which can withstand at least 10^6 programming cycles and 10^{12} read cycles without compromising the validity of the data.

2.3 Memory Window Design Considerations

In order to perform at reasonable speeds, the storage elements in a memory array cannot operate with a saturated memory window. Hence, a dynamic memory window must be chosen to define the necessary Erase and Write pulse widths for a given programming voltage. The positioning of the memory window is contingent upon balancing the decay rates of the Erase and Write states while simultaneously maximizing the read current. Since the decay rate for holes is approximately twice as great as that for electrons, the memory window should be appropriately placed to permit their respective threshold voltages to intersect at the zero crossing. Fig.2.7 demonstrates the relationship between the memory window placement and the data storage time. The decay rates are logarithmic and differ because of the different energy barrier heights of electrons and holes to back-tunnel from the nitride. The Erase state threshold voltage for a p-channel SONOS device after a read delay time, t_{RD} can be expressed as:

$$V_{TH(E)}(t_{RD}) = V_{TH(E)}(t_0) - r_E log(t_{RD}/t_0) \tag{2.17}$$

while the Write state threshold voltage is:

$$V_{TH(W)}(t_{RD}) = V_{TH(W)}(t_0) + r_W log(t_{RD}/t_0) \tag{2.18}$$

which defines the time dependent memory window to be:

$$V_{MW}(t_{RD}) = V_{TH(E)}(t_{RD}) - V_{TH(W)}(t_{RD}). \tag{2.19}$$

Rewriting this expression in terms of the decay rates, we find:

$$V_{MW}(t_{RD}) = V_{MW}(t_0) - (r_E + r_W) log(t_{RD}/t_0). \tag{2.20}$$

The memory window collapses to zero at a time t_S.

$$V_{MW}(t_S) = V_{MW}(t_0) - (r_E + r_W) log(t_S/t_(0)) = 0 \tag{2.21}$$

Thus, the maximum data storage time, t_S, becomes:

$$t_S = t_0 10^{(V_{MW}(t_0))/(r_E + r_W)} \tag{2.22}$$

Additional care must be taken to insure that the read detection scheme is sensitive enough to distinguish between the two logic states even as the memory window diminishes, else the data storage time will be severely limited.

The initial memory window is set by both the magnitude of the Erase and Write pulses as well as the order in which they are applied to the device. The memory is initially erased before data is written so the memory window will have a tendency to shift positively. In addition, the Erase pulse is generally longer than the Write pulse since the memory is sector erased but only single bit written. Hence, the shorter Write pulses will not be able to completely compensate for the threshold voltage shift caused by the Erase pulse. However, if during the device fabrication, the gate is doped n^+, the intrinsic threshold voltage will be shifted negatively by as much as 1V, allowing the memory window to be more appropriately centered to maximize data retention.

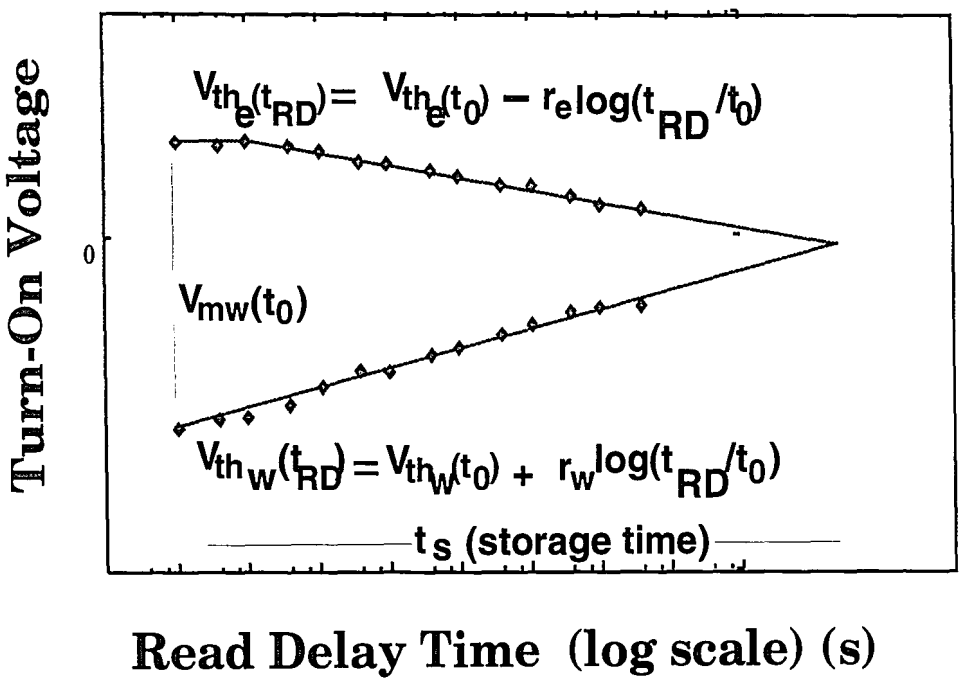


Figure 2.7: A hypothetical optimum dynamic memory window under 5V programming.

Chapter 3

The Memory Array Design

3.1 Approach

Once a dynamic memory window is chosen and true 5V programmability is demonstrated on discrete SONOS devices, the next goal is to implement a memory cell, and then, an entire SONOS memory test array. Fig.3.1 depicts the functional block layout of a 2 x 2 SONOS memory array using the device scaling guidelines developed by French [4] and Hu [21]. The array is unique since it contains p-channel SONOS transistors as the memory elements and is powered by a single unipolar 5V supply. A small array size is chosen because it can perform the same addressing and programming operations as larger EEPROMs while providing a higher yield in the fabrication.

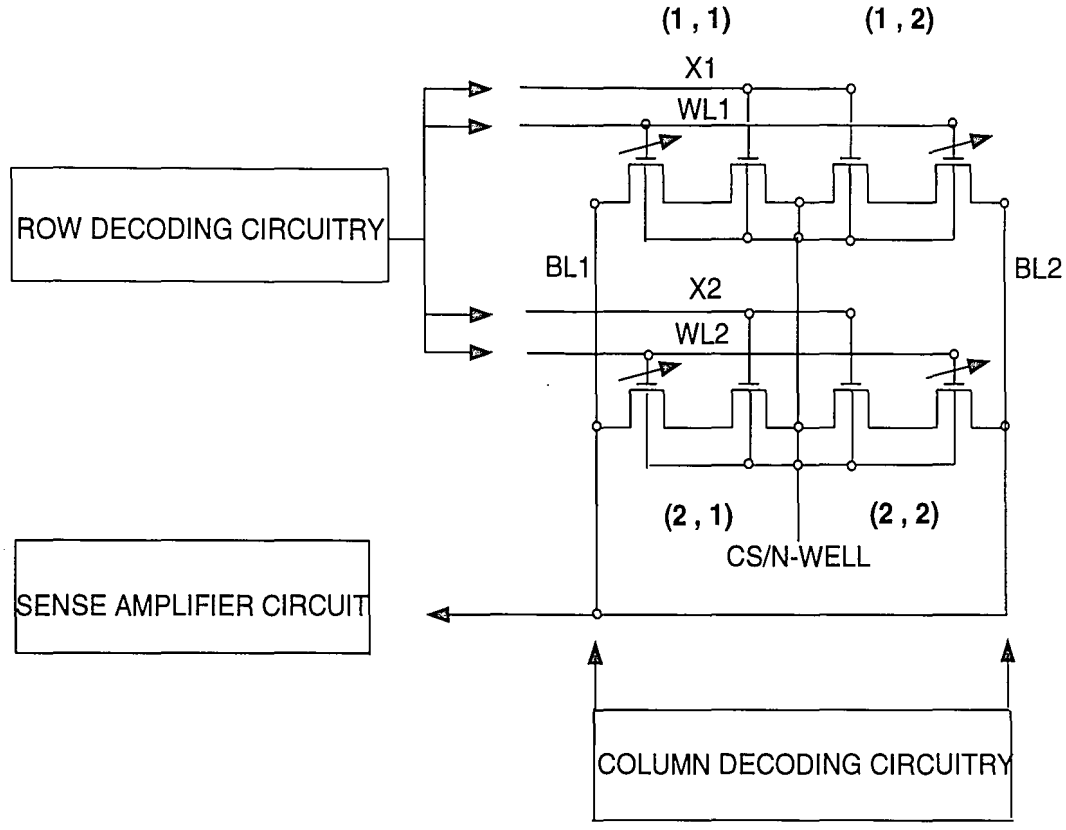
The architecture of the 2 x 2 array is similar to that used by Citizen's Watch in their 1Mb SONOS EEPROM [23]. However, instead of using an n-channel split-gate memory cell, we employ a cell design consisting of a p-channel select/address transistor in series with a p-channel SONOS memory transistor. The memory and select devices share a common n-well that is isolated from the addressing and sensing circuitry. This isolation is vital because the gate and bulk potentials of the memory devices are repetitively switched from one polarity to the other as the device undergoes different programming operations. While the alternating polarity of the gate to bulk voltage may be conducive to memory programming, it is also detrimental to the performance of the MOSFETs in the addressing circuitry.

3.2 Basic Operation

The biasing conditions described in chapter 2 for the SONOS device provide the basis for operating the cells of the memory array. The devices are operated upon by the assorted data lines that interface the actual memory array itself with the addressing circuitry. These data lines include the **Word Lines** (**WL1** and **WL2**), the **Bit Lines** (**BL1** and **BL2**), the **Select Lines** (**X1** and **X2**) and the common bulk input (**CS/n-well**). The **Word Lines** and **Bit Lines** apply the memory gate and source voltages, respectively, while **CS/n-well** provides the bulk bias. **X1** and **X2** control the series address transistors which only operate during the read mode where they provide a current path through the SONOS transistors and into the read detection or sensing stage which determine the logic state of the device.

By referring to Fig.3.1, we see that each memory element shares a common **Word Line** or **Bit Line** with its neighbors. Hence, when one memory cell is exercised, the others must be inhibited lest their data be inadvertently altered. The key feature of the addressing circuitry is that it must be able to both select any arbitrary cell as well as inhibit the other memory cells. This can be illustrated by examining the operating modes of cell **(1,1)**.

To write SONOS cell **(1,1)**, the addressing circuitry causes 0V to be applied to **WL1** while the n-well (**CS**) is pulsed to 5V, enabling both the source (**BL1**) and bulk to be at equal potentials. This allows programming of the device by altering the threshold voltage of the transistor negatively, forcing it into the low conduction state. Because none of the address/select transistors are turned on, the common drain terminals are left floating but they will soon settle to a voltage close to the source and bulk, and will thus, not significantly impede the operation of the device. While cell (1,1) is written, the other cells are inhibited by the decoding circuitry.



Operating Modes for Memory Cell Element (1 , 1)

Mode	X1	WL1	X2	WL2	BL1	BL2	CS
READ	GND	Vr	Vp	Vr	SA	Open	Vp
WRITE	Vp	GND	Vp	Vp	Vp	GND	Vp
ERASE	Vp	Vp	Vp	GND	GND	GND	GND

Figure 3.1: The SONOS memory array and its operating conditions

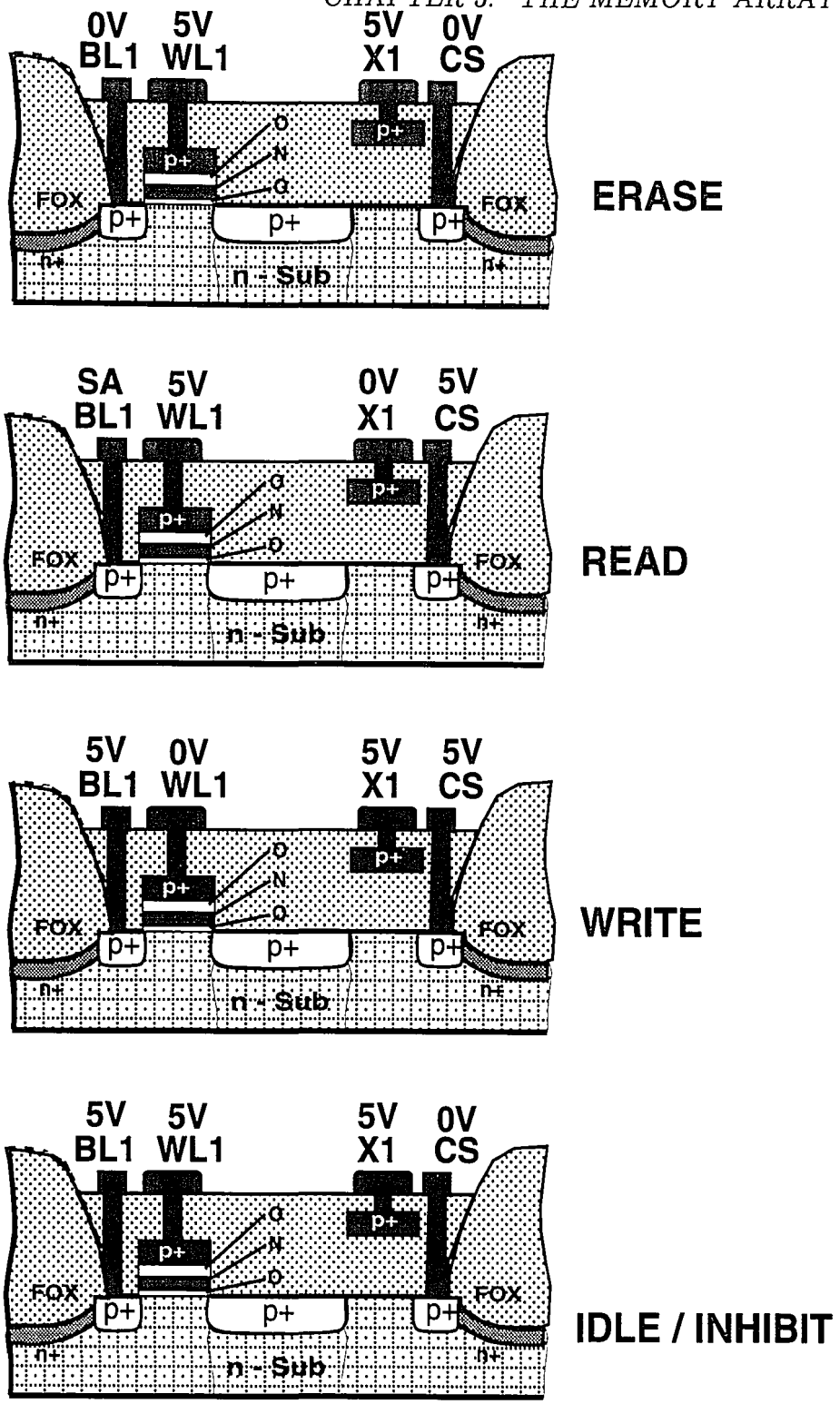


Figure 3.2: The cross section of a memory cell during its operational modes

Cell **(1,2)** is not affected by the voltage applied to **WL1** since **BL2** is set to the same potential, hence, there is no effective gate to source voltage. Cell **(2,1)** is also inhibited by a zero gate to source bias because **WL2** is set the same as **BL1**. Cell **(2,2)** is inhibited due to a zero gate to bulk differential voltage, preventing the formation of an inversion layer to allow programming.

Unlike the Write operation, which affects one cell at a time, the Erase operation is performed on an entire row. To erase row 1 (containing cells **(1,1)** and **(2,2)**), 5V is applied to **WL1** and 0V is applied to **CS**, **BL1** and **BL2**. Row 2 (containing cells **(2,1)** and **(2,2)**) are inhibited by keeping **WL2** at the same potential as **CS** and the Bit Lines. Row 1 now has a negative gate to source bias, enabling electron tunneling into the nitride to set the devices into the high conduction state (depletion mode) while row 2 has a 0V gate to source bias which inhibits tunneling, and hence, writing or erasing.

The Read operation is performed serially with one cell read at a time. To read cell **(1,1)**, **X1** is set to 0V and **CS/n-well** is pulsed to 5V, saturating the address transistor, and providing a current path from the CS source to the memory transistor. If the SONOS device is in the high conduction state, then a read voltage ($V_r \approx V_p = 5V$) applied to the gate will generate current flow proportional to the square of the programmed threshold voltage. The read current flows through **BL1** and into a read detection sensing stage to be translated into a data output. Note, no voltage is actually applied to **BL1** during the Read mode; it merely serves as a path from the memory cell to the sense amplifier. Since it is imperative that the Read operation be non-destructive, the Read voltage must be close to the source and bulk potential in order to minimize tunneling, and hence, data alteration.

The other cells are inhibited by decoupling their connection to the sense amplifer. While cell **(1,1)** is read, **BL1** is switched into the sense amplifier and **BL2** is left floating. Thus, even though the same Read voltage is applied to cells **(1,1)** and **(1,2)**, only cell **(1,1)** provides a path for current to flow to the output. Meanwhile

cells **(2,1)** and **(2,2)** cannot source current because the address transistors on row 2 remain off.

If the device in cell **(1,1)** is in the low conduction state, it will not source any current to the sense amplifier. A default output is then provided by the sense amplifier to indicate this.

3.3 Input/Output Signals

The addressing and programming of a particular memory cell are considered simultaneous operations. The combinatorial logic in the addressing circuitry uses the address control signals (**A0** and **A1**) with the programming control signals (**CS**, **ERASE** and **READ**/$\overline{PGM}$) to switch in the supply voltages (5V and 0V) to the appropriate Bit Lines, Word Lines and Select Lines. **A0** locates the row of a particular memory cell by pulsing high in order to access row 1 and pulsing low to access row 2. The column location is provided by **A1** which pulses high for column 1 and low for column 2. The programming mode is determined by the combination of **READ**/$\overline{PGM}$ and **ERASE**. If **READ** is high, then the addressing circuitry switches in the appropriate voltages to the location specified by **A0** and **A1**. If **READ** is low and **ERASE** is high, then the appropriate erase mode voltages are applied. Note the Write operation occurs by default when both **READ** and **ERASE** are low.

3.4 The Addressing/Decoding Circuitry

Fig.3.3 shows the circuit schematic for the memory array. Extensive simulations on the circuit were performed using the SPICE 3c1 software package. Each circuit module was tested individually for functionality and timing both before and after the VLSI layout. The complete circuit, itself, was also rigorously tested before and after layout, with special emphasis on the different data sensing schemes. Only the final results of these simulations will be presented in this thesis.

3.4.1 Row Decoder

The row decoder consists of identical circuits for each row, differing only in that the top row is addressed by **A0** while the bottom is addressed by $\overline{A0}$. For this reason, only the addressing circuitry of the top row will be explained.

The logic for the select line, **X1**, is functionally simple. The address/select transistors of row 1 remain off during the Erase and Write modes but will turn on during the Read mode. This function can be served with a simple nand gate whose inputs are **A0** and **READ**. When both inputs are true, **X1** goes low and activates the p-channel address transistors. Otherwise, **X1** remains high and the address transistors are off.

The logic for the Word Line, **WL1**, consists of three signal paths corresponding to the three different programming voltages that are applied to the Word Line. The Read programming voltage, Vr, is connected to the Word Line via p-channel MOSFET **mr**. This transistor is activated only when **READ** is high, and hence,

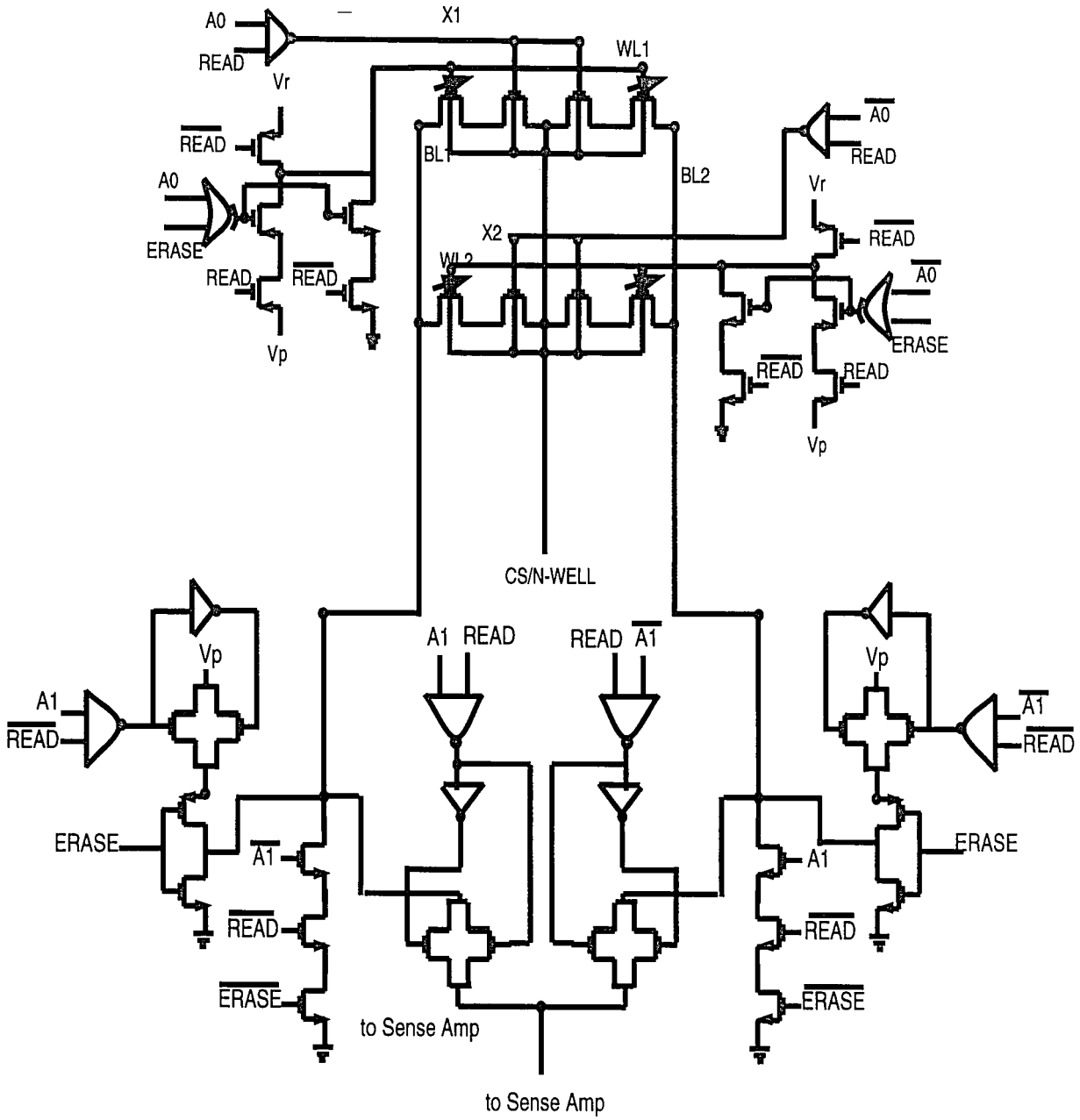


Figure 3.3: Circuit schematic for memory array

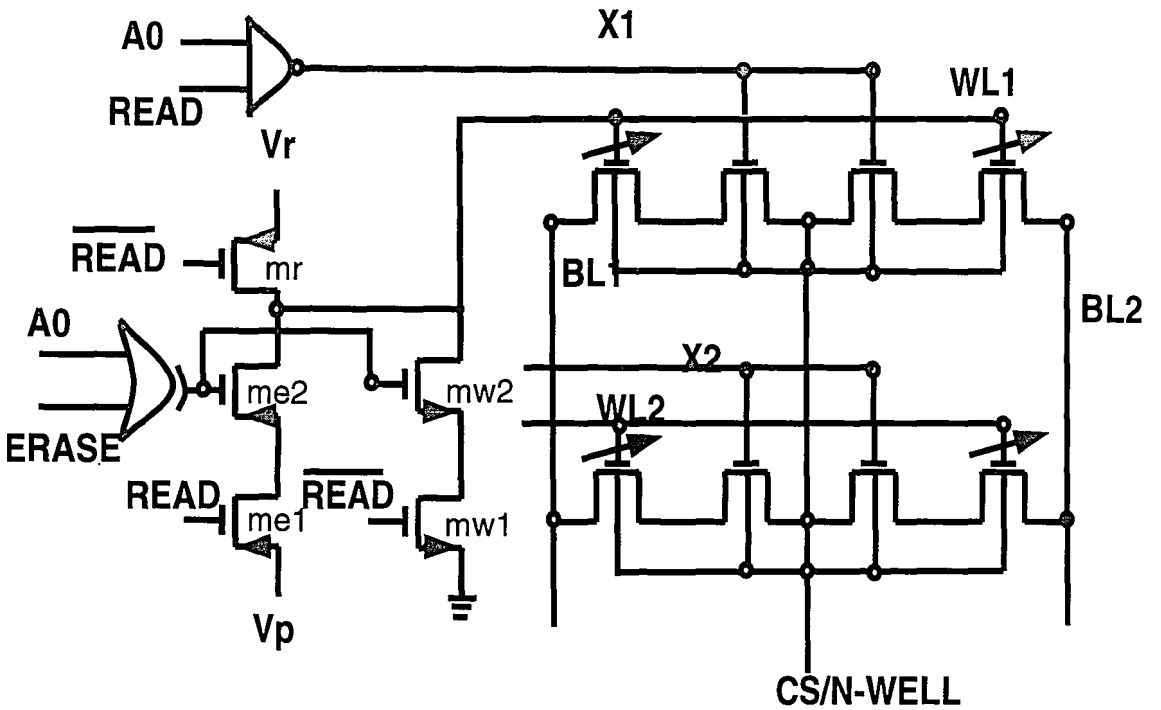


Figure 3.4: Row decoder circuit

$\overline{READ}$ (or **PGM**) is low. The Write voltage (0V) is connected to **WL1** through two n-channel MOSFETs, **mw1** and **mw2**. Both transistors will turn on simultaneously only when both **READ** and **ERASE** are low. The output of the exclusive-or gate (XOR) insures the latter condition holds true while **READ** insures the former. Vp, the voltage for the Erase mode, is connected to **WL1** via two p-channel MOSFETs, **me1** and **me2**. Transistor **me1** is activated by the XOR gate since both **A0** and **ERASE** are true while **me2** is activated since **READ** is low, hence, Vp is connected to **WL1**.

If **A0** is not true, then row 2 is selected and the addressing logic must insure row 1 is inhibited. The NAND gate always provides a 5V output whenever **A0** is low, preventing the select transistors in row 1 from activating. Meanwhile the XOR connects Vp to **WL1** during the Write operation and 0V to **WL1** during the Erase mode, providing a 0V gate to bulk bias which prevents programming.

3.4.2 Column Decoder

Like the row decoder, the column decoder also contains two modules which differ only in that **BL1** is addressed by **A1** and **BL2** is addressed by $\overline{A1}$. Again, like the row decoder, there are three different signal paths to the Bit Line. One path provides the Write and Erase programming voltages, the second path inhibits **BL1** when **BL2** is addressed and the third path connects to the sense amplifier when a memory cell in column 1 is being read. During the Read mode, if both **A1** and **READ** are true, then transmission switch **tr1** connects **BL1** to the sense amplifier; otherwise it remains uncoupled. During the Write mode, if **A1** is true but both **READ** and **ERASE** are not, Vp is connected to **BL1** by activating **mp1** as well as causing the nand gate to close transmission switch **tw1**. When **ERASE** is high, 0V is connected to BL1 via n-channel MOSFET **mg1**. Note that since an entire

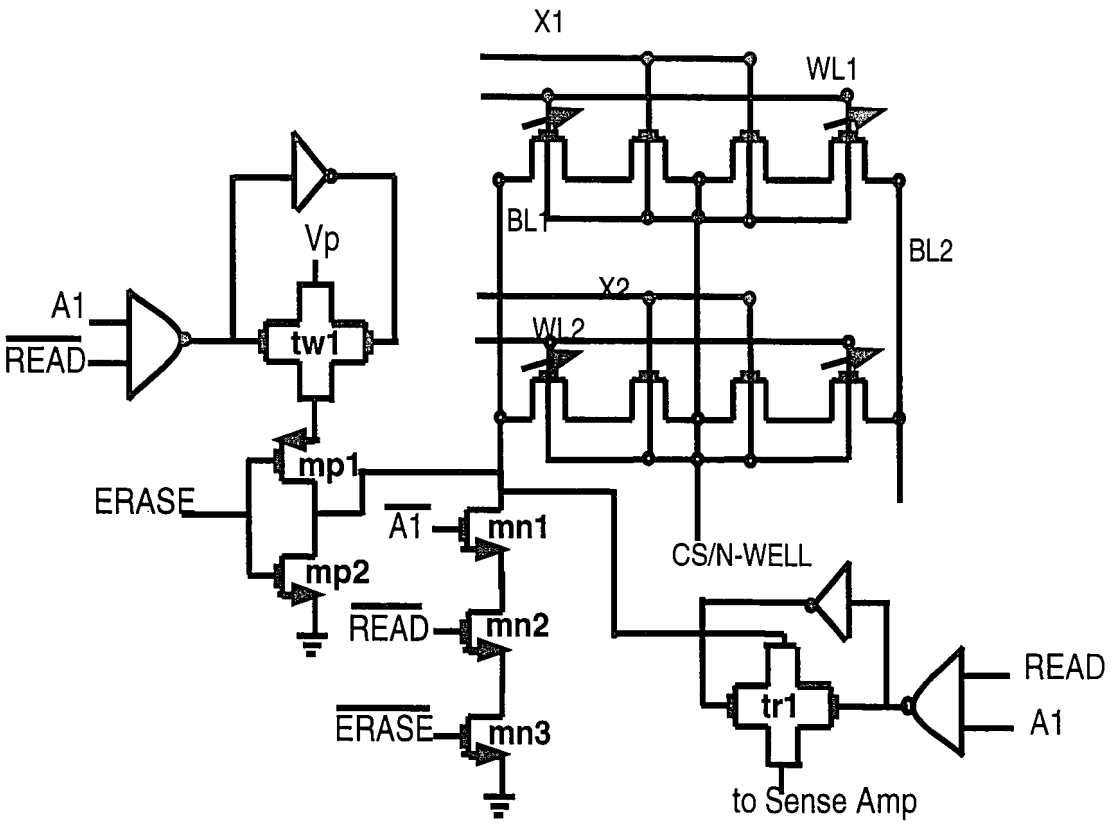


Figure 3.5: Column Decoder circuit

row is erased, both Bit Lines must have the same voltage applied to them. Thus, the column address **A1** need not be factored into the logic. During the Write mode, if **A1** is true, the nand gate outputs 0V and closes **tw1**. If **ERASE** is false during this period, **mp1** is activated and Vp is connected to **BL1**. The pass transistors **mn1**, **mn2** and **mn3** insure that **BL1** is inhibited when **BL2** is written (and vice versa).

3.4.3 Sense Amplifier

Sense amplifiers are used to convert the read current into an output voltage. They are especially useful in extending the longevity of the data. As the threshold voltage memory window for the SONOS devices decrease, so do their read currents. Sense amplifiers are able to tolerate significant degradation in the read currents while maintaining the integrity of the data.

The VLSI mask layout contains several different versions of the memory array, each containing a different sense amplifier. For one memory array, an off-chip current-to-voltage converter/transimpedance amplifier (fig. 3.6) is used to sense the current and convert it to an output voltage. A variable resistor is used in feedback with the opamp to provide higher gain if the read current degrades. If a virtual ground is assumed between the input terminals of the opamp, then the output voltage expression can be written as:

$$V_{out} = -I_{read}R_f \tag{3.1}$$

Typically, EEPROMs employ sense amplifiers that provide digital outputs, hence analog circuits such as transimpedance amplifiers are not commonly used.

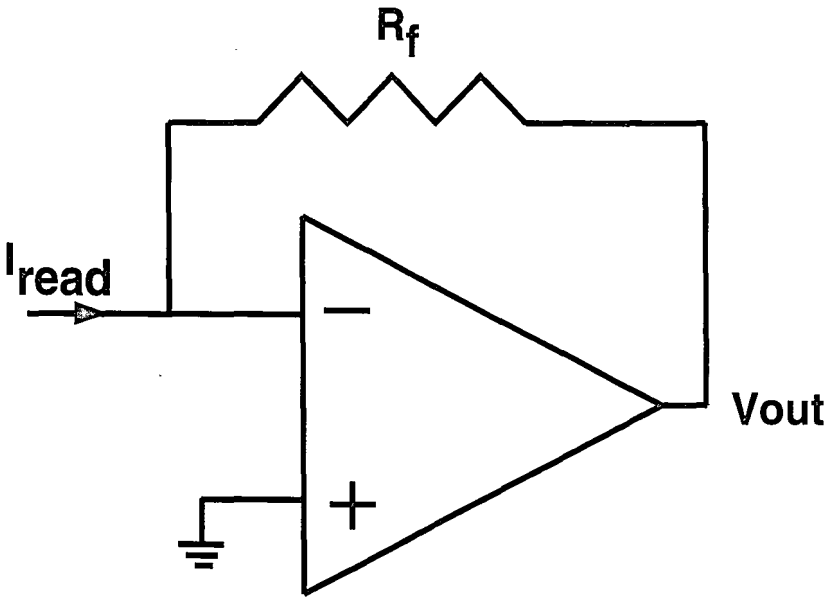


Figure 3.6: The transimpedance amplifier sensing scheme for analog output voltage.

The second sensing scheme employs a cross-coupled inverter configuration (Fig.3.7) to produce a digital output. Node A is precharged to 5V via transistor **m_p** and the input is precharged to ground via **m_g**. When **READ** is high, these transistors turn off and the read current is allowed to charge up the input capacitor (C_1) of the sense amplifier. If the memory device is in the high conduction state, then the capacitor will charge completely and cause **m2** to conduct, forcing node A to toggle to 0V. However, if the device is in the low conduction state then no current will flow into the sense amplifier and node A remains precharged to 5V. Since we define the high conduction state as logic 1 and the low conduction state as logic 0, node A is inverted to form the output of the sense amplifier.

The time required to charge the input capacitor, C_1, is restricted by the transient leakage through transistor **m4**. Before the array assumes the Read state, Vin is precharged to 0V. If current is generated in the Read mode, C_1 starts charging and V_{in} increases until it toggles the flip-flop. During this transition, **m4** progressively

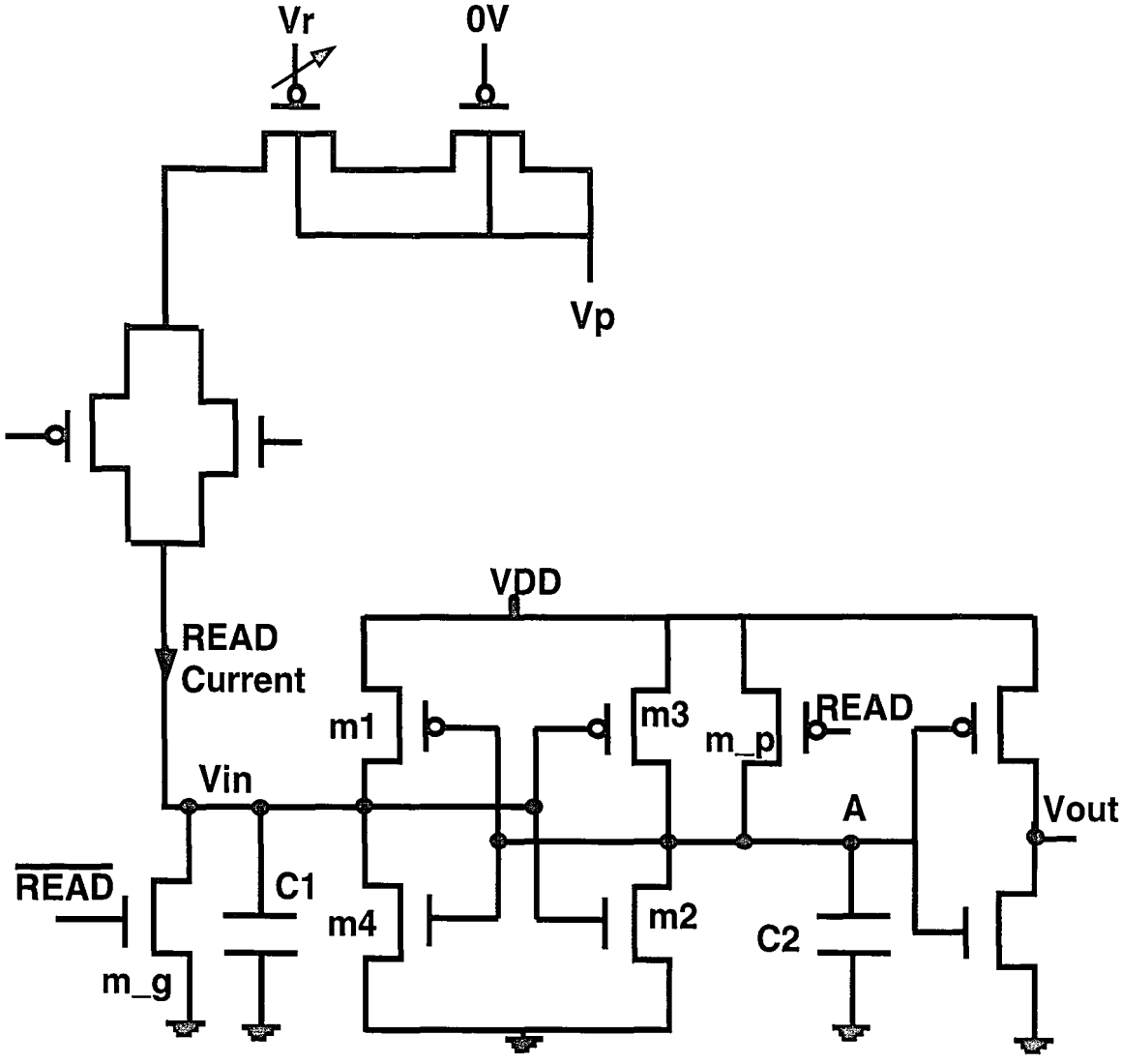


Figure 3.7: The cross-coupled inverter/flip-flop sense amplifier

decreases conduction until it is finally cut-off. Thus, until **m4** is completlely cut-off, it will continue to sink some of the charging current.

The equation depicting the transient currents at the input to the sense amplifier can be written as:

$$C_1 \frac{dV_{in}}{dt} + I_{D_{m4}} = I_{read} \tag{3.2}$$

which can be rewritten as,

$$\frac{dV_{in}}{dt} + \frac{I_{D_{m4}}}{C_1} = \frac{I_{read}}{C_1} \tag{3.3}$$

where $C_1 \frac{dV_{in}}{dt}$ is the current charging the input capacitor, $I_{D_{m4}}$ is the leakage current through **m4** and I_{read} is the current generated when a particular memory element is read. We assume the memory transistor to be in the high conduction state and is saturated when the read voltage is applied to it.

$$I_{read} = \frac{\beta_m}{2}(V_{GS_m} - V_{TM})^2 \tag{3.4}$$

Transistor **m4** is assumed to be in the linear region since it is trying to turn off during the charging transient.

$$I_D = \frac{\beta}{2}(2(V_{GS} - V_{TD})V_{DS} - V_{DS}^2) \tag{3.5}$$

The drain and source voltages of **m4** may be rewritten in terms of the potentials actually present.

$$V_{DS} = V_{in} \tag{3.6}$$

$$V_{GS} = V_G = V_{DD} \tag{3.7}$$

Thus, we have

$$I_D = \beta(V_{DD} - V_{TD})V_{in} - \frac{V_{in}^2}{2} \tag{3.8}$$

When V_{in} is small, the equation may be approximated by

$$I_D = \beta(V_{DD} - V_{TD})V_{in} \tag{3.9}$$

Reexamining Eqn.3.3, we can now make the following substitutions

$$\frac{dV_{in}}{dt} + \frac{\beta(V_{DD} - V_{TD})V_{in}}{C_1} = \frac{\beta_m}{2C_1}(V_{GS_M} - V_{TM})^2 \tag{3.10}$$

Laplace transforming, we obtain

$$sV_{in}(s) - V_{in}(0^+) + \frac{\beta(V_{DD} - V_{TD})}{C}V_{in}(s) = \frac{\beta_m(V_{GS_M} - V_{TM})^2}{2Cs} \tag{3.11}$$

$$V_{in}(s)(s + \frac{\beta(V_{DD} - V_{TD})}{C_1}) = \frac{\beta_m}{2C_1}\frac{(V_{GS_M} - V_{TM})^2}{s} \tag{3.12}$$

where $V_{in}(0^+) = 0$.

$$V_{in}(s) = \frac{\beta_m(V_{GS_M} - V_{TM})^2}{2C_1 s(s + \beta(V_{DD} - V_{TD})/C_1)} \tag{3.13}$$

By using partial fractions, we can simplify the expression.

$$V_{in}(s) = \frac{\beta_m(V_{GS_M} - V_{TM})^2}{2\beta(V_{DD} - V_{TD})}(\frac{1}{s} - \frac{1}{s + \beta(V_{DD} - V_{TD})/C_1}) \tag{3.14}$$

The expression can now be transformed back into the time domain to yield:

$$V_{in}(t) = \frac{\beta_m(V_{GS_M} - V_{TM})^2}{2\beta(V_{DD} - V_{TD})}(1 - e^{-\beta(V_{DD} - V_{TD})t/C_1}) \tag{3.15}$$

Since

$$g_{DS} = \beta(V_{DD} - V_{TD}) \tag{3.16}$$

and

$$V_{GS_M} = V_r - V_p \approx 0 \tag{3.17}$$

we can rewrite $V_{in}(t)$ as

$$V_{in}(t) = \frac{\beta_m}{2g_{DS}}V_{TM}^2(1 - e^{-(g_{DS}/C_1 t)}) \tag{3.18}$$

where $\frac{C_1}{g_{DS}}$ is a time constant τ For quicker charging times, the $\frac{g_{DS}}{C_1}$ ratio should be large. This constraint can be met by reducing conductance of **m4** with a smaller $\frac{W}{L}$ ratio.

The sensitivity of this read detection scheme can be calculated by noting that the input voltage must be at least half the supply voltage to trigger the flip-flop. Thus, C_1 must contain at least a minimum charge equal to $C_1\frac{V_{DD}}{2}$. Expressing this value in terms of the read current, we obtain the relation:

$$I_{READ}t_{READ} = C_1\frac{V_{DD}}{2} \tag{3.19}$$

Assuming negligible voltage drop across the address transistor, the read current can approximated by

$$I_{READ} = \frac{\beta_m}{2}V_{TM}^2 \tag{3.20}$$

Assuming:

- effective hole mobility = 250
- $\frac{W}{L} = 3$
- effective dielectric thickness = $100\mathring{A}$

we calculate β_m = 1.3 x 10^{-4}. If $t_{READ} = 150\mu$s and C_1 = 7pF, then $V_{TM(min)}$ = 0.03V, thus,defining the lower limit for the logic 1 state.

3.5 Array Layout

The mask layout was performed using the Mentor Graphics CAD software on a Sun workstation. From the layout, a series of fabrication masks will be constructed which will be used in our microelectronic processing facility to fabricate the memory array. The layout design is based upon the requirements of our n-well processing technology, necessitating the use of p-channel SONOS devices to provide the memory well isolation. We also employ a single poly-single metal process since we do not have the means to deposit or grow low temperature inter-poly oxides. For this

reason, we chose to design the memory cell as a two transistor cell as opposed to the equivalent split-gate topology which would reduce the cell size. In order to satisfy the resolution constraints imposed by the contact lithography and the wet chemical etching processes, our design rules are somewhat lax. The minimum channel length for each transistor is 8μm and most transistors, including the memory elements, have a width to length ratio of 3:1. This dimension provides an acceptable compromise between chip area and gain.

Figs.3.8-3.11 show the layout of the entire memory array as well as the individual test modules. In order to facilitate automatic testing, each layout employs a standard bonding pad arrangement to enable the Rucker & Kolls automatic wafer prober to use the same wafer card to probe the different circuits. Note, in addition to the input signals and the output of the sense amplifier, other key nodes of the array, such as the Bit Lines and Word Lines, are also connected to bonding pads. In this way, provisions are made to study the internal behavior of the circuit as well as its terminal characteristics.

3.6 Circuit Simulations

Prior to the actual layout, the addressing and sensing circuitry were extensively tested using the SPICE 3c1 simulation software. The basic SPICE model used for the MOSFETs is the MOSIS level 3 model which has been slightly modified to suit some of our particular processing conditions. The SONOS transistors have been modeled as MOSFETs with their threshold voltages preset to a particular programmed state. SPICE interprets the threshold voltage as a constant parameter and will not allow the gate voltage dependence to be inserted its definition within the circuit file. Thus, while the simulations are able to predict the transient behavior of the addressing and sensing circuitry, they are unable to analyze the transient variations in threshold voltage of the SONOS devices themselves. For simulation

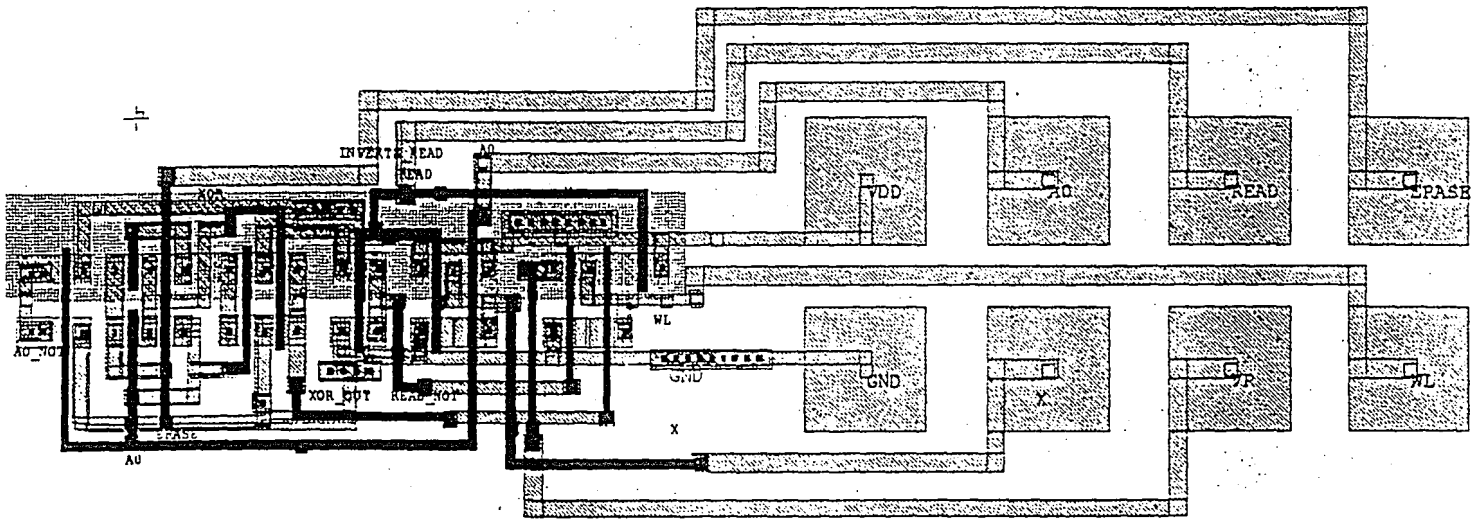


Figure 3.8: Row decoder test pattern

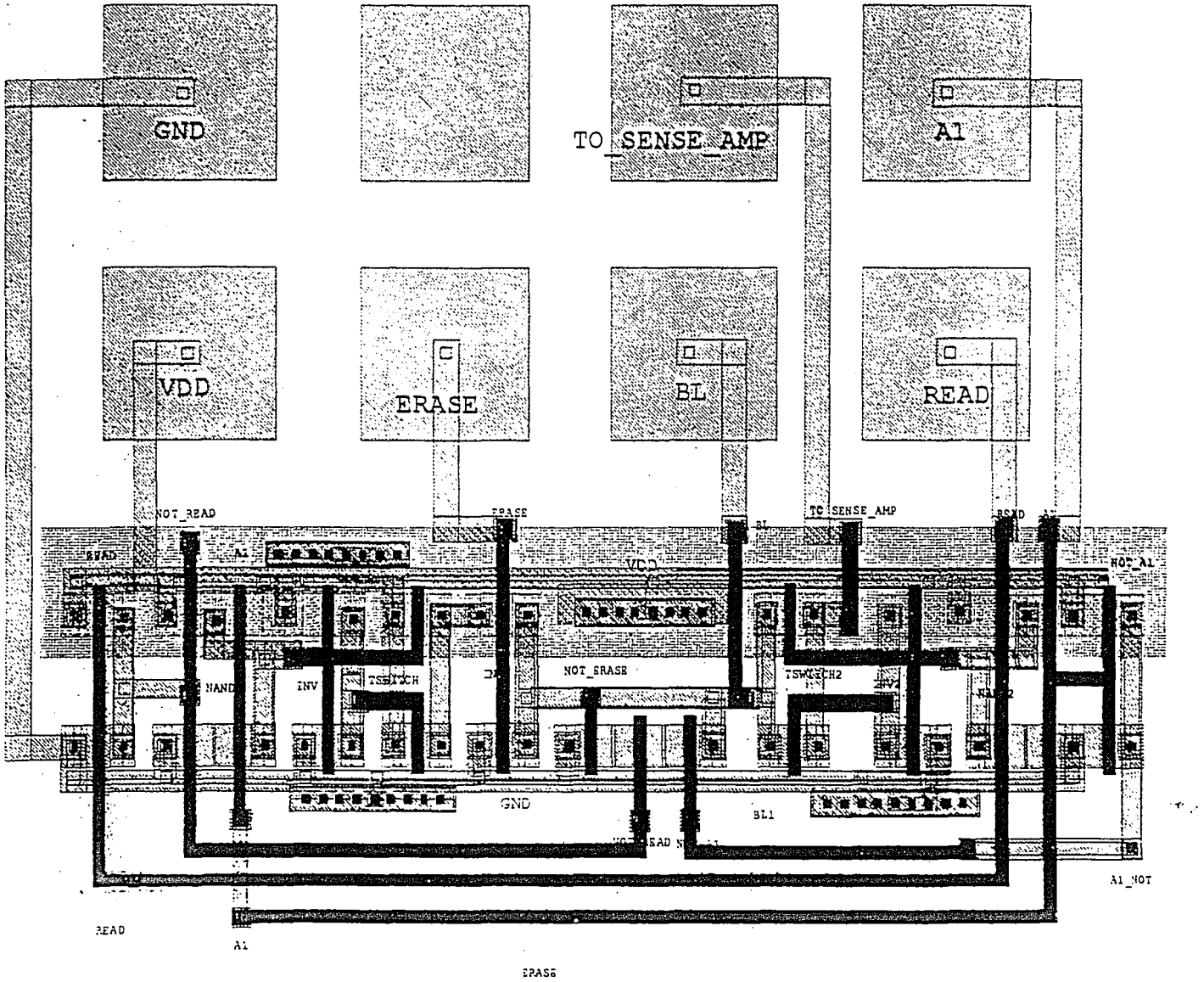


Figure 3.9: Column decoder test pattern.

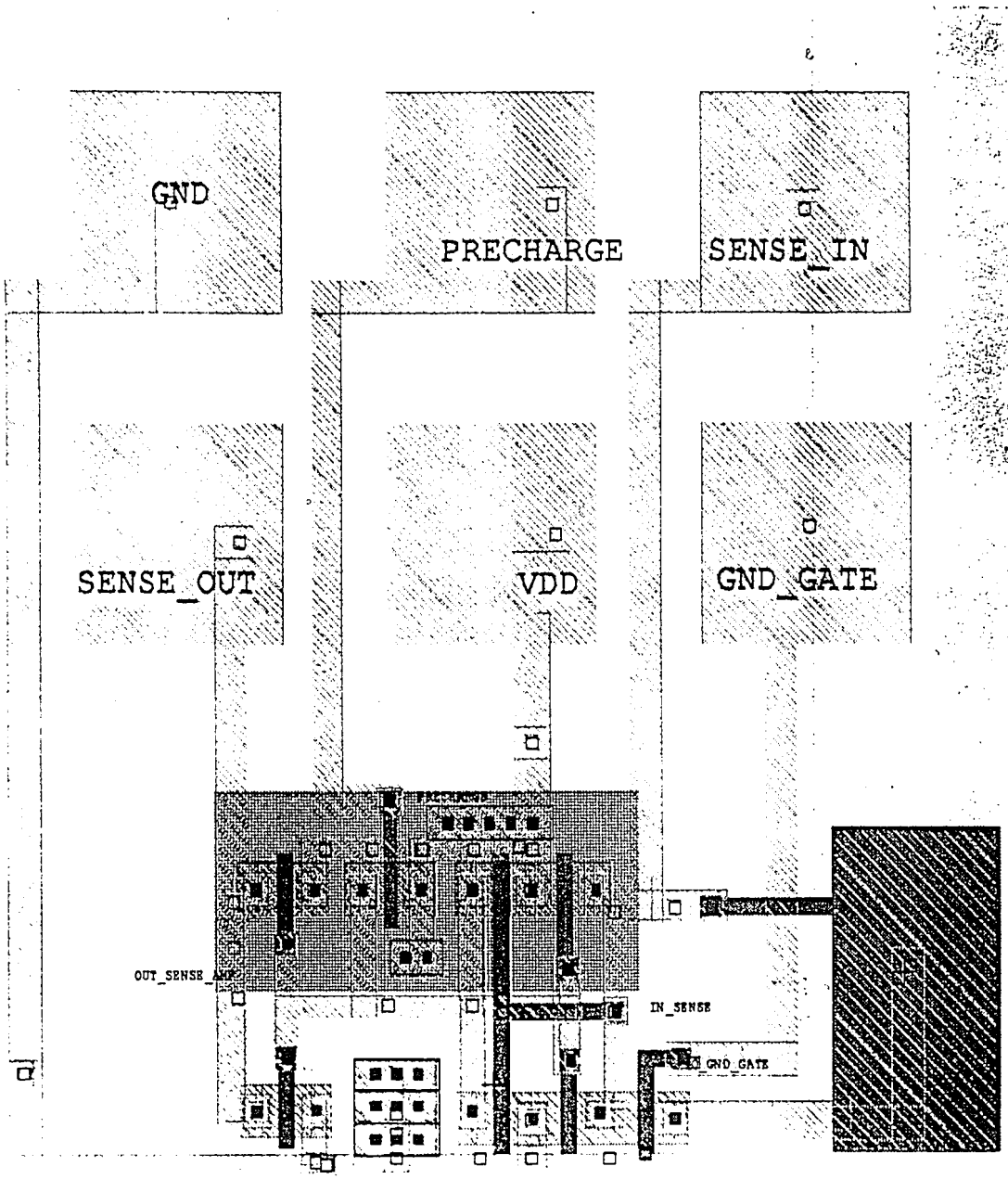


Figure 3.10: Sense amplifier test pattern

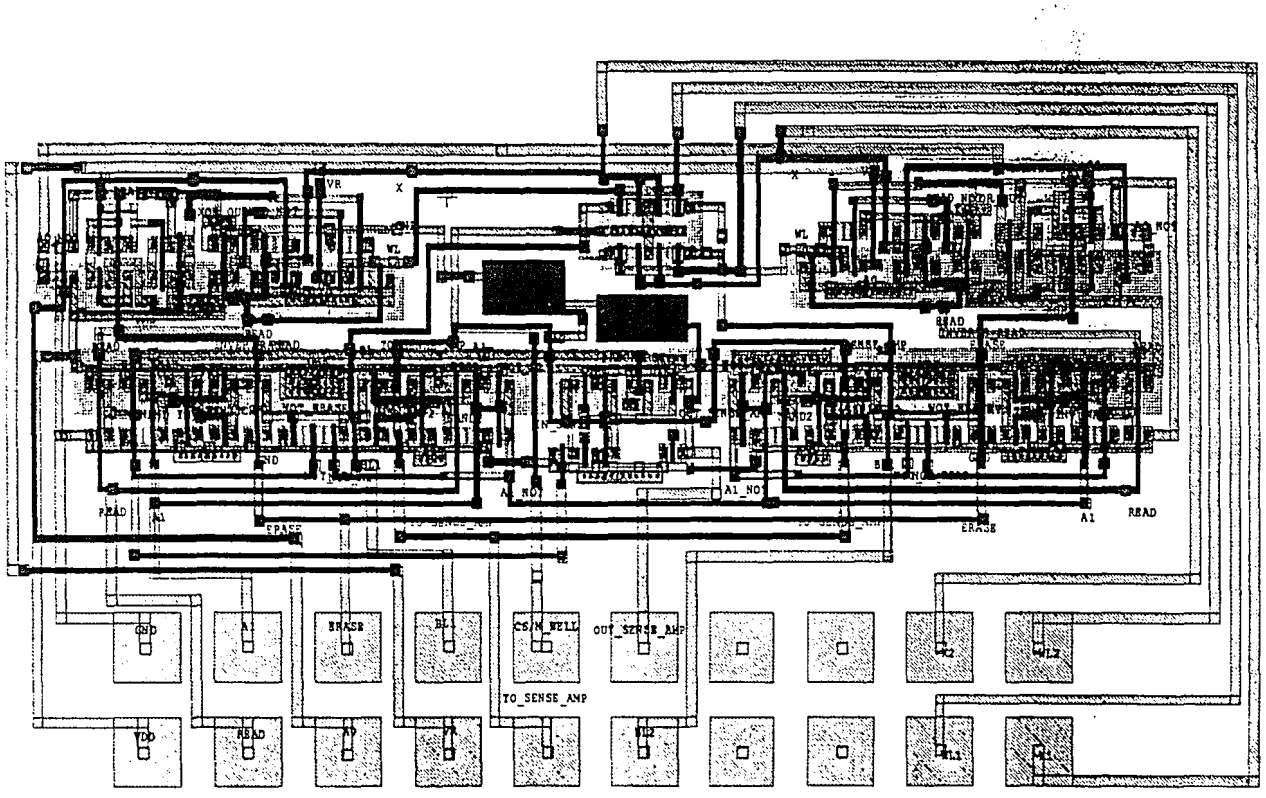


Figure 3.11: Memory array layout using the Mentor Graphics GDT software and Lehigh-SFC technology rules.

purposes, the programmed V_{TH} is set to the high conduction state in order to insure that current is always output to the sense amplifier when the device is read, whether or not it had been previously erased or written.

Circuit simulations were also performed on the actual layout itself. The Mentor Graphics GDT software package provides a means to extract a layout netlist that displays the equivalent circuit elements and the nodes to which they are connected. In addition to depicting the actual transistors, the netlist also contains the effective parasitic capacitance seen at each node including the interconnect and overlap capacitance. This netlist is compatible with the input circuit files for SPICE although additional editing of the netlist is necessary to specify the simulation conditions. Since the post-layout simulations provide a more realistic view of the circuit's performance, only their results will be presented.

The transient simulations of the extracted netlist are more prone to convergence problems than the pre-layout simulations. SPICE analyzes circuits by solving the matrix expressions for the node voltages and loop currents. Each parasitic capacitor in the layour becomes an element in the matrix. If a node is left floating or the transient charge stored in the capacitors is too small, the matrix becomes singular and the analysis diverges. In order to insure convergence, 10MΩ shunt resistors were inserted at each node to prevent them from floating. In addition, different tolerance parameters were specified to limit the number of calculation iterations such that divergent calculations are bypassed. After some initial experimentation, the following values were chosen:

- abstol=1e-10 = the absolute current error tolerance
- chgtol=1e-8 = the minimum charge tolerance
- reltol=0.05 = the relative error tolerance per iteration
- pivtol=1e-6 = the minimum value for a matrix element treated as a pivot

- trtol=1.0 = the transient error tolerance
- gmin=5e-8 = the minimum conductance value

The following timing diagrams depict how the memory array would actually be cycled in practical applications. While the address signals sequentially cycle through the four cells, each cell follows through an Erase/Read/Write/Read cycle. Referring to the timing diagrams, we see that A0 initally strobed high which implies that row 1 is addressed. During this time A1 is also strobed high, addressing cell (1,1). This cell then follows through a complete test cycle in which it is erased, read, written and read again. When A1 is strobed low, cell (1,2) is addressed, and also follows the same Erase/Read/Write/Read cycle. When A0 is pulsed low, row 2 is selected and cells (2,1) and (2,2) undergo the same programming sequence.

The key nodes that are examined in the timing diagrams are the Word Lines, Bit Lines, Select Lines, and the input and output terminals of the sense amplifier. since they are the input/output terminals of the memory cells. Each of these nodes is also directly accessible in the layout since there are separate bonding pads designated for each of them.

The timing simulations verify the functionality of the circuit, but indicate some nonidealities that must be considered. In both the row and column decoders, there is a small voltage drop in the Vp and ground signal paths due to the finite resistance therein. Thus, the programming time estimations based on discrete device measurements must be increased to compensate for the effective decrease in programming voltage.

In a practical implementation of the array, 5V programming requires programming pulse widths on the order of 10ms or 100ms to achieve an acceptable programming window ($\sim$ 1V). The simulations that use these pulse widths appear to have negligible propagation delays as well as rise and fall times. However, in order

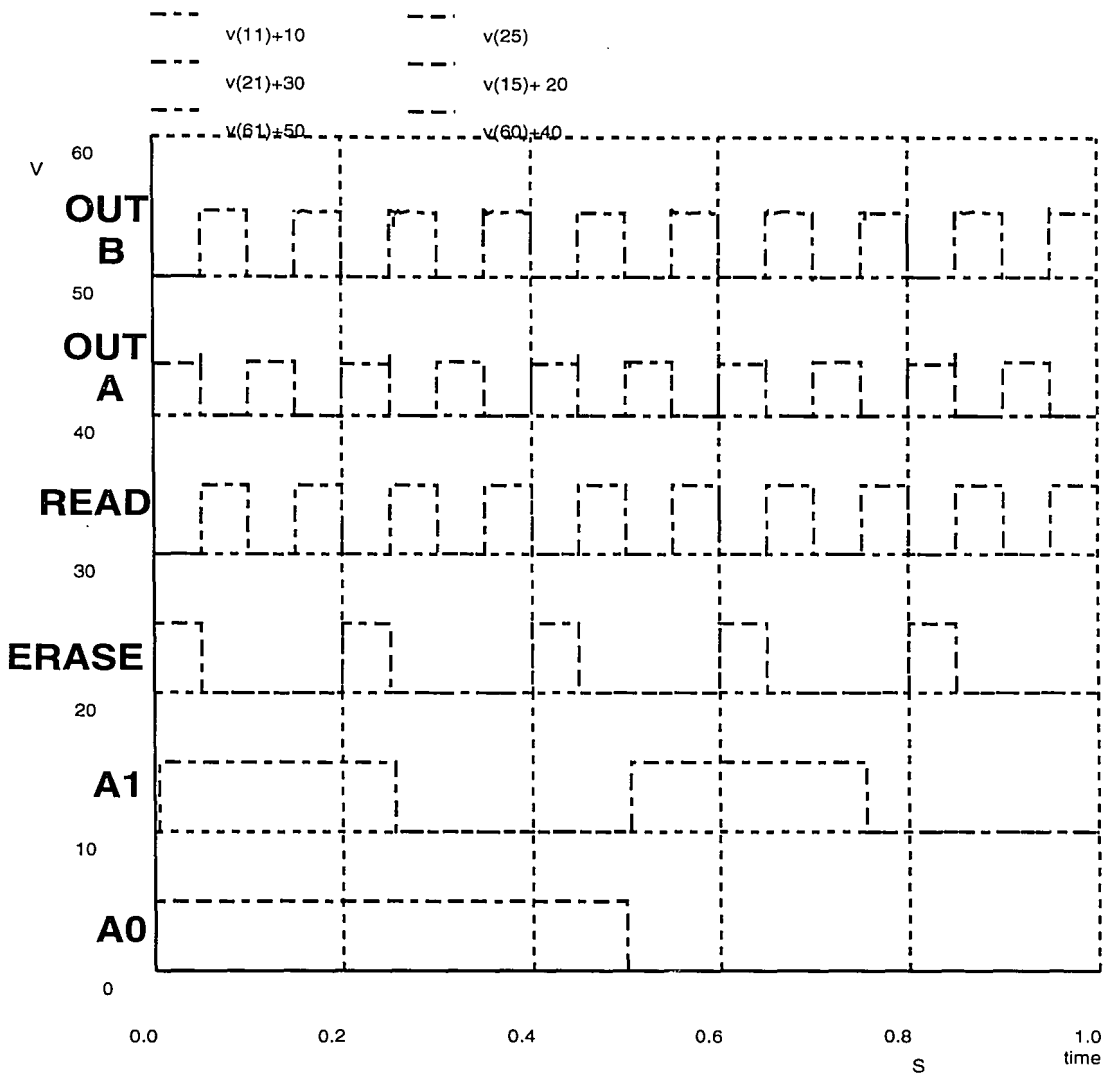


Figure 3.12: A timing diagram which demonstrates the flip flop sense amplifier's ability to convert the read current into a digital compatible voltage. Output A is inverted to form Output B because the former toggles to 0V when the memory device is in the high conduction state, which we have previously defined to be the logic 1 state.

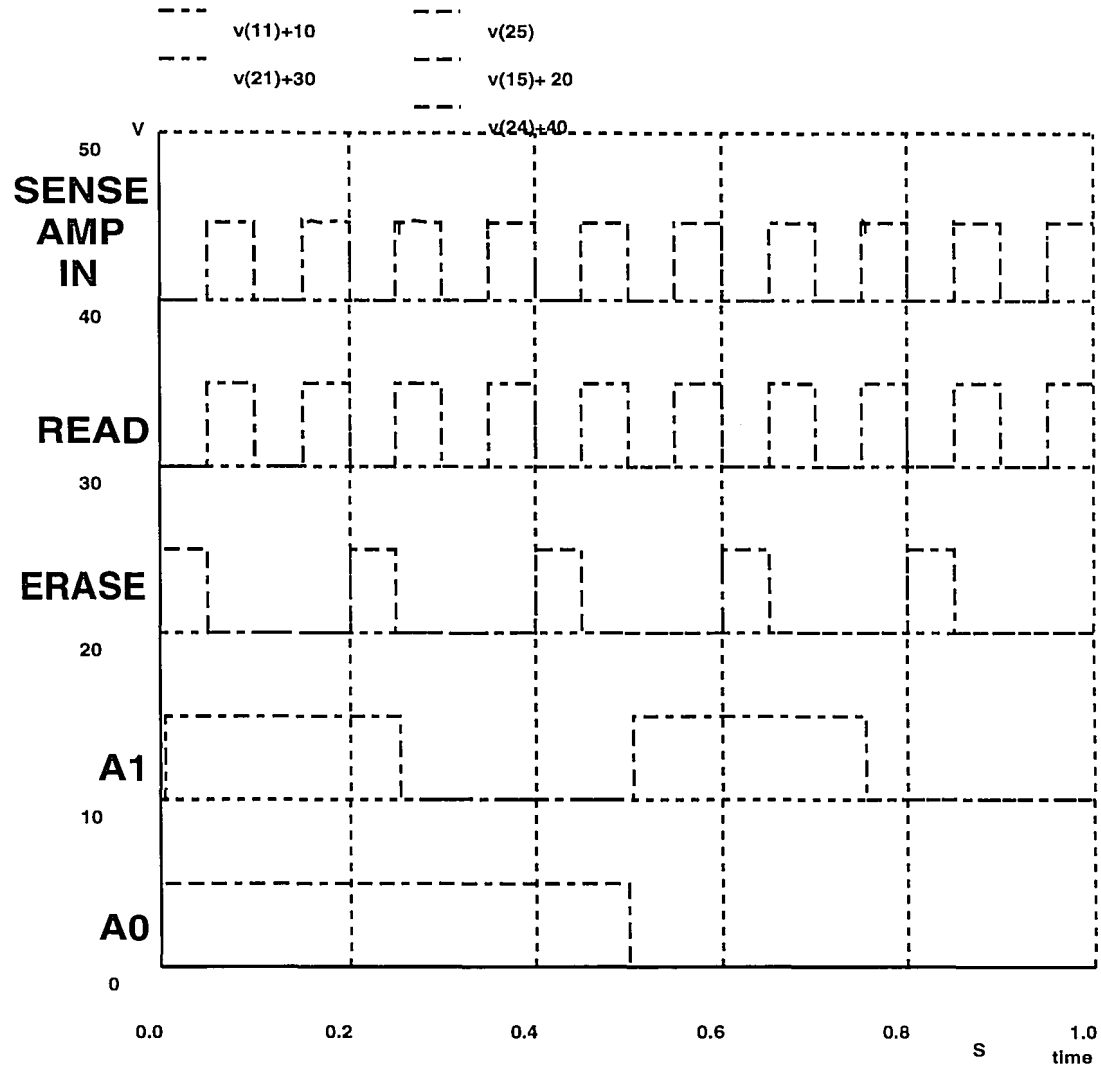


Figure 3.13: A timing diagram which shows the read current charging the input capacitor of the flip-flop sense amplifier to a digital voltage. Because SPICE requires all the device parameters to be predefined prior to simulation, the memory threshold is preset to high conduction state independent of the programming modes. Thus, the sense amp input always charges to 5V during the Read cycle.

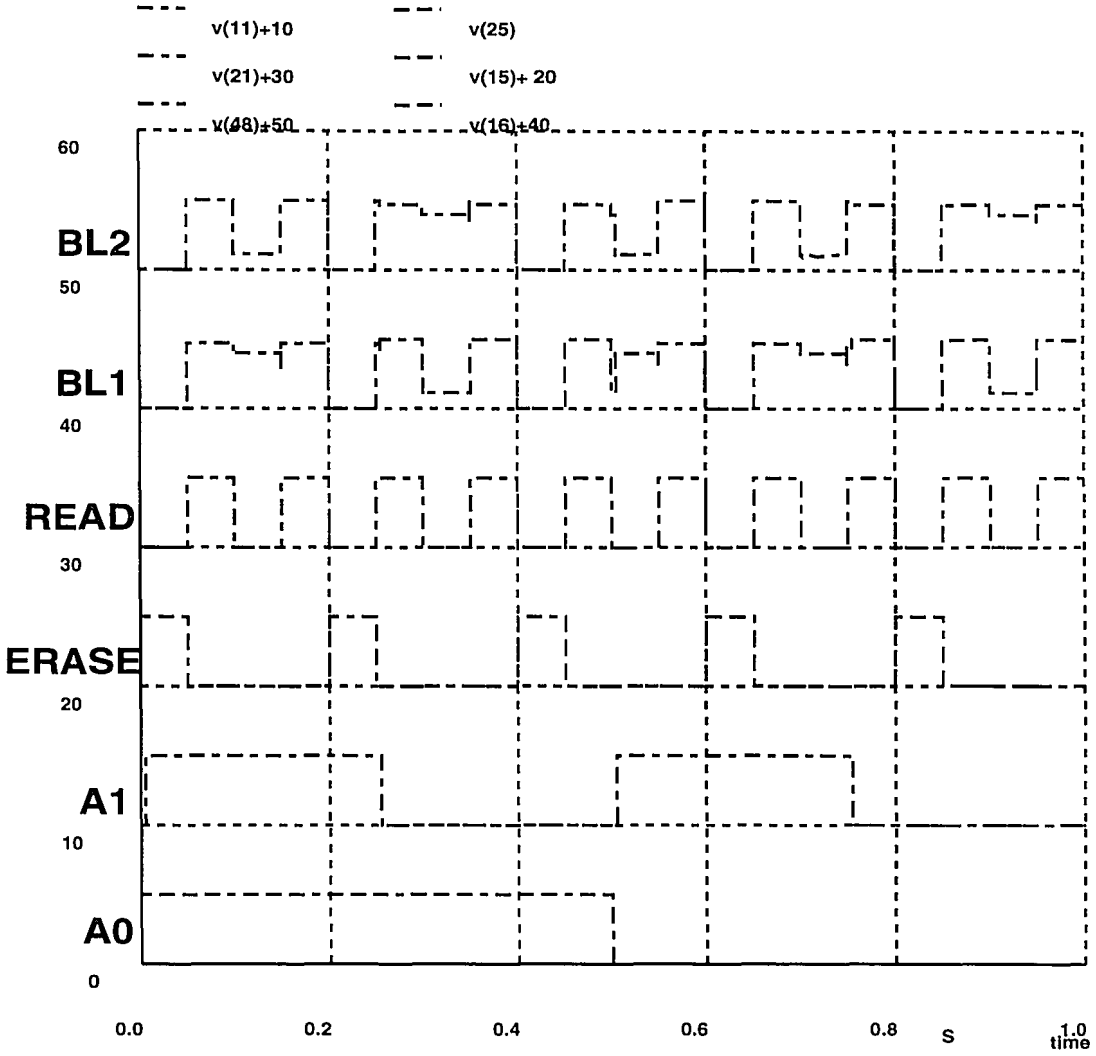


Figure 3.14: A timing diagram showing the Bit Line outputs. When the column address signal, A1, strobes high, BL1 is selected and undergoes the appropriate Write, Erase and Read voltages while depicting the different Inhibit voltages when A1 strobes low. The converse is true for BL2.

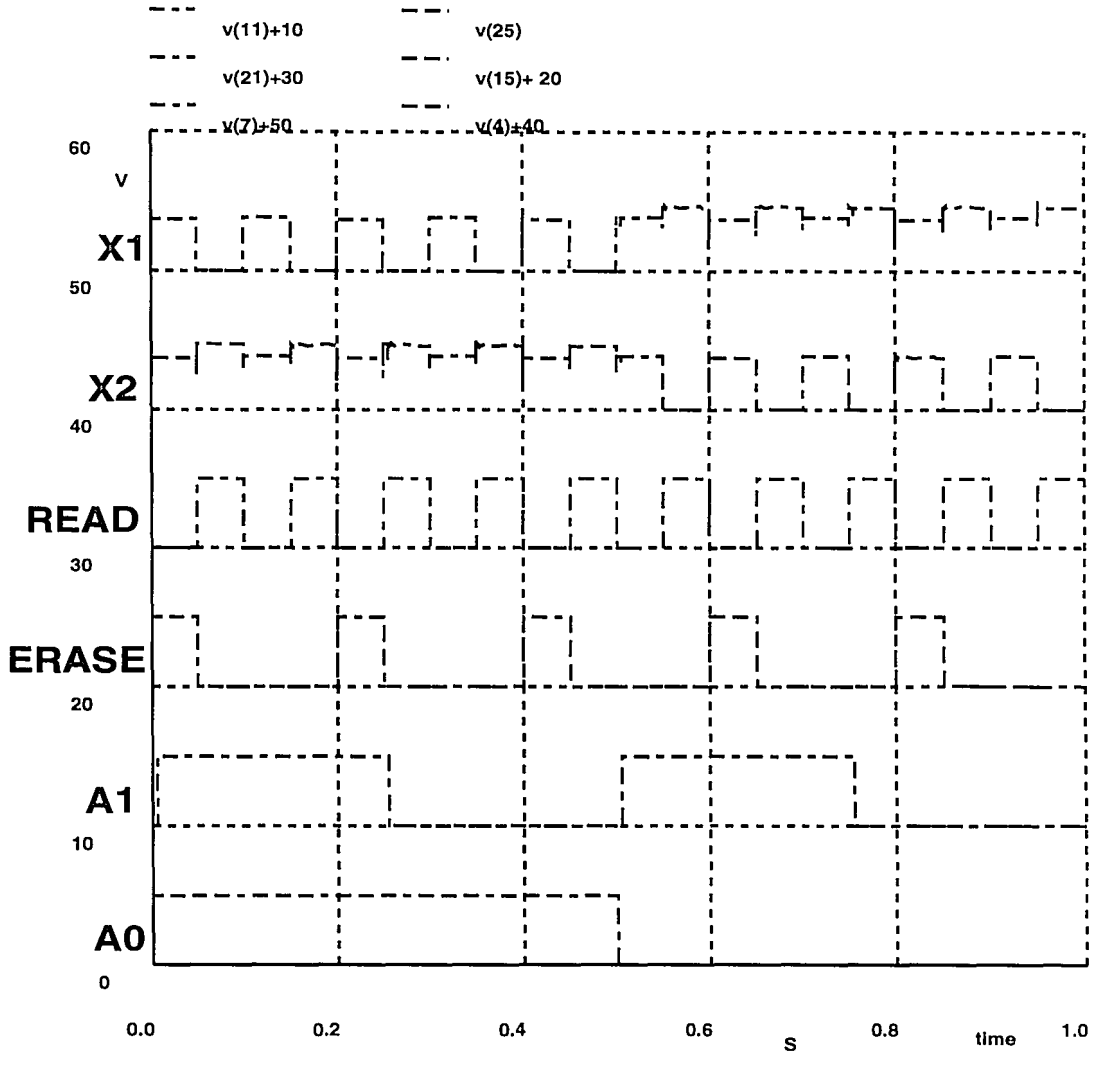


Figure 3.15: Timing diagram for the Select Lines. When A0 strobes high, row 1 is selected and X1 activates the series PMOS select transistors only when a memory cell in row 1 is read, else it inhibits their conduction by maintaining 5V on the memory gates. X2 behaves similarly when A0 strobes low.

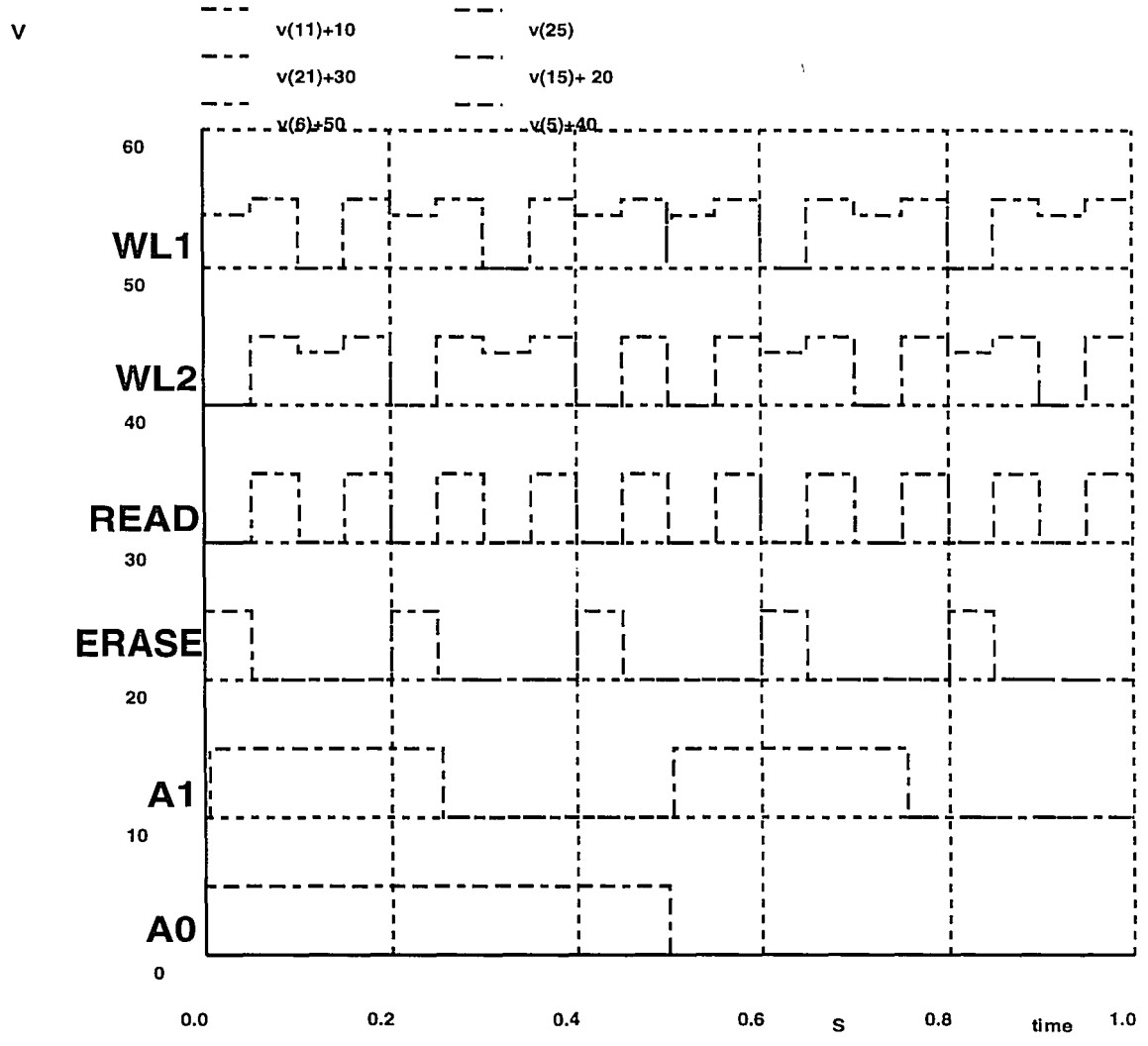


Figure 3.16: Timing diagram for the Word Lines. When A0 is high, row 1 is selected and WL1 depicts the voltages for the Erase, Read and Write operations, respectively, while WL2 inhibits row 2. The strobing of A1 during this time determines which particular cell of row 1 is addressed. When A0 is low, row 2 is selected while row 1 is inhibited, hence WL2 reflects the applied Erase, Read and Write voltages whereas WL1 shows the different inhibit voltages.

to determine the maximum possible programming speeds (even if they may not be physically realizable), pulse widths on the order of 50μs were used. In these simulations, it was noted that the rise and fall times of the outputs were approximately 30μs. Since, the programming pulse width should be significantly greater than the rise and fall times, the minimum pulse width should be no less than 150μs. While 150μs Erase and Write pulses are unrealistic for 5V programming, this figure of merit defines the minimum possible time to read a memory cell.

Chapter 4

Conclusions

4.1 Conclusions and Recommendations for Future Work

The attainment of true 5V programmability for EEPROMs is crucial to developing a highly reliable EEPROM and semiconductor disk technology. This thesis examined the some of the considerations regarding the memory window design to achieve 5V programmable SONOS devices and showed how they can be implemented in a small memory array. A CMOS addressing/decoding circuit was designed to switch 5V and 0V to the memory device terminals in order to achieve the bipolar programming voltages required to store and read data. The memory array design layout was implemented using the Mentor Graphics GDT software and was successfully simulated using the SPICE 3c1 program. The layout design rules were based upon our custom n-well CMOS/SONOS process sequence, thus necessitating the use of p-channel SONOS transistors to provide the memory well isolation. With the completion of the layout mask set, the circuit will fabricated and tested in-house.

While true 5V programmable nonvolatile memories is a worthy goal, it is important to note that the evolution of laptop and notebook computers will eventually necessitate scaling the programming voltage down to $\sim 1.3V$, where the nonvolatile memories can be powered by a nickel-cadmium battery. Before such SONOS EEPROMs can be implemented, however, the basic physics and technology of discrete devices must be investigated in order to produce memory elements which can operate

with this voltage.

There is also further work that must be done to improve our SONOS processing technology on the layout/circuit level. Theoretically, it is possible to design a $1\mu m^2$ SONOS memory cell using 0.25μm feature sizes, but this requires state of the art technology that is unavailable to us. To generate the required precision for small geometry (sub-micron) patterns, it is necessary to employ both electron beam lithography and reactive ion etching. Our in-house processing technology is capable only of contact lithography and wet chemical etching, both of which obscure the pattern resolution and thus, necessitate larger feature sizes on the mask set in order to produce reliable devices and circuits. If access to the high resolution processing equipment is permitted in the near future, then it is possible to produce a SONOS EEPROM with a 1μm cell size.

Further improvements need to be made on the effectiveness of the SPICE simulations, as well. The current simulations demonstrate the functionality of the addressing and sensing circuitry and show that the slewing of the programming pulses caused by the parasitic capacitances is minimal compared to the programming time. However, SPICE cannot model the transient behavior of the SONOS device itself because the threshold voltage is predefined as a constant parameter. Some efforts have made to skirt around this problem by artificially altering the threshold voltage via inserting a capacitor parallel to the gate dielectric and charging it with a current source whose magnitude is dependent upon the gate to bulk voltage. The stored charge can alter the threshold voltage in a similar manner as the silicon nitride layer of a SONOS device. While SPICE is flexible in allowing nonlinear dependence on its current sources, it cannot incorporate time dependence into the voltage-controlled current source expression. Hence, an accurate equivalent circuit for a SONOS device cannot be developed using conventional SPICE to depict the transient change in threshold voltage when there is a gate bias.

With the recent acquisition of the XSPICE software from GeorgiaTech, there is

renewed hope in creating a SPICE compatible model for SONOS. XSPICE enables the user to define a new device from its fundamental behavioral equations. Some initial investigations have shown promise in modeling the time dependent threshold voltage shift, and when the actual SONOS model is completed, it can be incorporated into the circuit simulations to attain a complete transient analysis of the memory array.

One final challenge is offered to the future researcher: to improve the scaling of the memory cell by eliminating the need for the select transistor. Select transistors prevent erroneous programming at the cost of increasing the cell area. While it is possible to reduce this area by overlapping the gates of the two transistors, it is certain that the future of SONOS EEPROM scaling lies in the development of a true single transistor memory cell.

References

[1] Yasuo Itoh, Masaki Momodomi, Riichiro Shirota, Yoshihisa Iwata, Ryozo Nanayama. An Experimental 4Mb CMOS EEPROM with a NAND Structured Cell. *IEEE International Solid-State Circuits Conference*, 1989.

[2] Sebastiano D'Arrigo, Giuliano Santin, Manzur Gill, Rinn Cleavelin, Stefano Spagliccia. A 5V-Only 256kBit CMOS Flash EEPROM. *IEEE International Solid-State Conference*, 1989.

[3] Richard D. Pashley, Stefan K. Lai. Flash Memories: the Best of two Worlds. *IEEE Spectrum*, December 1989.

[4] M.L. French, H. Sathianathan ,and M.H. White. A SONOS Nonvolatile Memory Cell for Semiconductor Disk Application. *Nonvolatile Semiconductor Memory Technology Review*, 1993.

[5] M.L. French, C.-Y. Chen and M.H. White. New Results on Scaled SONOS Nonvolatile Memory Devices. *1992 IEEE Non-Volatile Semiconductor Memory Workshop*, 1992.

[6] A.K. Banerjee. Automated Measurements of Memory Devices. Master's thesis, Lehigh University, May 1993.

[7] Chenming Hu. *Nonvolatile Semiconductor Memories: Technologies, Design and Applications.* IEEE Press, 1991.

[8] Raul-Adrian Cernes, Gheorghe Samachisa, Chien-sheng Su, Hui-Fang Tsai. A 1Mb Flash EEPROM. *IEEE International Solid-State Circuits Conference*, 1989.

[9] Georghe Samachisa, Chien-Sheng Su, Yu-Sheng Kao George Smarandoiu. A 128K Flash EEPROM using Double Polysilicon Technology. *IEEE International Solid-State Circuits Conference*, 1987.

[10] Virgil Kynett, Jim Anderson, Greg Atwood, Pat Dix, Mick Fandrich. A 90ns 100K Erase/Program Cycle Megabit Flash Memory. *IEEE International Solid State Circuits Conference*, 1989.

[11] Satyen Mukherjee, Thomas Chang, Richard Pang, Mark Knecht, Dan Hu. A Single Transistor EEPROM Cell and its Implementation in a 512K CMOS EEPROM. *IEDM Digest*, 1985.

[12] Yasushi Terada, Kazuo Kobayashi, Takeshi Nakayama, Masanori Hayashikori. 120ns 128k x 8b / 64k x 16b CMOS EEPROMs. *IEEE International Solid-State Circuits Conference*, 1989.

[13] D. Guterman, B. Houch, L. Starnes, B. Yeh. New Ultra-High Density Textured Poly-Si Floating Gate EEPROM Cell. *IEDM Digest*, pages 826–828, 1986.

[14] S. K. Lai and V. K. Dham. Comparison and Trends in Today's Dominant EEPROM Technologies. *IEDM*, pages 580–583, 1986.

[15] Yoshiaki Kamigaki, Shin-Ichi Minami, Takaaki Hagiwara, Kazunori Furusawa, Takeshi Furuno, Ken Uchida, Masaaki Terasawa and Koubu Yamazaki. Yield and Reliability of MNOS EEPROM Products. *IEEE Journal of Solid-State Circuits*, 24(6):1714–1722, December 1989.

[16] P. C. Y. Chen. Threshold-Alterable Si-Gate MOS Devices. *IEEE Transactions on Electron Devices*, ED-24(5):584–586, May 1977.

[17] A.K. Agarwal. *Study of Traps in the Metal-Insulator-Semiconductor (MIS) System with the Three Terminal Gated-Diode Structure.* PhD thesis, Lehigh University, 1984.

[18] E. Suzuki, H. Hiraishi, K. Ishii and Y. Hayashi. A Low Voltage Alterable EEPROM with Metal-Oxide-Nitride-Oxide-Semiconductor (MONOS) Structures. *IEEE Transactions on Electron Devices*, ED-30(2):122–128, 1983.

[19] C. Chao and M. H. White. Charaterization of Charge Injection and Trapping in Scaled SONOS/MONOS Memory Devices. *Solid State Electronics*, 30(3):307–319, 1987.

[20] F. R. Libsch. A Computerized Dynamic Characterization for MNOS Transistors. Master's thesis, Lehigh University, 1984.

[21] Yin Hu. *Physics and Characterization of the Reliability of Nonvolatile Semiconductor Memory Devices.* PhD thesis, Lehigh University, October 1992.

[22] Yoshiaki Kamigaki, Shin-ichi Minami, Kazunori Furusawa, Ken Uchida, Takaki Hagiwara. New Scaling Guidelines for MNOS Memory Devices. *10th IEEE Non-Volatile Semiconductor Memory Workshop*, 1989.

[23] Takaaki Nozaki, Toshiaki Tanaka, Yoshiro Kijiya, Eita Kinoshita, Tatsuo Tsuchiya, Yutaka Hayashi. A 1-Mb EEPROM with MONOS Memory Cell for Semiconductor Disk Application. *IEEE Journal of Solid-State Circuits*, 4(26):497–501, April 1991.

[24] A. Roy. Retention , Endurance and Interface Traps in MONOS Memory Transistors. Master's thesis, Lehigh University, October 1985.

[25] Y. Hu, A. K. Banerjee and M. H. White. Characterization of Si/SiO2 Interface States in Ultra-Thin Oxides. *23rd Semiconductor Interface Specialist Conference*, December 1992.

Appendix A

CMOS/SONOS Fabrication Sequence

1. Starting Material 3 in p-type 2-3 Ohm/cm

1. N-Well Implant Formation

   - Wet Oxidation for 1000 Å, 950 °C, 25 min
   - Mask Level **NW**
   - BHF Etch, 10:1, 2 min
   - Implant, Phosphorus, 4.8×10^{12}, 100KeV
   - Dry Oxidation, 500 Å, 1200 °C, 5 min
   - Implant Anneal, 1200 °C, 240 min

1. Active Device

   - LPCVD Nitride, 200mTorr, 10:1 ratio, 54 min, 1000 Å
   - Mask Level **AD**
   - Plasma Etch Nitride (CF_4)

1. Channel Stop Implant

   - Mask Level **FI**
   - Implant, BF_2, 5×10^{11}, 145KeV
   - Wet Field Oxidation, 6500 Å, 1100 °C, 60 min

- BHF Etch, 10:1, 1 min
- Hot H_3PO_4, 170°C, 35 min
- BHF ETCH, 10:1, 1.5 min
- Wet Oxidation, 900 °C, 20 min
- BHF Etch, 10:1, 1 min
- Wet Pad Oxidation, 900 °C, 15 min
- Implant, Boron, 9×10^{11}, 70KeV
- Anneal, 950 °C, 30 min

1. Gate Dielectric Formation

- Triple Wall Dry Oxidation, 900 °C, 800 Å
- Mask Level **MW**
- BHF Etch, 10:1, 2 min
- Triple Wall Dry Oxidation, 720 °C, 20 Å, 9 min
- LPCVD Nitride, 250 mTorr, 735 °C, 120 Å, 5 min, 10:1
- Wet Blocking Oxidation, 1000 °C, 40 Å, 50 min

1. Gate Material

- LPCVD Polysilicon, 800 mTorr, 180 sccm SiH_4, 625 °C, 30 min, 5000 Å
- $POCl_3$ Doping, 900 °C, 25 min Pre-deposition, 30 min Drive-in
- BHF Etch, 10:1 , 15 sec.
- Mask Level **PY**
- Plasma Polysilicon/Gate Dielectric Etch (SF_6)
- BHF Etch, 100:1, 1 min

- Hot H_3PO_4 Etch, 170 °C, 3.5 min
- BHF Etch, 100:1, 1 min

1. Source/Drain Formation

   - Dry Pad Oxidation, 900 °C, 15 min, 200-300 Å
   - Mask Level **N+**
   - Source/Drain Implant, Phosphorus, 2×10^{15}, 100KeV
   - Mask Level **P+**
   - Source/Drain Implant, Boron, 5×10^{15}, 30KeV
   - Anneal and Drive-in, 950 °C, 60 min
   - BHF Etch, 10:1, 1 min

1. Contact Window Formation

   - Wet Oxidation, 900 °C, 30 min, 1000 Å
   - Mask Level **CW**
   - BHF Etch, 10:1, 3-5 min
   - Dilute HF Etch (HF Dip), 30 sec

1. Metallization

   - RF Sputtering Aluminum, 110 mTorr, 60 min
   - Mask Level **MET**
   - PAN Etch, 45 °C, 2 min
   - Backside Clean-up
   - Backside RF Sputtering Aluminum, 110 mTorr, 60 min
   - PMA, 450 °C, 30 min

Appendix B

Sample Spice Code

The sample SPICE code below simulates the programming sequence of the memory array. The ciruit file is generated from the layout itself.

B.1 Sample SPICE code

```
*
*
*                          2/26/93
*
*
*
*
*
*.TECH lehigh-sfc-new

* SPICE net list for CELL masters
* Ledi: Version 5.2_1.10 sun4os4 Built Tue Nov  3 08:06:32 EST 1992


* start main CELL masters
Mcolumn_try2_i0_tpo_nand 9 8 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tp1_nand 9 11 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn0_name 12 8 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tn1_nand 9 11 12 107 TN W=24U L=8U
Mcolumn_try2_i0_tp2_inv 13 9 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn2_inv 13 9 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tp3_tgate 14 9 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn3_tgate 14 13 10 107 TN W=24U L=8U
Mcolumn_try2_i0_tp0_inv_erase 16 15 14 108 TP W=24U L=8U
Mcolumn_try2_i0_tn0_inv_erase 16 15 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tn1 17 16 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tn2 18 8 17 107 TN W=24U L=8U
Mcolumn_try2_i0_tn3 16 19 18 107 TN W=24U L=8U
Mcolumn_try2_i0_tp1_nand2 20 11 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tp2_nand2 20 21 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn4_nand2 22 11 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tn5_nand2 20 21 22 107 TN W=24U L=8U
Mcolumn_try2_i0_tp6_inv2 23 20 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn7_name 23 20 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tp1_tgate2 24 20 16 108 TP W=24U L=8U
Mcolumn_try2_i0_tn4_tgate2 24 23 16 107 TN W=24U L=8U
Mcolumn_try2_i0_tp0 8 21 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn0 8 21 3 107 TN W=24U L=8U
Mcolumn_try2_i0_tp1 19 11 10 108 TP W=24U L=8U
Mcolumn_try2_i0_tn4 19 11 3 107 TN W=24U L=8U
```

```
Mrow2_i0_tp0_X1_nand 7 21 10 108 TP W=24U L=8U
Mrow2_i0_tp1_X1_nand 7 25 10 108 TP W=24U L=8U
Mrow2_i0_tn0_X1_nand 26 21 3 107 TN W=24U L=8U
Mrow2_i0_tn1_X1_nand 7 25 26 107 TN W=24U L=8U
Mrow2_i0_tp2 6 27 28 108 TP W=24U L=8U
Mrow2_i0_tp3 6 29 30 108 TP W=24U L=8U
Mrow2_i0_tp4 30 21 10 108 TP W=24U L=8U
Mrow2_i0_tn2 6 29 31 107 TN W=24U L=8U
Mrow2_i0_tn3 31 27 3 107 TN W=24U L=8U
Mrow2_i0_xor2_1_i0_tp0 15 32 33 108 TP W=24U L=8U
Mrow2_i0_xor2_1_i0_tp1 33 15 32 108 TP W=24U L=8U
Mrow2_i0_xor2_1_i0_tn0 15 25 33 107 TN W=24U L=8U
Mrow2_i0_xor2_1_i0_tn1 33 15 25 107 TN W=24U L=8U
Mrow2_i0_xor2_1_i0_tp2 32 25 10 108 TP W=24U L=8U
Mrow2_i0_xor2_1_i0_tn2 32 25 3 107 TN W=24U L=8U
Mrow2_i0_xor2_1_i0_tp3 29 33 10 108 TP W=24U L=8U
Mrow2_i0_xor2_1_i0_tn3 29 33 3 107 TN W=24U L=8U
Mrow2_i0_tp0 27 21 10 108 TP W=24U L=8U
Mrow2_i0_tn0 27 21 3 107 TN W=24U L=8U
Mrow2_i0_tp1 34 25 10 108 TP W=24U L=8U
Mrow2_i0_tn1 34 25 3 107 TN W=24U L=8U
Mrow2_i1_tp0_X1_nand 4 21 10 108 TP W=24U L=8U
Mrow2_i1_tp1_X1_nand 4 34 10 108 TP W=24U L=8U
Mrow2_i1_tn0_X1_nand 35 21 3 107 TN W=24U L=8U
Mrow2_i1_tn1_X1_nand 4 34 35 107 TN W=24U L=8U
Mrow2_i1_tp2 5 36 28 108 TP W=24U L=8U
Mrow2_i1_tp3 5 37 38 108 TP W=24U L=8U
Mrow2_i1_tp4 38 21 10 108 TP W=24U L=8U
Mrow2_i1_tn2 5 37 39 107 TN W=24U L=8U
Mrow2_i1_tn3 39 36 3 107 TN W=24U L=8U
Mrow2_i1_xor2_1_i0_tp0 15 40 41 108 TP W=24U L=8U
Mrow2_i1_xor2_1_i0_tp1 41 15 40 108 TP W=24U L=8U
Mrow2_i1_xor2_1_i0_tn0 15 34 41 107 TN W=24U L=8U
Mrow2_i1_xor2_1_i0_tn1 41 15 34 107 TN W=24U L=8U
Mrow2_i1_xor2_1_i0_tp2 40 34 10 108 TP W=24U L=8U
Mrow2_i1_xor2_1_i0_tn2 40 34 3 107 TN W=24U L=8U
Mrow2_i1_xor2_1_i0_tp3 37 41 10 108 TP W=24U L=8U
Mrow2_i1_xor2_1_i0_tn3 37 41 3 107 TN W=24U L=8U
Mrow2_i1_tp0 36 21 10 108 TP W=24U L=8U
Mrow2_i1_tn0 36 21 3 107 TN W=24U L=8U
Mrow2_i1_tp1 42 34 10 108 TP W=24U L=8U
Mrow2_i1_tn1 42 34 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tpo_nand 44 43 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tp1_nand 44 19 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn0_name 45 43 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tn1_nand 44 19 45 107 TN W=24U L=8U
Mcolumn_try2_i2_tp2_inv 46 44 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn2_inv 46 44 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tp3_tgate 47 44 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn3_tgate 47 46 10 107 TN W=24U L=8U
Mcolumn_try2_i2_tp0_inv_erase 48 15 47 108 TP W=24U L=8U
Mcolumn_try2_i2_tn0_inv_erase 48 15 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tn1 49 48 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tn2 50 43 49 107 TN W=24U L=8U
Mcolumn_try2_i2_tn3 48 51 50 107 TN W=24U L=8U
Mcolumn_try2_i2_tp1_nand2 52 19 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tp2_nand2 52 21 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn4_nand2 53 19 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tn5_nand2 52 21 53 107 TN W=24U L=8U
Mcolumn_try2_i2_tp6_inv2 54 52 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn7_name 54 52 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tp1_tgate2 24 52 48 108 TP W=24U L=8U
Mcolumn_try2_i2_tn4_tgate2 24 54 48 107 TN W=24U L=8U
Mcolumn_try2_i2_tp0 43 21 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn0 43 21 3 107 TN W=24U L=8U
Mcolumn_try2_i2_tp1 51 19 10 108 TP W=24U L=8U
Mcolumn_try2_i2_tn4 51 19 3 107 TN W=24U L=8U
Marray2_layout_i0_tp0 55 6 16 108 psonos W=20U L=8U
Marray2_layout_i0_tp1 55 7 56 108 TP W=20U L=8U
Marray2_layout_i0_tp2 57 7 56 108 TP W=20U L=8U
```

```
Marray2_layout_i0_tp3 57 6 48 108 psonos W=20U L=8U
Marray2_layout_i0_tp4 16 5 58 108 psonos W=20U L=8U
Marray2_layout_i0_tp5 58 4 56 108 TP W=20U L=8U
Marray2_layout_i0_tp6 56 4 59 108 TP W=20U L=8U
Marray2_layout_i0_tp7 59 5 48 108 psonos W=20U L=8U
v_dummy 24 240 0v
Msense2_i0_tn0 60 240 3 107 TN W=24U L=8U
Msense2_i0_tn1 240 60 3 107 TN W=24U L=72U
Msense2_i0_tp0 60 240 10 108 TP W=24U L=8U
Msense2_i0_tp1 240 60 10 108 TP W=24U L=8U
Msense2_i0_tp2_precharge_sense_amp 60 21 10 108 TP W=24U L=8U
Msense2_i0_tn2 240 8 3 107 TN W=24U L=8U
Msense2_i0_tn3_inv_sense_amp 61 60 3 107 TN W=24U L=8U
Msense2_i0_tp3_p_inv_sense_amp 61 60 10 108 TP W=24U L=8U
*.INOUT inout0 1
*.INOUT inout1 2
*.GND gnd0 3
*.INOUT inout2 4
*.INOUT inout3 5
*.INOUT inout4 6
*.INOUT inout5 7
Vnwell 108 0 5
Vbulk 107 0 0
*
*
*
.OPTIONS abstol=0.10n chgtol=1e-8 reltol=0.05 pivtol=1e-6 trtol=1.0 gmin=5e-8
*
*
cload 24 0 4pf
cout 61 0 4pf
*
v_read 21 0 pulse(0v 5v 50ms 0 0 100us 50.1ms)
v_a0 25 0 pulse(0v 5v 0 0 0 500ms 1000ms)
v_a1 11 0 pulse(0v 5v 0ms 0 0 250ms 500ms)
v_erase 15 0 pulse(0v 5v 0ms 0 0 50ms 100.2ms)
v_cs 56 0 5v
v_r 28 0 5v
.tran 50us 500ms
*
*
*
*********************************************************************************
.model tn nmos level=3 vto=0.78 kp=56.1e-6 gamma=0.597 phi=0.692
+ is=5e-16 pb=0.8 cgso=1.95e-10 cgdo=1.95e-10 rsh=36 cj=2.50e-4 mj=0.76
+ cjsw=5e-10 mjsw=.34 js=1e-5 tox=400e-10 nsub=8e15 nss=1e11 nfs=1e12
+ tpg=+1 xj=0.15u ld=0.20u uo=650 vmax=5.1e4 fc=0.5 delta=1.4 theta=0.06
+ eta=0.06 kappa=0.4

.model tp pmos level=3 vto=-0.75 kp=22.0e-6 gamma=0.517 phi=0.677
+ is=5e-16 pb=0.8 cgso=1.90e-10 cgdo=1.90e-10 rsh=100 cj=1.95e-4 mj=0.535
+ cjsw=3.5e-10 mjsw=.30 js=1e-5 tox=400e-10 nsub=6e15 nss=1e11 nfs=1e12
+ tpg=-1 xj=0.05u ld=0.20u uo=255 vmax=3.0e4 fc=0.5 delta=1.0 theta=0.06
+ eta=0.06 kappa=0.4
*********************************************************************************
*
.model psonos pmos level=3 vto=+1.50v kp=22.0e-6 gamma=0.517 phi=0.677
+ is=5e-16 pb=0.8 cgso=1.90e-10 cgdo=1.90e-10 rsh=100 cj=1.95e-4 mj=0.535
+ cjsw=3.5e-10 mjsw=.30 js=1e-5 tox=100e-10 nsub=6e15 nss=1e11 nfs=1e12
+ tpg=-1 xj=0.05u ld=0.20u uo=255 vmax=3.0e4 fc=0.5 delta=1.0 theta=0.06
+ eta=0.06 kappa=0.4
*********************************************************************************
*
*
*
C4 4 0 2646.92e-15
C5 5 0 2559.77e-15
C6 6 0 3158.72e-15
C7 7 0 3397.33e-15
C8 8 0 1169.12e-15
```

```
C9 9 0 331.35e-15
C10 10 0 5986.50e-15
C11 11 0 926.42e-15
C12 12 0 77.00e-15
C13 13 0 136.89e-15
C14 14 0 81.47e-15
C15 15 0 2007.32e-15
C16 16 0 1388.26e-15
C17 17 0 80.08e-15
C18 18 0 80.85e-15
C19 19 0 1159.51e-15
C20 20 0 347.13e-15
C21 21 0 4784.62e-15
C22 22 0 77.00e-15
C23 23 0 161.76e-15
C24 24 0 2626.66e-15
C25 25 0 1656.19e-15
C26 26 0 79.31e-15
C27 27 0 311.76e-15
C28 28 0 3214.56e-15
C29 29 0 396.09e-15
C30 30 0 80.08e-15
C31 31 0 80.85e-15
C32 32 0 327.67e-15
C33 33 0 258.13e-15
C34 34 0 2519.76e-15
C35 35 0 79.31e-15
C36 36 0 311.76e-15
C37 37 0 396.78e-15
C38 38 0 80.08e-15
C39 39 0 80.85e-15
C40 40 0 327.67e-15
C41 41 0 258.13e-15
C42 42 0 133.49e-15
C43 43 0 368.59e-15
C44 44 0 331.35e-15
C45 45 0 77.00e-15
C46 46 0 129.97e-15
C47 47 0 81.47e-15
C48 48 0 1787.40e-15
C49 49 0 80.08e-15
C50 50 0 80.85e-15
C51 51 0 293.90e-15
C52 52 0 347.13e-15
C53 53 0 77.00e-15
C54 54 0 161.86e-15
C55 55 0 56.55e-15
C56 56 0 1184.90e-15
C57 57 0 54.60e-15
C58 58 0 55.90e-15
C59 59 0 53.30e-15
C60 60 0 495.87e-15
C61 61 0 758.85e-15
VGND0 3 0 0
*
*
r1 1 0 10meg
r2 2 0 10meg
r3 3 0 10meg
r4 4 0 10meg
r5 5 0 10meg
r6 6 0 10meg
r7 7 0 10meg
r8 8 0 10meg
r9 9 0 10meg
r10 10 0 10meg
r11 11 0 10meg
r12 12 0 10meg
r13 13 0 10meg
r14 14 0 10meg
```

```
r15 15 0 10meg
r17 17 0 10meg
r18 18 0 10meg
r19 19 0 10meg
r20 20 0 10meg
r21 21 0 10meg
r22 22 0 10meg
r23 23 0 10meg
r24 24 0 10meg
r25 25 0 10meg
r26 26 0 10meg
r27 27 0 10meg
r28 28 0 10meg
r29 29 0 10meg
r30 30 0 10meg
r31 31 0 10meg
r32 32 0 10meg
r33 33 0 10meg
r34 34 0 10meg
r35 35 0 10meg
r36 36 0 10meg
r37 37 0 10meg
r38 38 0 10meg
r39 39 0 10meg
r40 40 0 10meg
r41 41 0 10meg
r42 42 0 10meg
r43 43 0 10meg
r44 44 0 10meg
r45 45 0 10meg
r46 46 0 10meg
r47 47 0 10meg
r48 48 0 10meg
r49 49 0 10meg
r50 50 0 10meg
r51 51 0 10meg
r52 52 0 10meg
r53 53 0 10meg
r54 54 0 10meg
r55 55 0 10meg
r56 56 0 10meg
r57 57 0 10meg
r58 58 0 10meg
r59 59 0 10meg
r60 60 0 10meg
r61 61 0 10meg
*
*
.END
```

Vitae

Harikaran Sathianathan was born on August 8, 1969 in the Maldive Islands to Dr. Manickavasagar and Mrs. Sivadevi Sathianathan. After attending various English medium schools throughout Asia, he joined Lehigh University in 1987. While enrolled as an undergraduate, he was initiated into the Tau Beta Pi engineering honor society and received the Sherman Fairchild Summer Fellowship for Solid State Research for two consecutive summers. After graduating with a B.S. degree in Electrical Engineering in 1991, he continued on with his graduate studies under Dr. Marvin White in the area of nonvolatile semiconductor semicondcutor memory design and is currently a co-author for a paper to be presented at the Nonvolatile Memory Technology Review. After completing his M.S. degree in June, 1993, he plans to extend his graduate studies to the PhD level.

END

OF

TITLE