VIRTUAL-GROUND SENSING TECHNIQUES FOR FAST, LOW-POWER, 1.8V TWO-BIT-PER-CELL FLASH MEMORIES

A DISSERATATION

SUBMITTED TO THE DEPARTMENT OF

ELECTRICAL ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Binh Quang Le

November 2003

©Copyright by Binh Quang Le 2004
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Robert W. Dutton, Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Joseph M. Kahn

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Bruce A. Wooley

Approved for the University Committee on Graduate Studies:

Abstract

Fast and accurate read operation in 1.8V, two-bit-per-cell virtual ground flash memories requires techniques to substantially reduce the read margin loss due to the adjacent cell leakage current, the complementary-bit disturbance and also due to the cycle-induced mobility degradation. The read margin loss caused by the combined effect of these three disturbance factors is serious enough to eliminate the read margin window, which is already small when the power supply voltage is about 1.8V and when each memory cell stores 2 bits. This work introduces for the first time the sense current recovery technique to counteract the adjacent cell leakage current effect, the differential feedback cascoded control of bitline voltage to minimize the complementary-bit disturbance, and the auto-calibrated control of the wordline voltage in the read mode to reduce the mobility degradation effect as well as to ease the design of the sensing circuitry. A 1.8V, 256Mb, two-bit-per-cell virtual-ground flash memory employing all three techniques has been integrated using 0.13μm nitride-storage technology. These three sensing techniques are essential for the memory in order to achieve 30.4ns initial read access and 200MHz internal burst sensing speed. The die size for the prototype test chip is $52mm^2$ and the cell size is $0.121\mu m^2$.

Acknowledgments

I would like to acknowledge many people who have directly or indirectly contributed to the successful development of the work in this thesis. First, I would like to thank my adviser, Professor Robert W. Dutton, for his invaluable guidance and support. His questions and suggestions have always inspired me to think about a topic more thoroughly and deeper. He taught me how to write technical papers for journals and conferences effectively with sound arguments and focusing. I have benefited from his vast technical expertise in many areas, not only in circuit design. I am very grateful and respectful to him who always acts in the best interest of his students. This work would not been possible without his support and guidance. I dedicate this dissertation to him, Professor Robert W. Dutton.

I would like to thank my first associate adviser, Professor Krishna Saraswat for his help and support for me to continue with my Ph.D. program. He and Professor Robert W. Dutton will always be my role models and I have strong desire to help others to follow their dreams the same way Professor Saraswat and Professor Dutton have helped me.

I also would like to thank my second associate adviser, Professor Joseph M. Kahn for serving on my oral exam and reading committees. I feel fortunate to have his excellent

teaching in Signals and Systems many years ago at U.C. Berkeley. The knowledge I learned from him has been the foundation for many other circuit design and signal processing classes I took at Stanford University.

I sincerely thank Professor Bruce A. Wooley for serving on both my orals and reading committees. I am grateful to him for what he has taught me in EE313 and EE315, which has been extremely useful in building the sensing path for the memory described in this dissertation.

I would like to thank Professor Mark A. Horowitz for bringing me into the Ph.D. program and for the knowledge he taught me in EE371, which has been critical in optimizing the speed of the memory data path.

I would like to acknowledge many people in the Memory Group at Advanced Micro Devices, which has just become FASL, LLC. First and foremost, I wish to acknowledge Mr. Michael VanBurkirk, Group Vice President of Worldwide Engineering. His novel thinking and optimistic attitude has been an inspiration to me for all of my time at AMD. For this dissertation, he has been very supportive both financially and intellectually from the beginning. I am very grateful to him for this lifetime opportunity to work on this exciting flash memory.

I would like to express my gratitude to Dr. Pau-Ling Chen, Director of Engineering, for his constant support in many years for my M.S. and Ph.D. programs. Working for him has been an exciting and rewarding experience for me. His supervision for the project described in the dissertation has been critically important. Special thanks to him for clearing many obstacles during the development of the project.

I sincerely thank to Mr. Lee Cleveland, Design Manager, for his tremendous contribution to the project. I am deeply grateful to him for asking me to develop the sensing path for this flash memory. It was this chance that led me to almost all the ideas I have developed in the dissertation.

I gratefully acknowledge Mrs. Darlene Hamilton, Product Development Manager, for her strong support for my Ph.D. program and for her wise advice on technical as well as non-technical issues. The advice have helped me greatly in getting over many obstacles in my life.

I am deeply indebted to many other people in the Memory Group who played very important roles in making the project a successful one. Special thanks to Michael Achter, Xin Guo, Chin-Ghee Ch'ng who are the members of the read path design team. I wish to acknowledge the design, layout, technology, test and product engineering teams at FASL LLC. in Sunnyvale (USA), in Penang (Malaysia) and in Japan.

I would like to thank Professor Dutton's staff members, Fely Barrera, Miho Nishi and Dr. Daniel for their generous help and support. I have been very fortunate to be part of Professor Dutton's research group: Reza Navid, Yi-Chang Lu, Jae Wook Kim, Jung-Hoon Chun, Minghui Han, Choshu Ito and others. I would like to thank them all for their friendship and their helpful discussions. They are very bright and intelligent students; the more I work with them, the more I can learn from them. Special thank to Ting-Yen Chiang in Professor Saraswat's group for his long-time friendship, kindness and his useful advice.

I am also grateful to Professor Nathan W. Cheung at U.C. Berkeley and Professor Konrad Stein at Golden West College for supporting me to pursuit graduate education at Stanford University. They have been my favorite professors and have profound influence on my academic path.

I am deeply indebted to my family for their love and unconditional support. Even in the most difficult times for the family, my parents, brothers and sister have been steadfast to keep me advancing forward. I am very grateful to them for the encouragement and support they have given me for all my life.

Finally, I would like to express my gratitude to my wife, Diem Le, who have devoted her entire time for the family. Without her help, her support and sacrifice, I would not have finished this long and difficult Ph.D. program. She also helped reviewing the dissertation, finding out numerous errors and typos. At last, I would like to thank my little sons Binh An Quang Le, Brian Quang Thai Le and Anthony Quang Phu Le. They are my endless source of happiness, thus in a way they have been contributed to the success of the project.

Table of Contents

Abstract . . . . **v**

Acknowledgments . . . . **vii**

Table of Contents . . . . **xi**

List of Tables . . . . **xv**

List of Figures . . . . **xvii**

Chapter 1 Introduction **1**

1.1 Two-bit-per-cell virtual-ground architecture 3

1.2 Design challenges 6

1.2.1 Read margin loss due to the side-leakage current 6

1.2.1.1 “0” read margin loss 6

1.2.1.2 “1” read margin loss 10

1.2.2 Read margin loss due to the Complementary-Bit Disturbance 11

1.2.3 Read margin loss due to the mobility degradation 14

1.2.4 Read speed considerations 17

1.3 Contributions 18

1.4 Dissertation organization 18

Chapter 2 Sense current recovery technique **21**
2.1 Preliminary Solution to the side-leakage problem 22
2.1.1 Fixing the “1” read margin loss -- A Preliminary Solution 22
2.1.2 Minimizing the “0” read margin loss -- A Preliminary Solution 23
2.2 Sense Current Recovery Technique 25
2.2.1 Recover the read margin loss by using multiple drains and multiple protecting bitlines 26
2.2.2 Eliminating the voltage mismatch between the drain bitlines and the protecting bitlines 29
2.2.3 HSPICE simulation results for the Sense Current Recovery Technique 33
2.2.3.1 Simulation result for the worst case “1” read margin loss 33
2.2.3.2 Simulation result for the worst case “0” read margin loss 35
2.3 Summary 37

Chapter 3 Multiple-drain-bitline and multiple-protecting-bitline column decoding **39**
3.1 Chip architecture 40
3.2 Column decoding architecture 40
3.2.1 Local bitline decoding block 43
3.2.2 Sector-y logic block 46
3.2.3 Global bitline decoding block 46
3.2.4 Source-Drain-Protecting logic block 50
3.2.5 Simulation results and sizing for the column decoding 52
3.2.6 Bitline decoding for the edges 57
3.3 Summary 59

Chapter 4 Differential feedback cascoded bitline voltage control **61**
4.1 Analysis of the traditional cascode amplifier 62
4.2 Differential feedback cascoded bitline voltage control 64
4.2.1 Differential feedback cascode amplifier - A simplified version 65

4.2.1.1 The differential amplifier . . . 66
4.2.1.2 The reference voltage generator . . . 67
4.2.1.3 Power consumption of the new cascode amplifier . . . 68
4.2.2 Differential feedback cascode amplifier - The complete version . . . 69
4.2.2.1 Eliminating disturbance to the sensing node . . . 70
4.2.2.1 Eliminating the bitline voltage overshoot . . . 71
4.3 HSPICE simulation result for the differential feedback cascode amplifier . . . 73
4.4 Summary . . . 75

Chapter 5 Auto-calibrated wordline voltage control 77
5.1 A/D Converter block . . . 78
5.1.1 Designing the resistor chain . . . 80
5.1.2 Designing the fast reference voltage circuit . . . 81
5.1.3 Designing the comparator . . . 86
5.1.4 Simulation result for the A/D converter block . . . 87
5.1.5 Alternative FVREF design - Speed/Accuracy trade-off . . . 88
5.2 Vboost block . . . 92
5.3 Summary . . . 98

Chapter 6 Read path simulation and measured results 99
6.1 Read path simulation results . . . 100
6.2 Read path measured results . . . 104
6.3 Summary . . . 112

Chapter 7 Conclusion 113
7.1 Contributions . . . 114
7.2 Recommendations for future work . . . 115

Appendix 117
A.1 Verilog Code for Local bitline decoding . . . 117
A.2 Verilog Code for Global source bitline decoding . . . 119

A.3 Verilog Code for Global drain bitline decoding 120
A.4 Verilog Code for Global protecting bitline decoding 123

Bibliography **125**

List of Tables

Table 3.1 Truth table for the sector-y logic block 47

Table 3.2 Truth table for the source-drain-protecting logic block - Source decoding 52

Table 3.3 Truth table for the source-drain-protecting logic block - Drain decoding 53

Table 3.4 Truth table for the source-drain-protecting logic block - Protecting decoding 54

Table 5.1 Target thermometer codes for the A/D converter block at different supply voltages 78

Table 5.2 Voltage levels of VDIV nodes at different VCCs 81

Table 5.3 Simulation result for the A/D converter block 87

Table 5.4 A/D converter block simulation result, using the simple fast reference circuit 91

Table 6.1 Measured performance 111

List of Figures

Figure 1.1 Flash memory usage for cellular phones 2
Figure 1.2 a) Multilevel memory cell, and b) nitride-storage memory cell 2
Figure 1.3 Virtual-ground array architecture 4
Figure 1.4 Read margin window and read margins 5
Figure 1.5 Total side-leakage current 7
Figure 1.6 Simulation result for the side-leakage current 8
Figure 1.7 A previous design for reducing the side-leakage current 9
Figure 1.8 “1” read margin loss 10
Figure 1.9 Threshold distributions of Bit A 12
Figure 1.10 Traditional cascode amplifier 13
Figure 1.11 Read margin loss due to the cycling induced mobility degradation 15
Figure 1.12 Simple wordline booster 15
Figure 1.13 Improved simple wordline booster 16
Figure 1.14 Feedback regulation booster 17

Figure 2.1 Preliminary Solution for fixing “1” read margin loss 22
Figure 2.2 “0” read margin loss 23

Figure 2.3 Preliminary Solution to minimize the “0” read margin loss 24
Figure 2.4 Sense Current Recovery Technique - Recover “1” read margin loss 26
Figure 2.5 Sense Current Recovery Technique - Recover “0” read margin loss 28
Figure 2.6 Sense Current Recovery Technique - Solving voltage mismatch problem 29
Figure 2.7 Protecting feedback amplifier 30
Figure 2.8 Schematic for testing the protecting feedback amplifier performance . . . 31
Figure 2.9 HSPICE simulation result for the protecting feedback amplifier 32
Figure 2.10 Worst case “1” read margin loss 33
Figure 2.11 Threshold distribution of the erased and programmed bits 34
Figure 2.12 HPICE simulation for the worst case “1” read margin loss 35
Figure 2.13 Worst case “0” read margin loss 36
Figure 2.14 HPICE simulation for the worst case “0” read margin loss 37

Figure 3.1 Chip architecture 40
Figure 3.2 Bank column decoding architecture 41
Figure 3.3 Data block arrangement 42
Figure 3.4 Physical address for 32 bits on a wordline for a data block 43
Figure 3.5 Sector select level shifter 44
Figure 3.6 Local bitline decoding for one data block 45
Figure 3.7 Basic global bitline decoding for one data block 48
Figure 3.8 Global bitline decoding for GBLn_(0) and GBLn_(7) 49
Figure 3.9 Global bitline decoding for data block interface 51
Figure 3.10 Logic simulation for the sector-y logic block 55
Figure 3.11 Logic simulation for the source-drain-protecting logic block 56
Figure 3.12 Column decoding for the right edge 57
Figure 3.13 Column decoding for the left edge 58

Figure 4.1 Traditional cascode amplifier 62
Figure 4.2 Differential feedback cascode amplifier - A simplified version 65
Figure 4.3 The differential amplifier a) with current mirror load b) with resistive

load . . . 66
Figure 4.4 Bandgap reference voltage generator . . . 67
Figure 4.5 Resistive reference voltage generator . . . 68
Figure 4.6 Current reference for the new cascode amplifier . . . 69
Figure 4.7 The full version of the differential feedback cascode amplifier . . . 70
Figure 4.8 Bitline voltage overshoot . . . 72
Figure 4.9 Simulation result for the traditional cascode amplifier . . . 73
Figure 4.10 Simulation result for the differential feedback cascode amplifier . . . 74

Figure 5.1 A/D Booster . . . 78
Figure 5.2 A/D converter block . . . 79
Figure 5.3 Typical booster timing in an initial read access . . . 79
Figure 5.4 Traditional bandgap reference circuit . . . 82
Figure 5.5 Fast bandgap reference circuit . . . 84
Figure 5.6 HSPICE simulation result for fast bandgap reference circuit . . . 85
Figure 5.7 Comparator . . . 86
Figure 5.8 Simple fast reference circuit . . . 89
Figure 5.9 HSPICE simulation result for the simple fast reference circuit . . . 90
Figure 5.10 A simple view of the Vboost block . . . 92
Figure 5.11 Actual Vboost block schematic . . . 93
Figure 5.12 Booster cell . . . 94
Figure 5.13 High voltage switch . . . 95
Figure 5.14 Simulation result for the wordline booster without using the A/D converter . . . 96
Figure 5.15 Simulation result for the A/D wordline booster - 0-bit error . . . 97
Figure 5.16 Simulation result for the A/D wordline booster - 1-bit and 2-bit errors . . 98

Figure 6.1 Read path simulation set-up diagram . . . 100
Figure 6.2 Read path simulation result . . . 102
Figure 6.3 Simulation waveforms for the sensing node SAIN and the reference node . . . 104

Figure 6.4 Microphotograph of the 1.8V, two-bit-per-cell, 256Mb flash memory. . 105

Figure 6.5 Measurement set-up 106

Figure 6.6 Simulation result for the output buffer 107

Figure 6.7 Initial read access - delay from node START to node END 108

Figure 6.8 Burst read access - delay from clock to node DSI 108

Figure 6.9 Measured performance of the differential feedback cascode amplifier and the protecting feedback amplifier 109

Figure 6.10 Read voltage margin 110

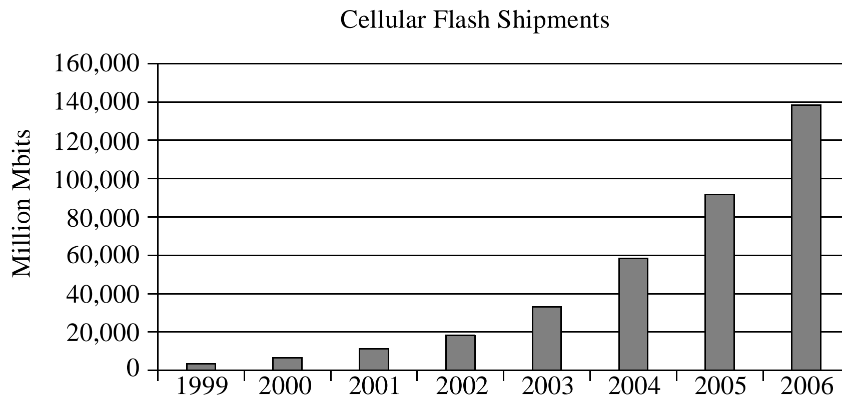
Figure 6.11 Wordline booster output level at different supply voltages 111

Chapter 1

Introduction

Since its first volume shipment in 1988, flash memory year-to-year growth rate has been remarkable. Unlike DRAM, flash memory can retain data without the aid of a power supply. This nonvolatility as well as other important characteristics such as in-system updateability, fast speed and low power consumption help flash memory find its way into numerous applications. Flash memory is used in personal computers, printers, GPS systems, cars, digital cameras, digital camcorders, modems, routers, cellular phones, etc. Figure 1.1 shows the explosive usage of flash memory for cellular phones in the coming years. The market for flash memory in general is growing rapidly, especially for low-cost and high- density applications. Thus, two-bit-per-cell flash memory becomes very important because it offers both low cost and high density.

There are two main technologies for the two-bit-per-cell flash memory. The first one is the Multilevel-cell technology, which has been reported in many papers such as [1], [2] and [3]. The second one is the nitride-storage technology, which is used to develop the flash memory described in this dissertation.



- Cell Phone Flash memory content increasing from 29Mbits to 100Mbits in 2004

Figure 1.1: Flash memory usage for cellular phones
(Data from Web-Feet Research, 2002 Non-volatile Memory Conference)

In the Multilevel-cell technology, memory cell data is stored on a floating gate as shown in Figure 1.2a. Two bits can be stored in a cell by programming, which is the process of putting electrons into the floating gate to raise its threshold voltage, to achieve one of four possible threshold voltages.

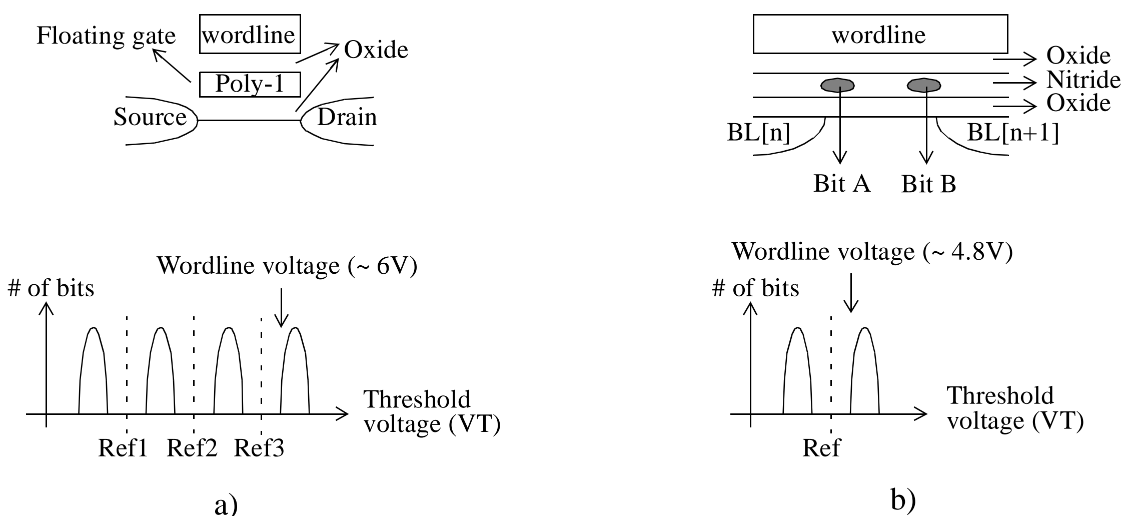


Figure 1.2: a) Multilevel memory cell, and b) nitride-storage memory cell

Because there are 4 threshold voltage values, the last threshold range is high, and therefore the wordline voltage in a read operation must be high accordingly. This leads to slower read access and increased power consumption.

In the nitride-storage technology, memory cell data is stored in the Nitride layer at 2 different locations as shown in Figure 1.2b. The bit on the left of the cell is called Bit A, and the bit on the right is called Bit B. Each bit has one of only 2 possible threshold voltage ranges. Note that two bits can be stored in the same Nitride layer at 2 different locations because the Nitride is a non-conductive material and charge will stay where it has been stored. Because there are only 2 threshold voltage ranges, the wordline voltage in a read operation is lower, about 4.8V. This leads to faster read access and less power consumption. Due to the symmetrical arrangement of the bits in a nitride-storage memory cell, the array must have a virtual-ground architecture, which is the topic of the next section.

1.1 Two-bit-per-cell virtual-ground architecture

The virtual-ground array architecture is shown in Figure 1.3. The array is divided horizontally into 64 data blocks, where each data block is 16 cells wide. This division is implemented so that in each read operation, 64 bits can be read out at the same time, one bit from each data block. In Figure 1.3, to read Bit A of Cell1_1 (in Data block 1), the wordline WL(1) is boosted to about 4.8V; a drain voltage (~1.2V) and a ground voltage (0V) are applied to the bitlines BL1_(1) and BL1_(0), respectively. To read Bit B of this same cell, all the bias voltages are the same, except that BL1_(0) now becomes the drain and BL1_(1) becomes the source. The term “virtual ground” comes from the fact that there is no fixed ground line in the array; any bitline can be decoded as the ground bitline, depending on what cell and what side of the cell is read. Note that when Bit A of Cell1_1 of Data block 1 is read, Bit A of Cell2_1 of Data block 2 is also read at the same time, thus there is a potential leakage from the drain bitline BL1_(1) to the source bitline BL2_(0) of

Data block 2, if all the cells between these 2 bitlines are erased cells (low threshold voltage cells). This potential leakage and its consequences will be discussed extensively in Chapters 1 and 2.

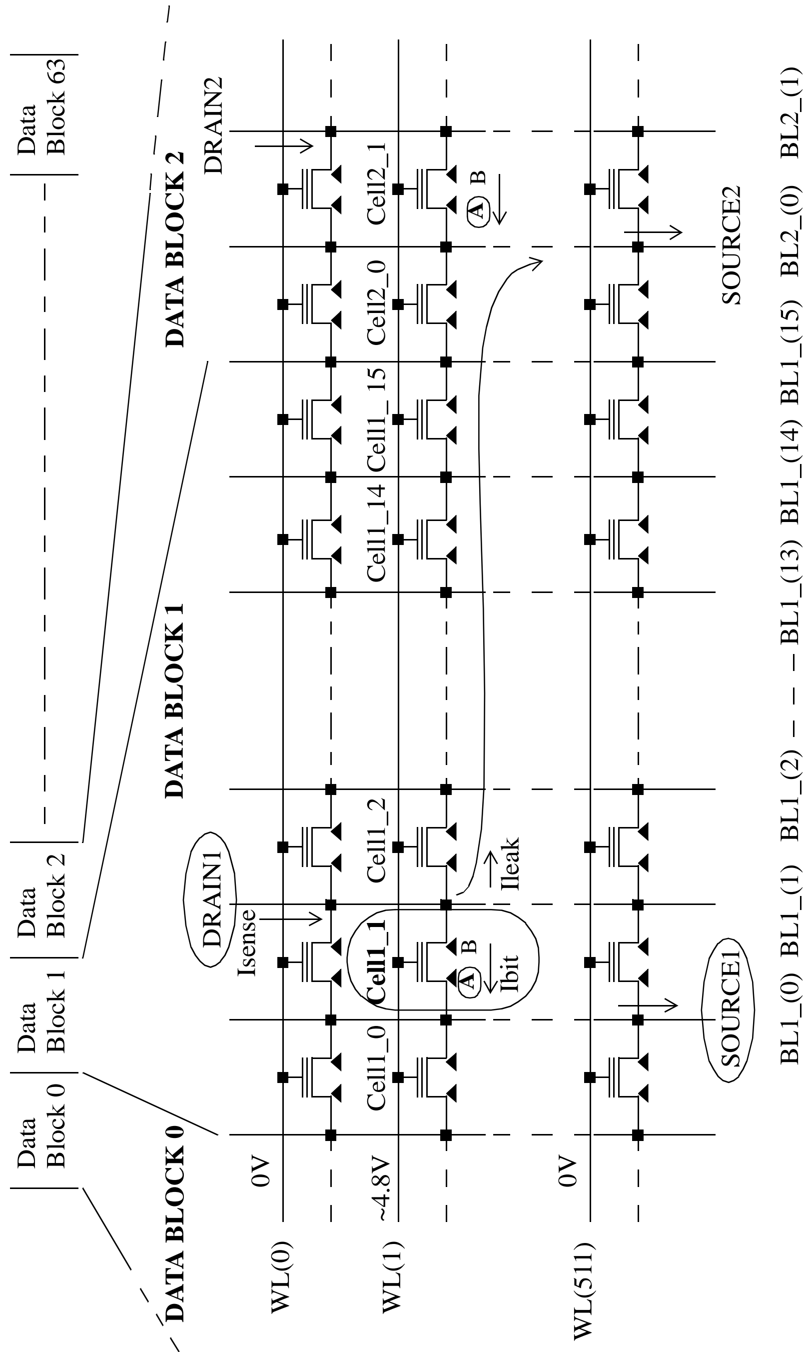


Figure 1.3: Virtual-ground array architecture

As shown in Figure 1.2b, each bit of a nitride-storage memory cell has one of only two possible threshold voltage ranges; the bit is either an erased bit (also called a “1” bit) or a programmed bit (also called a “0” bit). To change a bit from “1” to “0”, a programming operation is performed, in which electrons are injected locally into the nitride layer to raise the threshold voltage of the bit. An erase operation is the inverse of a programming operation, in which a bit is changed from “0” to “1”. In a read access, the current of the bit being read is compared with a reference current. If the bit current is more than the reference current, that bit is an erased bit, which has a low threshold voltage VT. Vice versa, if the bit current is less than the reference current, that bit is a programmed bit, which has a high threshold voltage VT. The absolute value of the difference between the reference current and the bit current is called the read margin. Figure 1.4 shows the read margins for a “0” bit and a “1” bit. The difference between the programmed bit current and the erased bit current is called the read margin window.

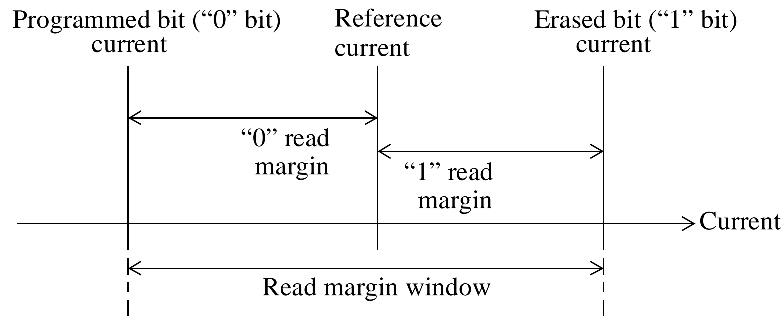


Figure 1.4: Read margin window and read margins

In Figure 1.4, suppose that the reference current, the erased bit current and the programmed bit current are 26μA, 30μA and 22μA, respectively, then the read margin window is 8μA, the “0” read margin is 4μA and the “1” read margin is also 4μA. Apparently, the larger the read margin is, the better. Thus for a “0” bit, smaller current is preferred because this leads to larger “0” read margin, while for a “1” bit, larger current is preferred because it makes the “1” read margin larger.

1.2 Design challenges

Designing a 1.8V, two-bit-per-cell nitride-storage flash memory is very challenging due to the substantial read margin loss caused by the adjacent cell leakage current, the disturbance from the other bit in the same memory cell and the cycle-induced mobility degradation. These three read margin loss mechanisms can wipe out the entire read margin window, and if there are no effective sensing techniques to suppress or eliminate them, the design of the memory is impossible. The next three sections will describe these read margin loss mechanisms in detail.

1.2.1 Read margin loss due to the side-leakage current

The adjacent cell leakage current, which is well-known in virtual-ground flash memories, causes the biggest read margin loss. It is inherent in the virtual-ground array architecture by itself, in which the current of the bit being read is affected by the status of the neighboring cells. The problem is difficult to solve because this so-called "side-leakage" can not be predicted by the sensing circuitry which is residing outside of the memory array; the side-leakage current at any reading address can exist at one time but disappear at the other time when the cells neighboring the cell being read are programmed. The side-leakage current, depending on its direction, can cause the read margin loss for both "0" bits and "1" bits.

1.2.1.1 "0" read margin loss

This section explains why the side-leakage current significantly reduces or even eliminates the read margin for a "0" bit. In Figure 1.3, suppose that the bit being read is

Bit A of Cell1_1, and also suppose that this bit is a programmed bit ("0" bit), which should have smaller read current than the reference current. This bit uses the drain bitline BL1_(1) in the Data block 1 to send its current to the sensing circuitry. Ideally, the current sent to the sensing circuitry Isense should be equal to the current of the bit being read, but due to the side-leakage current, Isense is not equal to the bit current anymore. As shown in Figure 1.3, there is a potential side-leakage current from the drain bitline BL1_(1) of the Data block 1 to the source bitline BL2_(0) of the Data block 2. In the worst case where all the cells between these bitlines are over-erased cells, which have low threshold voltages, the side-leakage can be very large. The steady-state side-leakage current in this case can be estimated easily. An over-erased cell has a resistance of about 16 Kohms, and because there are 15 cells between these bitlines, the steady-state side-leakage current (DC component) is about $\frac{(1.2V - 0V)}{15 \times 16Kohms} = 5\mu A$, where the drain voltage is assumed to be 1.2V. Figure 1.3 reveals that Isense = Ibit + Ileak, thus the read margin loss for this "0" bit is Ileak, which is 5µA. If the reference current is 26µA and Bit A current is 22µA, then Isense is 22µA + 5µA = 27µA, which is larger than the reference current. This is very serious because the side-leakage current has wiped out the entire "0" read margin, and the reading result for Bit A, instead of "0", is "1", which is totally wrong. Even more serious, the side-leakage has a large transient current (ac) component, which is much larger than the steady-state current of 5µA. Figure 1.5 shows the total leakage current Ileak, which includes both dc and ac components.

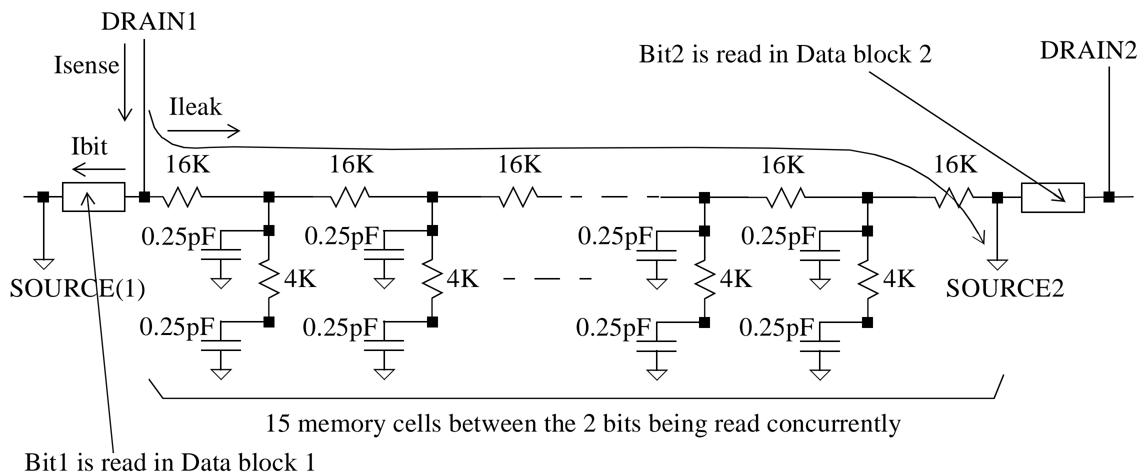


Figure 1.5: Total side-leakage current

The large magnitude of the transient component is caused by the capacitance of the floating bitlines between the DRAIN1 bitline and the SOURCE2 bitline. Each bitline has a capacitance to ground of about 0.5pF, a resistance of about 4Kohms, and is modeled as an RC Π-network, with capacitors of 0.25pF and a resistor of 4Kohms as shown in Figure 1.5. The HSPICE simulation results for Figure 1.5 are shown in Figure 1.6. To find the magnitude of the side-leakage current Ileak only, Ibit1 in Figure 1.5 is set to zero; in the simulation, the bitline DRAIN1 is ramped from 0V to 1.2V in 10ns.

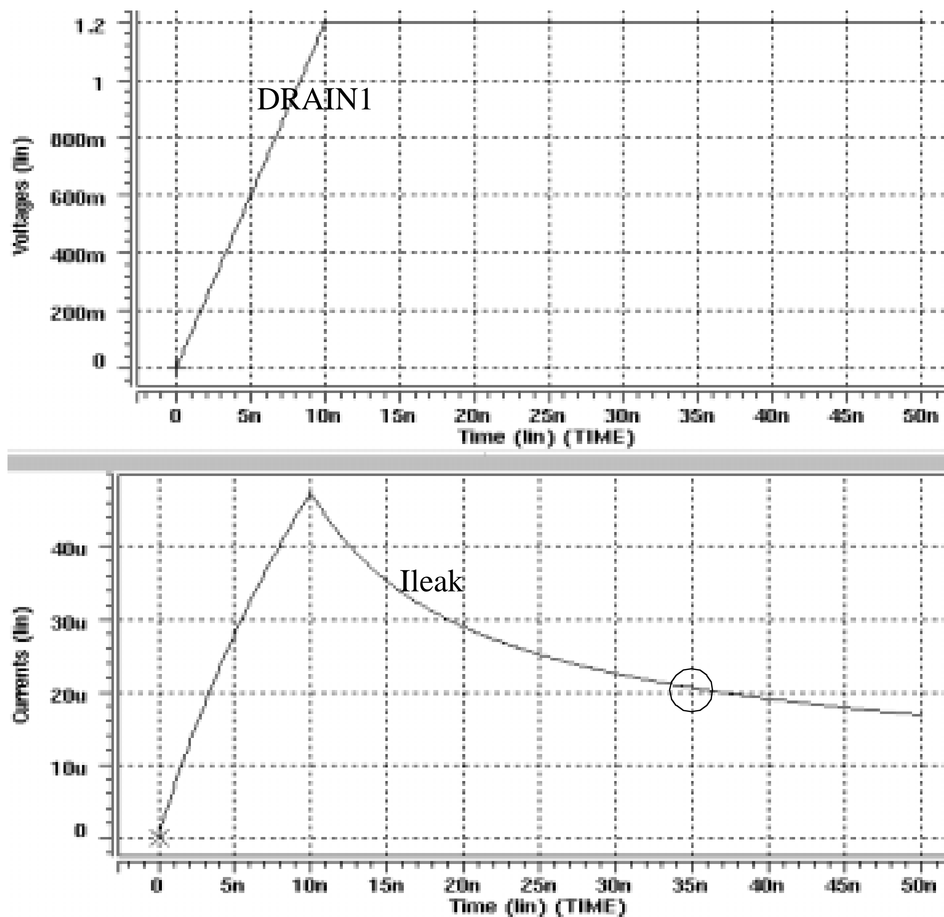


Figure 1.6: Simulation result for the side-leakage transient current

Figure 1.6 reveals that the total side-leakage current is very large, about 20μA at 35ns after the bitline DRAIN1 is activated. The sensing operation will fail with such a large side-leakage current magnitude.

Many previous designs have tried to reduce this side-leakage current such as in [4] and [5]. In [4], the virtual-ground array is physically divided horizontally into separate 8-cell-wide slices. This approach eliminates the dc component of the side-leakage but the ac component is still large, and thus the approach is not suitable for flash memory with fast read access. In addition, the physical isolation between the slices increases the memory array size.

Another approach used to reduce the side-leakage current is discussed in [5], and is shown in Figure 1.7. A protecting amplifier, which is similar to the cascode amplifier that is used to provide the drain voltage for the cell being read, is used to drive the bitline next to the drain bitline (denoted as "D") to a similar voltage to that of the drain; this bitline is called the protecting bitline and is denoted as "P" in Figure 1.7.

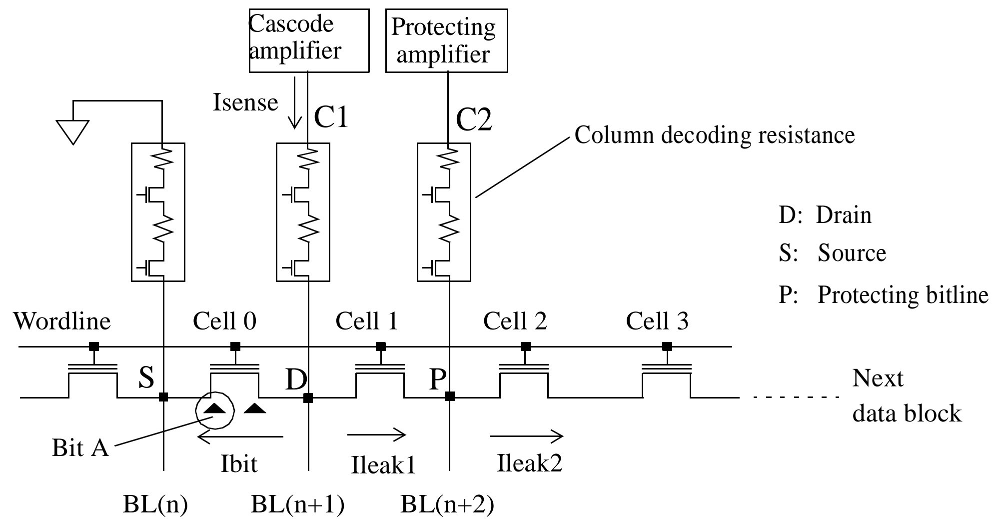


Figure 1.7: A previous design for reducing the side-leakage current

The approach shown in Figure 1.7 indeed reduces the leakage Ileak1 from the drain bitline BL(n+1) to the right (called the forward leakage), but simulation result reveals that Ileak1 is still large (about 3.4µA) when Ileak2 is 20µA -- a value of 20µA is obtained from the simulation shown in Figure 1.6. Because the read margin for a "0" bit is about 4µA,

Ileak1 of 3.4μA is still not acceptable, although it is much smaller than the 20μA in the case when the protecting bit line is not used. Another drawback of the approach shown in Figure 1.7 is that, in many other cases, it causes the read margin loss for “1” bits (or erased bits). This effect is discussed in the next section.

1.2.1.2 “1” read margin loss

Figure 1.8 shows the case when the approach discussed in [5] causes the read margin loss for a “1” bit. Suppose the bit being read is Bit A of Cell 0 and assume that Bit A is an erased bit (“1” bit). Also assume that Cell 1 is an erased cell and Cell 2 is a programmed cell, thus there is a potential side-leakage through Cell 1 but not Cell 2.

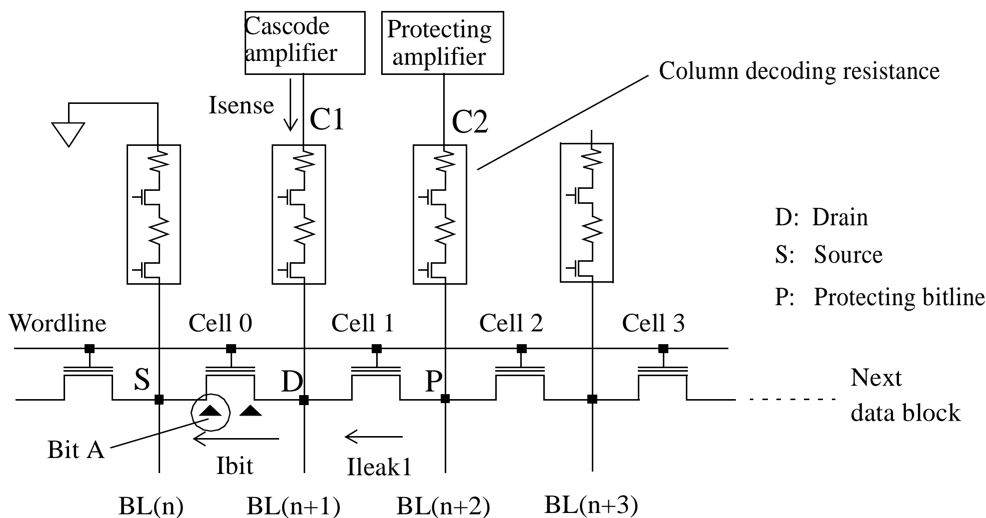


Figure 1.8: “1” read margin loss

Because Bit A of Cell 0 is erased, its current Ibit is large, thus node D is pulled down more than node P, creating a voltage difference between nodes D and P, thus a leakage current Ileak1 is created, flowing from the protecting bitline P to the drain bitline D (called the reverse leakage). Figure 1.8 shows that the current Isense is now equal to (Ibit - Ileak1),

not Ibit, therefore the read margin loss for this “1” bit (Bit A of Cell 0) is Ileak1. In the worst case, when Cell 1 is an over-erased cell, which has very low threshold voltage, Ileak1 can be about 5μA. This reverse leakage apparently wipes out the entire read margin for a “1” bit, assuming that the original read margin for a “1” bit is 4μA.

To reduce this reverse leakage current [6], in Figure 1.8, BL(n+3) can be driven by the protecting amplifier instead of BL(n+2), which is now left floating; any reverse leakage current from BL(n+3) to BL(n+1) must pass through two cells, Cell 1 and Cell 2, therefore its magnitude is smaller. This approach, however, causes slow read access because while the dc component of leakage is reduced, the ac component, which charges the high capacitive float bitline BL(n+2), is increased.

In short, the side-leakage current Ileak1 causes a significant read margin loss for “0” bits if it flows from the drain bitline to the protecting bitline and causes substantial read margin loss for “1” bits if it flows from the protecting bitline to the drain bitline. The forward or reverse leakage Ileak1 can be even larger due to the mismatch in voltage between nodes C1 and C2; even though the nodes are driven by almost identical amplifiers, the mismatch can still occur because the currents drawn from the amplifiers are not equal.

1.2.2 Read Margin loss due to the Complementary-Bit Disturbance

The second mechanism that also causes substantial loss of read margin is from the disturbance of the other bit stored in the same memory cell. Due to the interaction between the two bits in a memory cell, a bit can belong to 4 possible threshold voltage distributions 11, 10, 01, 00 as depicted in Figure 1.9. Distributions 11 and 10 are for erased bits and distributions 01 and 00 are for the programmed bits. Suppose that Bit A is an erased bit and Bit B is also an erased bit, then Bit A belongs to the distribution 11,

which is the distribution for the bits with largest current (or lowest threshold voltage); in this case Bit A has a large “1” read margin. If Bit B becomes programmed then the current produced when reading Bit A is less, and the distribution for Bit A is now 10; the “1” read margin for Bit A has been reduced because of the disturbance from Bit B. Conversely, if Bit A is a programmed bit and Bit B is also a programmed bit, then Bit A belongs to the distribution 00, which is the distribution for the bits with smallest current (or highest threshold voltage); in this case, Bit A has a large “0” read margin. When Bit B becomes programmed, the current produced when reading Bit A is more, and Bit A now belongs to the distribution 01. The “0” read margin for Bit A has been reduced due to disturbance from Bit B.

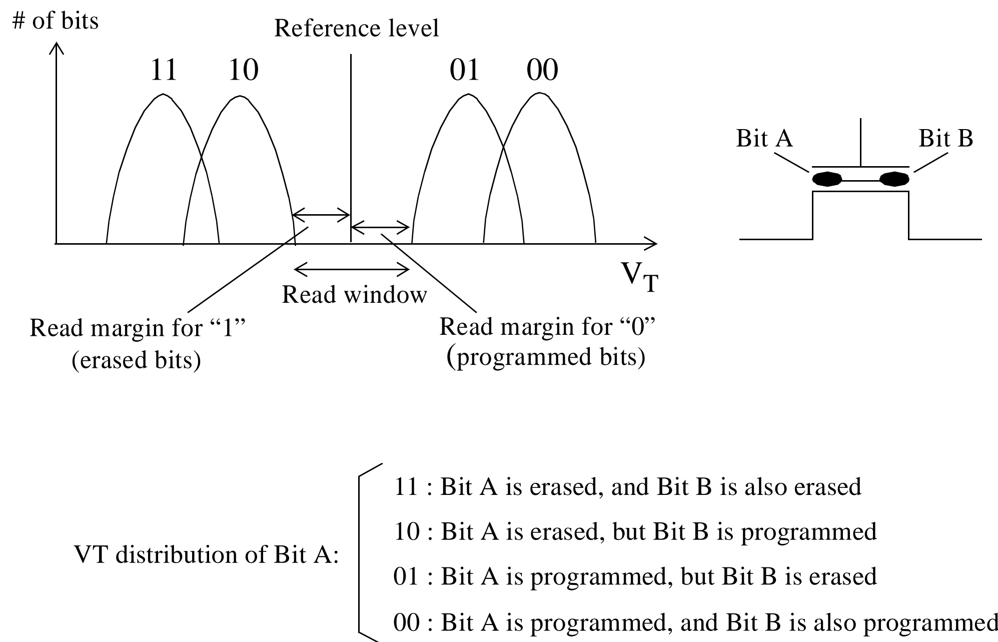


Figure 1.9: Threshold voltage distributions of Bit A

The disturbance caused by the other bit in the same cell is usually called the complementary-bit disturbance (CBD effect). This disturbance is serious because it substantially reduces the read margin for both “1” and “0” bits, especially at low supply voltage of 1.8V when the read margin is already small.

The complementary-bit disturbance is less if the cell is operated well in the saturation region. In saturation, as described in numerous text books such as [7], there is a pinch-off region at the drain, and thus the effect of the negative charge at Bit B to the drain current is less because there is no channel at the drain anyway. Of course, the charge at Bit B is not concentrated at one point, but spreads out in a certain distance, thus the larger the pinch-off region, the better Bit B is shielded, leading to smaller CBD effect and better read margin. In short, to minimize the read margin loss caused by the Complementary-Bit disturbance, the drain bitline voltage *needs to be raised as high as possible and its variation with process, temperature and cell current should be as small as possible*. Because the channel length of the memory cell is very short, carrier velocity saturation effect actually occurs, and therefore the drain voltage at which the memory cell starts to enter the saturation region Vdsat is smaller [8]. Vdsat is about 1.15V and thus to minimize the CBD effect, the cascode amplifier should raise the drain bitline voltage to at least 1.15V. A traditional cascode amplifier such as the one in [1] or [9] is shown in Figure 1.10.

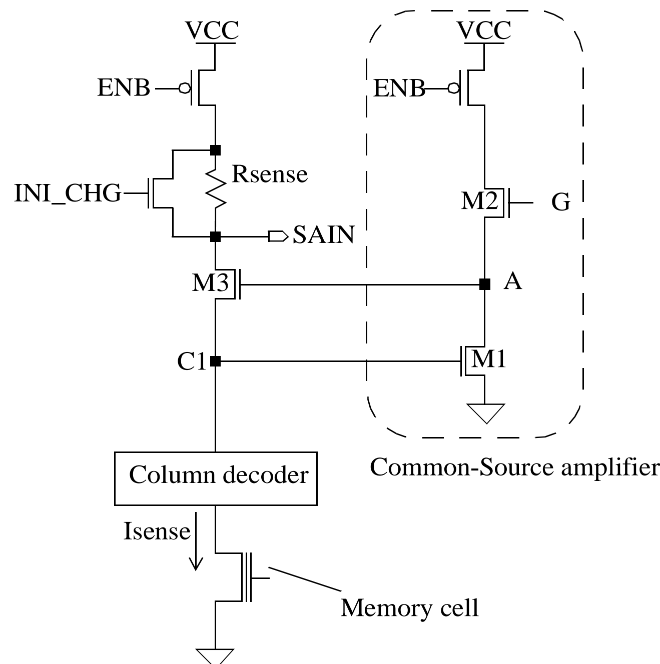


Figure 1.10: Traditional cascode amplifier

The cascode amplier is usually used to convert the sensing current Isense to a voltage (SAIN) for sensing. It is also used to provide a stable voltage V_{C1} at node C1 for the drain bitline during a read operation. The traditional cascode amplifier shown in Figure 1.10 is fast and simple, but it is not very good for the nitride-storage flash memory. First of all, for this type of cascode amplifier, it is very difficult to raise the bitline voltage V_{C1} to a high level (about 1.15V). Note that C1 is the input of the Common-Source amplifier, and if the common mode input level V_{C1} increases, the common mode output level (node A) will decrease, causing V_{C1} to drop back somewhat because node A controls the gate of transistor M3. Secondly, V_{C1} of the traditional cascode amplifier varies greatly with process because V_{C1} depends on many process parameters such as the threshold voltages of transistors M1, M2 and M3, which are not well-controlled. In summary, a new cascode amplifier that can overcome the drawbacks of the traditional cascode amplifier, is needed to make the 1.8V nitride-storage two-bit-per-cell flash memory work.

1.2.3 Read margin loss due to the mobility degradation

The third mechanism that causes the read margin loss is the cycling induced mobility degradation, which is depicted in Figure 1.11. In the figure, the Id-Vg curves for an erased bit before and after cycling as well as for the reference bit are shown. The read margin, which is the difference between the bit current and the reference current, is reduced after cycling due to the change in slope of the Id-Vg curve; this slope change is an indication of the cycling induced mobility degradation. Even worse, if the wordline voltage varies too much, for example to the point V_2, the read data will be incorrect because the bit current at this point becomes smaller than the reference current. Therefore, to minimize the read margin loss caused by the cycling induced mobility degradation, the wordline voltage *needs to be controlled accurately.*

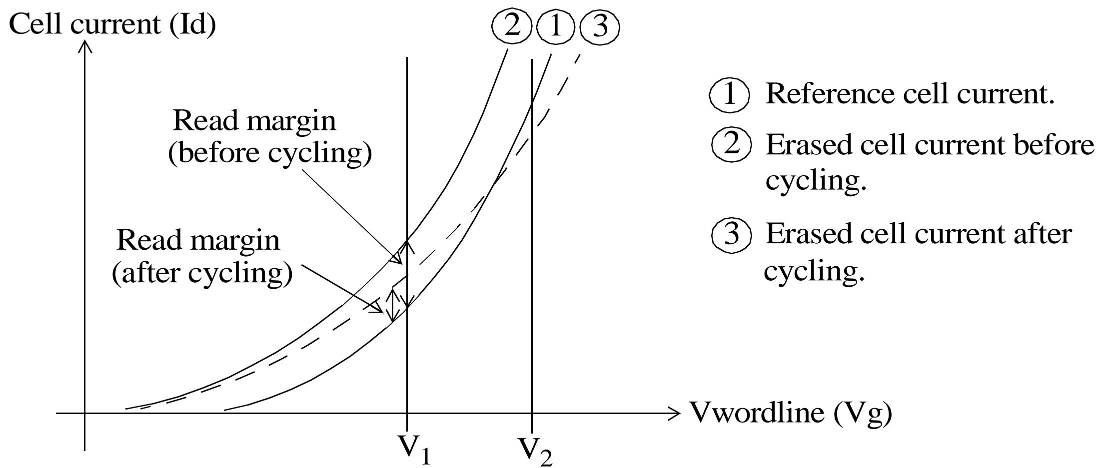


Figure 1.11: Read margin loss due to the cycling induced mobility degradation

The circuit used to generate the wordline voltage in read mode is called the wordline booster. Because the boosted voltage is higher than the supply voltage VCC, the wordline booster usually uses capacitors to multiply the power supply voltage to the target voltage, thus the main variation in the boosted level is caused by the variation in the supply voltage VCC. Figure 1.12 presents a simple wordline booster.

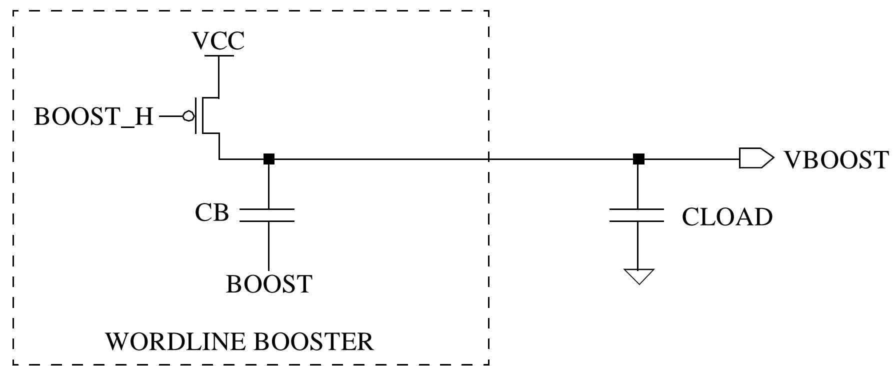


Figure 1.12: Simple wordline booster

The boosted level VBOOST can be calculated easily by applying the charge conservation principle: $VBOOST = \left(1 + \frac{CB}{CB + CLOAD}\right)VCC$. Assume that CB = CLOAD, then

the variation in VBOOST will be 1.5 times the variation in VCC. For VCC ~ 1.6V, to generate a high VBOOST (~4.8V), CB has to be a combination of 3 capacitors; each capacitor is precharged to VCC first then all 3 capacitors are connected in series to generate the high voltage VBOOST. The variation in VCC (from 1.6V to 2.0V), therefore, is amplified more, and the variation of VBOOST can be as large as 1.2V. An attempt to improve the accuracy of the booster is as follows. Instead of precharging the boosting capacitors to VCC, the capacitors are precharged from an accurate reference voltage VREF, as shown in Figure 1.13.

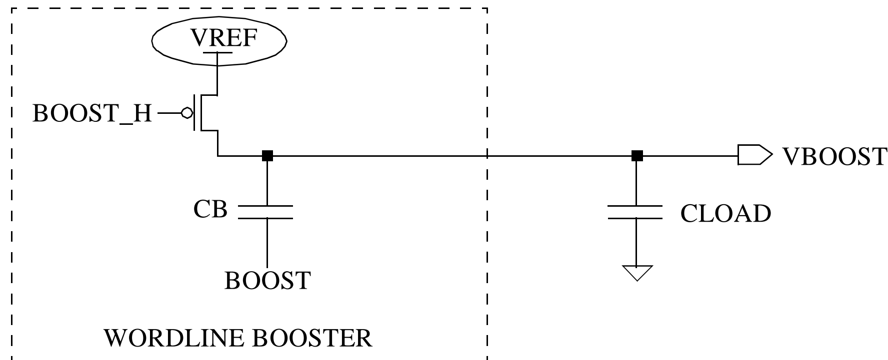


Figure 1.13: Improved simple wordline booster

The problem with this design is that VREF is a weak signal, thus it needs to be buffered before being used to charge the capacitors. The buffer, which can be just a unity feedback amplifier, is difficult to design because it has to drive a very big capacitive load (20pF - 30pF) to the VREF level in about only 10ns.

The feedback regulation technique discussed in [10] can be used to design an accurate wordline booster shown in Figure 1.14. However, the speed of this design is slow due to the use of the feedback amplifier to drive the big discharge transistor M1.

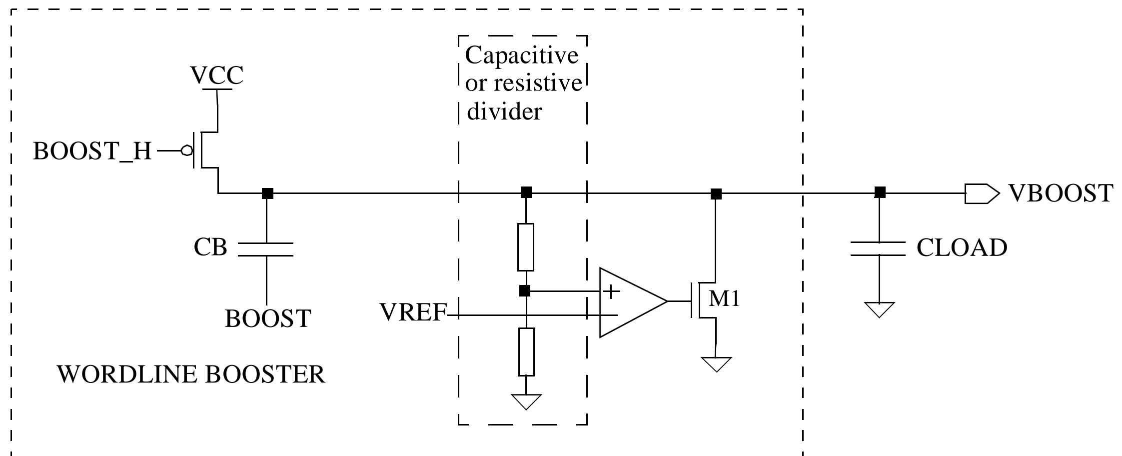


Figure 1.14: Feedback regulation booster

Thus, a new accurate wordline booster that can quickly generate the wordline voltage is needed. The accuracy of the wordline level helps to reduce the read margin loss due to the cycle-induced mobility degradation, and the fast speed of the wordline booster is required to achieve fast read operation.

1.2.4 Read speed considerations

The challenges in designing the 1.8V, two-bit-per-cell, 256Mbits nitride-storage flash memory are not only in developing new sensing techniques to solve the margin loss problem, but also in finding new fast and low power sensing techniques. The design must satisfy the initial read access time of 45ns and a burst read access time of 133MHz. The initial access time is measured from the time the new address is applied until the time data is valid at the output pins (I/O pads). During this initial access, the wordline is boosted to a high voltage, thus the initial access time is longer. Burst read accesses are faster than the initial access because these accesses are performed when the selected wordline has been

boosted to a high voltage, thus no additional time is spent for the wordline decoding path and the wordline boosting operation; in burst read, only the bitline decoding path is switching.

1.3 Contributions

The most important contributions of the research described in this dissertation are the development of new sensing techniques that substantially reduce the read margin loss caused by the side-leakage current, the Complementary-Bit Disturbance (CBD) and the cycle-induced mobility degradation, making the first design of the 1.8V, two-bit-per-cell, 256Mbits nitride-storage flash memory possible. The new sensing techniques are the sense current recovery and decoding techniques, the differential feedback cascoded bitline voltage control and the auto-calibrated wordline voltage control. These new sensing techniques not only provide enough read margin to this low voltage (1.8V), two-bit-per-cell flash memory, but also help the memory to achieve a fast initial read access of 30.4ns and an internal burst sensing speed of 200MHz at low power consumption.

1.4 Dissertation organization

This dissertation is organized into seven chapters, including this introduction chapter.

Chapter 2 introduces the Sense Current Recovery Technique, which is the most important new sensing technique developed in this research. The Sense Current Recovery Technique is used to greatly suppress the read margin loss caused by the side-leakage current.

Chapter 3 describes in detail the design of the column decoding “S-S-D-D-D-P-P-P”, which has been developed to support the Sense Current Recovery Technique discussed in Chapter 2. “S-S-D-D-D-P-P-P” means that in a read operation for a memory bit, two source bitlines, three drain bitlines and three protecting bitlines are provided for that bit.

Chapter 4 focuses on the Differential Feedback Cascoded Bitline Voltage Control technique, which is a new sensing technique used to minimize the read margin loss due to the disturbance from the other bit in the same memory cell (CBD effect). The key part of this technique is a new differential feedback cascode amplifier, which has many advantages over the traditional cascode amplifier when used in the 1.8V, two-bit-per-cell nitride-storage flash memory.

Chapter 5 describes the Auto-Calibrated Wordline Voltage Control technique, in which a new A/D wordline booster plays a critical role. The technique is used to reduce the read margin loss caused by the cycle-induced mobility degradation as well as to ease the design of the sensing circuitry, especially the cascode amplifier.

Chapter 6 presents the final read path simulation results as well as the measured performance of the 1.8V, two-bit-per-cell, 256Mbits flash memory. Chapter 7 provides the the conclusion remarks and suggests areas for future work.

Chapter 2

Sense current recovery technique

This chapter describes the Sense Current Recovery Technique, a new sensing technique developed in this research to substantially reduce the side-leakage current, which has been a notorious problem for virtual-ground memories. The technique is fast, effective and easy to implement. In Section 2.1, a preliminary solution to the side-leakage problem is described because it leads naturally to the Sense Current Recovery Technique. Section 2.2 focuses on the two main components of the Sense Current Recovery Technique, which are the use of multiple bitlines and multiple protecting bitlines to recover the read margin loss caused by the side-leakage current and the use of a unity feedback amplifier to eliminate the voltage mismatch between the drain bitlines and the protecting bitlines. HSPICE simulation results are presented at the end of Section 2.2 to show the effectiveness of this new sensing technique. Section 2.3 summarizes the chapter and provides some implications of the Sense Current Recovery Technique on other operation modes besides the read operation.

2.1 Preliminary Solution to the side-leakage problem

The Preliminary Solution described here is one of the early versions of the Sense Current Recovery Technique. The way it tries to fix the “1” read margin loss (for erased bits) and minimize the “0” read margin loss (for programmed bits) leads directly to the Sense Current Recovery Technique in Section 2.2.

2.1.1 Fixing the “1” read margin loss -- A Preliminary Solution

As mentioned in Section 1.2.1.2 of Chapter 1, the “1” margin loss is caused by the side-leakage current Ileak1 flowing from the protecting bitline “P” to the drain bitline “D” as shown in Figure 1.8, which is similar to the bottom part of Figure 2.1 shown below.

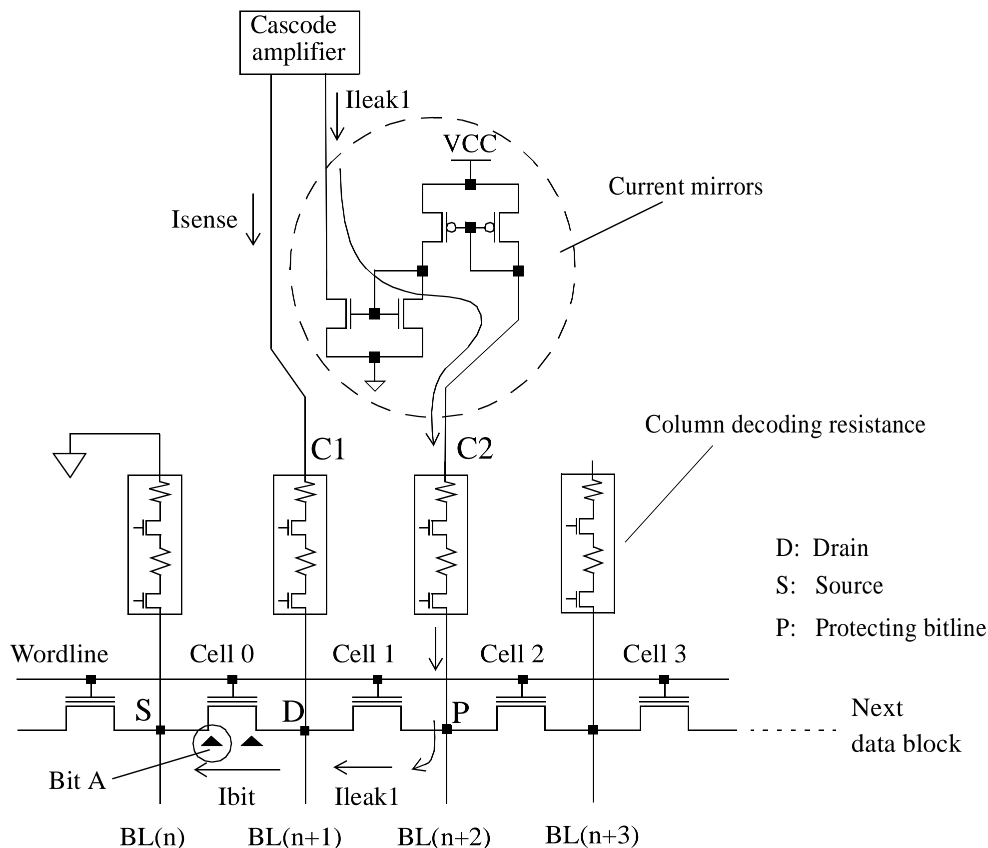


Figure 2.1: Preliminary Solution for fixing “1” read margin loss

In Figure 2.1, assume that the bit being read is Bit A of Cell 0, and that Cell 2 is programmed hard enough, preventing any side-leakage current flowing through it. Also assume that Cell 1 is an erased cell thus it does not stop the side-leakage running from node P to node D. Suppose that Bit A is an erased bit ("1" bit), then it should have a large current Ibit; node D will be pulled down more than node P, creating a side-leakage current Ileak1 from P to D. Ideally, the current sent to the sensing circuitry Isense should be equal to the current of Bit A, but when the side-leakage Ileak1 is created, Isense = Ibit - Ileak1, not Ibit, and therefore the read margin loss for Bit A is Ileak1. If Ileak1, by some mean, is added back to Isense, then there is no "1" margin loss anymore. This is the main idea of the Preliminary Solution for fixing the "1" read margin loss shown in Figure 2.1. In the figure, 2 current mirrors are used to send the side-leakage current Ileak1 to the cascode amplifier, where it will be added back to Isense, so the margin loss for "1" is eliminated. However, if A is a programmed bit ("0" bit), the action of the current mirrors actually makes the "0" read margin loss larger, and that is not good.

2.1.2 Minimizing the "0" read margin loss -- A Preliminary Solution

Figure 2.2 shows the "0" read margin loss mechanism. The figure is the same as Figure 1.7 in Section 1.2.1.1, but placed here again for convenience.

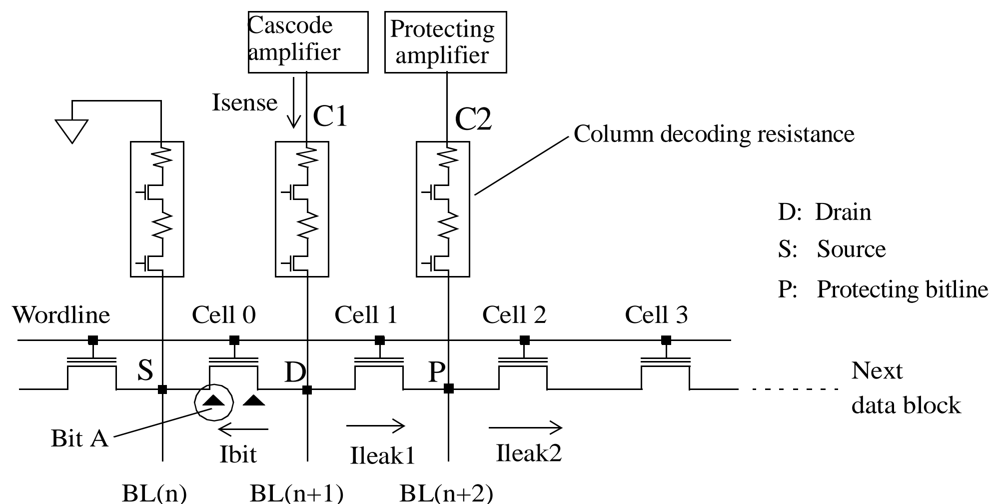


Figure 2.2: "0" read margin loss

In Figure 2.2, the bit being read is Bit A. Assume Bit A is a “0” bit (a programmed bit), which should have a small current Ibit. Also assume that Cell 1, Cell 2 and all the cells on the right of Cell 2 are over-erased cells. This condition, as discussed in Chapter 1, will cause a large side-leakage current Ileak2 of about 20µA. Therefore node P is pulled down lower than node D, creating a leakage current Ileak1 flowing from node D to P. Therefore, the current Isense is not equal to Ibit as expected, but Isense = Ibit + Ileak1 as shown in Figure 2.2, at node D. Thus the “0” read margin loss is Ileak1. If Ileak1 is substracted from Isense, then this “0” margin loss is eliminated. The two current mirrors discussed in the previous section (Section 2.1.1) help to eliminate the “1” read margin loss, but in the case for “0” bits, it causes more margin loss. For a “0” bit, instead of substrating Ileak1 from Isense, the current mirrors blindly add Ileak1 to Isense, without knowing that Ileak1 now flows from D to P, not P to D. To minimize the negative effect of the current mirrors for “0” bits, the Preliminary Solution tries to minimize Ileak1, thus if Ileak1 is wrongly added to Isense, the error is still small.

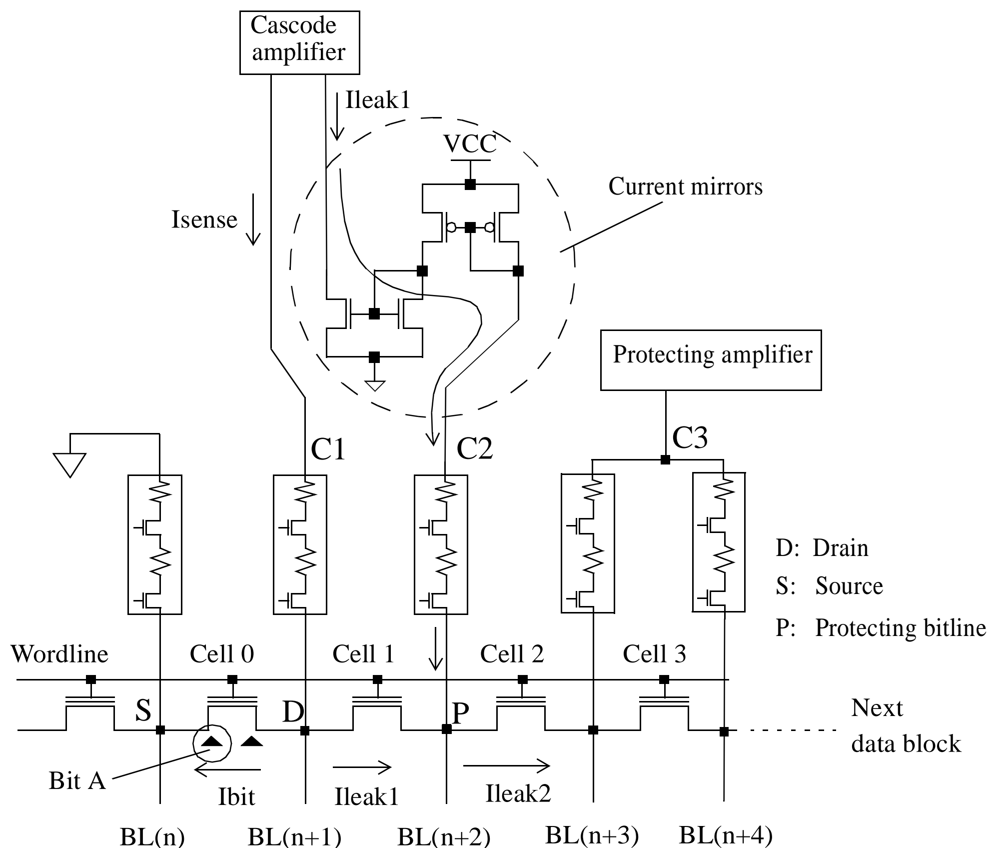


Figure 2.3: Preliminary Solution to minimize the “0” read margin loss

Figure 2.3 shows the Preliminary Solution for minimizing Ileak1, in which BL(n+3) is driven by a protecting amplifier to a voltage similar to the voltage at node P, thus Ileak2 is reduced due to the buffering (protecting) action of BL(n+3). When Ileak2 reduces, Ileak1 will be reduced because it is Ileak2 that causes Ileak1. It is even better if both BL(n+3) and BL(n+4) are driven by the protecting amplifier. With the protection of two additional bitlines, Ileak2 is minimized significantly leading to a large reduction in Ileak1.

In short the Preliminary Solution tries to recover the "1" read margin loss by using current mirrors to add the side-leakage current back to Isense, and tries to minimize the "0" read margin loss by using additional protecting bitlines. The Preliminary Solution, however, has some major drawbacks. First of all, when the supply voltage is at 1.6V (the minimum supply voltage level given in the Specifications), there is no voltage headroom to build the PMOS current mirror because the threshold voltage of these PMOSs is high. Secondly, the current mirror response is slow, degrading the memory read speed. Finally, in Figure 2.3, there is no guarantee that the voltage at node C3 is equal to that of node C2, thus Ileak2 may not be reduced to an acceptable level. All these drawbacks are solved by the Sense Current Recovery Technique, which is the focus of Section 2.2.

2.2 Sense Current Recovery Technique

The Sense Current Recovery Technique is the complete solution to solve the read margin loss problem caused by the side-leakage current in the virtual-ground memory array. It is very effective in recovering the read margin loss for both "1" and "0" bits, which is what the Preliminary Solution tries to achieved. The Sense Current Recovery Technique, however, does not have the drawbacks that the Preliminary Solution suffers such as lacking of voltage headroom, slow speed and bitline voltage mismatch. These drawbacks have been mentioned at the end of Section 2.1.2. Section 2.2.1 and Section 2.2.2 focus on the two essential ideas of the Sense Current Recovery Technique, which are the use of

multiple drains, multiple protecting bitlines to recover the read margin loss, and the use of a unity feedback amplifier to eliminate the voltage mismatch between the drain bitlines and the protecting bitlines. Section 2.2.3 presents the HSPICE results to show the effectiveness of the Sense Current Recovery Technique.

2.2.1 Recover the read margin loss by using multiple drains and multiple protecting bitlines

The Preliminary Solution recovers the "1" margin loss by adding the leakage current Ileak1 back to the Isense current, using current mirrors. However, the current adding operation can be performed by just connecting the bitline BL(n+2) to BL(n+1) at node C1 through the column decoding network as shown in Figure 2.4. Effectively, adding operation is performed by using just a "wire". Note that in Figure 2.4, "S1", "S2", "D1", "D2", "D3", "P1", "P2" and "P3" are two source bitlines, three drain bitlines and three protecting bitlines, respectively.

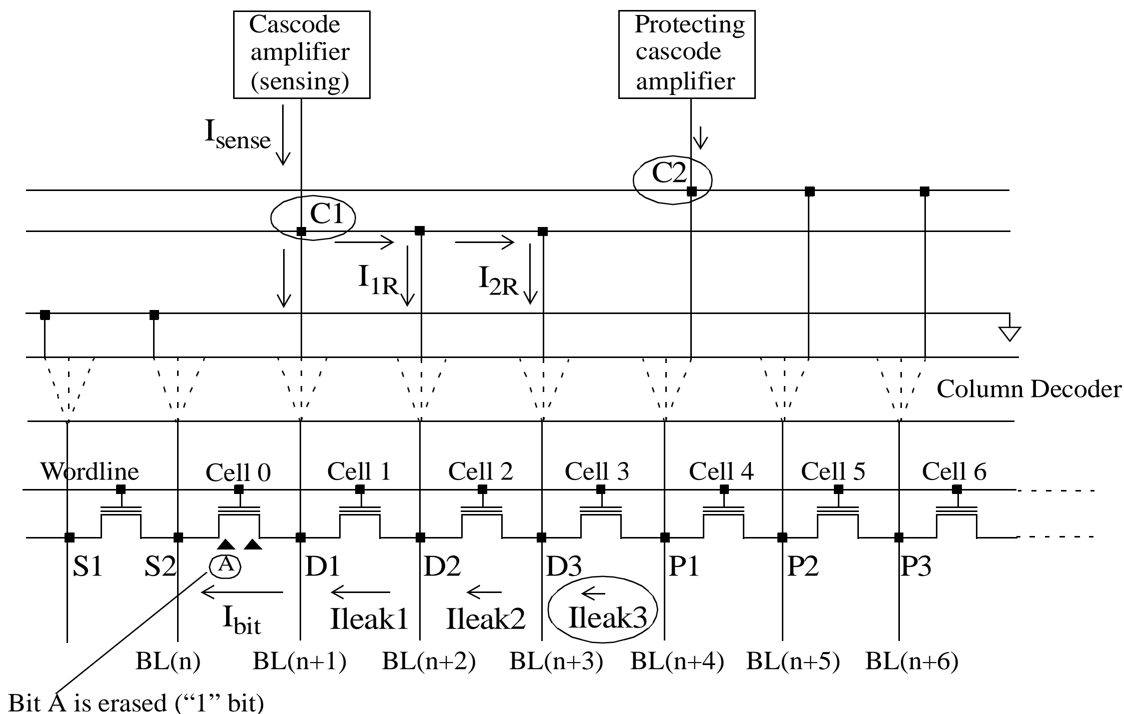


Figure 2.4: Sense Current Recovery Technique - Recover "1" read margin loss

Using a “wire” is the most accurate and simplest way. In addition, a “wire” does not have the voltage headroom problem as the current mirror does. The wire is also the fastest component, thus this approach is the fastest current-adding approach.

Because BL(n+2) is now connected to BL(n+1), it effectively becomes the second drain because its current is also sent to the cascode amplifier. Similarly, BL(n+3) can also be connected to the same node C1 through the column decoder to recover the “1” read margin loss even more, effectively forming three drain bitlines. The current recovery action resulted from the use of multiple drain bitlines is now analyzed a little further. Because the current of the erased Bit A is large, there is a large voltage drop at node D1. Although the cascode amplifier is strong, node D1 still drops because of the resistance of the column decoding path. Node D2 drops less because it is buffered by node D1, and node D3 drops even less because it is protected from the large current Ibit (of Bit A) by both D1 and D2. This situation leads to the relationship Ileak1 >> Ileak2 >> Ileak3. Ileak1 and Ileak2 are recovered by using the column decoder to connect the three drain bitlines. The remaining current loss is Ileak3, but it is very small. Note that Cell 1 to Cell 5 in Figure 2.4 are assumed to be erased cells.

Not only is the read margin loss for “1” bits recoverd, the Sense Current Recovery Technique also helps to recover most of the read margin loss for “0” bits by using multiple protecting bitlines as shown in Figure 2.5. The protecting bitlines are coupled together at node C2 through the column decoder. If Bit A is programmed, Ibit should be small, but Isense could have been erroneously large due to the side-leakage if there was only one protecting bitline. The use of multiple protecting bitlines significantly reduces the read margin loss for “0” bits. In Figure 2.5, the remaining margin loss for Bit A, which is a “0” bit, is Ileak4; note that Ileak4 is the current at the interface of the drain bitline group and the protecting bitline group. Ileak4 is very small even though Ileak7 is very large (about 20µA as shown in Figure 1.6 of Chapter 1) because node P1 is well protected from Ileak7 by nodes P2 and P3.

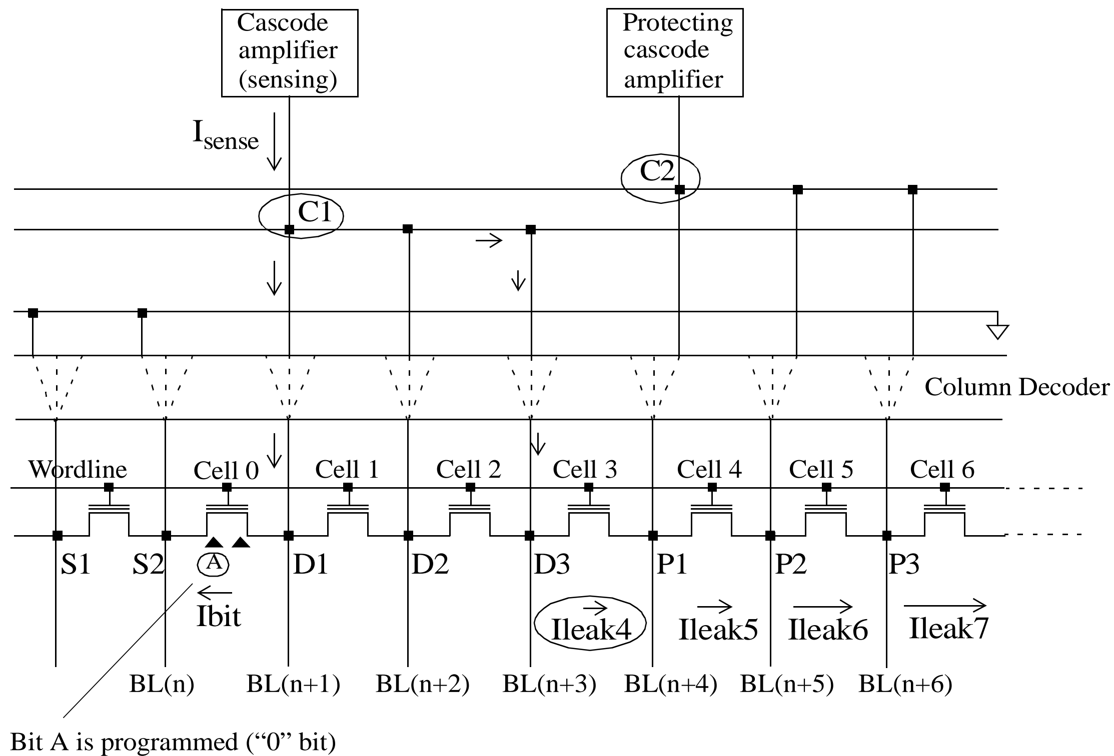


Figure 2.5: Sense Current Recovery Technique - Recover “0” read margin loss

Figure 2.4 and 2.5 reveal that there is no floating bitline between the drain bitline group and the protecting bitline group, which is an important characteristic of the Sense Current Recovery Technique. As mentioned in Section 1.2.1.2 of Chapter 1, a previous design [6] tried to float the bitline between the drain bitline and the protecting bitline to reduce the side-leakage current, but doing so causes a slow read access due to the increase of the side-leakage ac current component, the current to charge the high capacitive floating bitline. The Sense Current Recovery Technique is fast because the floating bitline situation does not exist. All the bitlines involving in the read operation are driven either from the cascode amplifier or the protecting amplifier.

2.2.2 Eliminating the voltage mismatch between the drain bitlines and the protecting bitlines

This section focuses on the second important idea of the Sense Current Recovery Technique, which is the use of a unity feedback amplifier to eliminate the voltage mismatch between the drain bitline group and the protecting bitline group as shown in Figure 2.6.

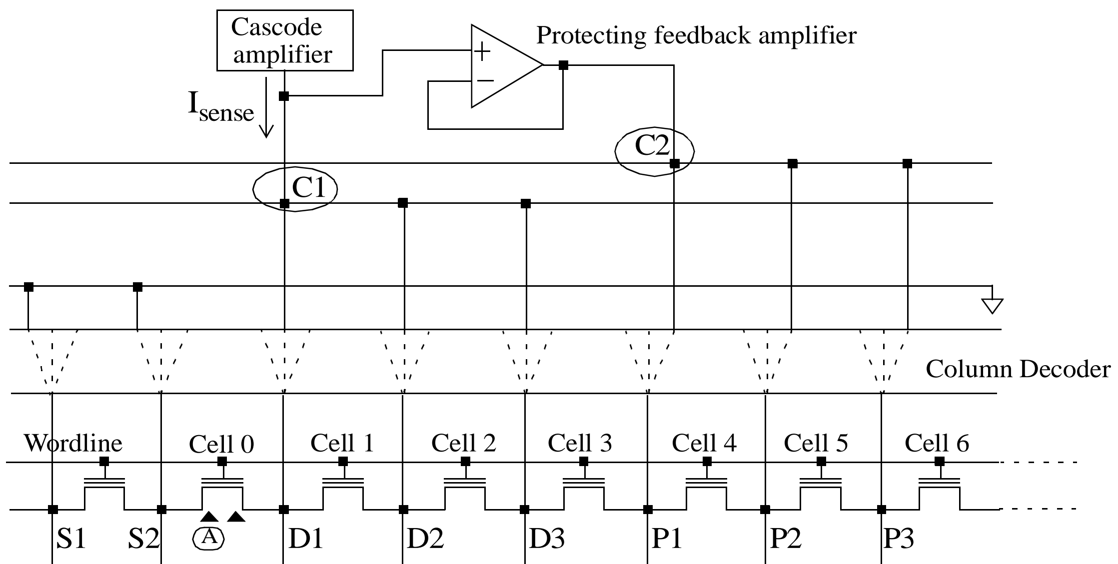


Figure 2.6: Sense Current Recovery Technique - Solving voltage mismatch problem

Recall in the previous section that by using multiple drains, multiple protecting bitlines, the read margin loss is substantially recovered. The remaining margin loss for “1” is Ileak3, and the remaining margin loss for “0” is Ileak4 shown in Figure 2.4 and Figure 2.5, respectively. Ileak3 and Ileak4 are very small if the voltage at node C1 equals to the voltage at node C2. In fact, there is a large mismatch between them, about 60mV in magnitude, even though the cascode amplifier and the protecting cascode amplifier in figures 2.4 and 2.5 are very similar. This mismatch occurs because the current drawn out

of the cascode amplifier can be different from the current drawn out of the protecting cascode amplifier. The mismatch can not be compensated because it depends on the data pattern stored in the memory cells. For example, if Bit A is an erased bit, it will draw about 30µA out of the cascode amplifier, while the protecting cascode amplifier may need to provide only 2µA for the side-leakage current.

In Figure 2.6, by using a unity gain feedback amplifier, the voltage at node C2 is forced to be very close to the voltage at node C1, thus Ileak3 and Ileak4 are suppressed much further. To have a good tracking between C1 and C2, the open-loop gain of the unity feedback amplifier should be high, so the number of stages of this amplifier should be more than one. However, this amplifier is also a feedback amplifier, in which a good frequency response is required, therefore the number of stages should be limited. A simple two-stage op-amp in [11] can be used for this purpose. Figure 2.7 shows the actual protecting feedback amplifier, in which a PMOS current mirror load is used to maximize the open-loop gain.

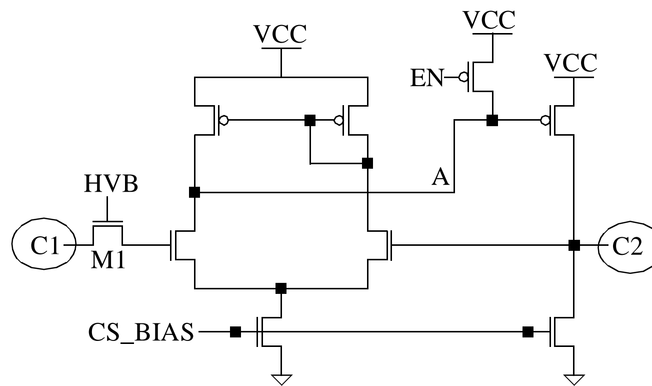


Figure 2.7: Protecting feedback amplifier

As discussed in [12], a simplified formula for gain is $g_m / g_o = \frac{2L}{V_{gs} - V_T}\left(\frac{dx_d}{dV_{ds}}\right)^{-1}$, thus to have high gain, all the transistors in Figure 2.7, except M1, are thin-oxide enhancement transistors. In addition, the bias current is controlled to be small enough to reduce the power consumption and also to have high gain, but the bias current should not be too small

because speed is also an important factor; the maximum dc bias current for the protecting feedback amplifier shown in Figure 2.7 has been controlled to be less than 70µA. Note that the use of the high voltage transistor M1 is for protecting the thin oxide transistor in the protecting feedback amplifier when C1 is raised to high voltages in some modes such as programming or erase mode. Transistor M1 is an intrinsic transistor and has a threshold voltage close to 0, or even a negative threshold.

Like most other two-stage op-amps, the protecting feedback amplifier in Figure 2.7 needs frequency compensation. There are many good ways to compensate this amplifier such as [13], [14] and [15], but in the 256Mb flash memory described here, a traditional and simple compensation method is used to reduce the area of this amplifier. The dominant pole frequency is lowered by connect a capacitive load CL at node C2, and the first nondominant pole frequency is increased as high as possible by minimizing the capacitive loading at node A in Figure 2.7. This approach is convenient because the capacitive load CL comes at no layout or area expense. It is the capacitance of the 3 protecting bitlines (and all other parasitic capacitance associated with this node), which can be 2pF.

The schematic shown in Figure 2.8 is used to test the performance of the protecting feedback amplifier. Node C1 represents the group of three drain bitlines, which has a total capacitance of about 2pF. Node C2 represents the group of three protecting bitlines.

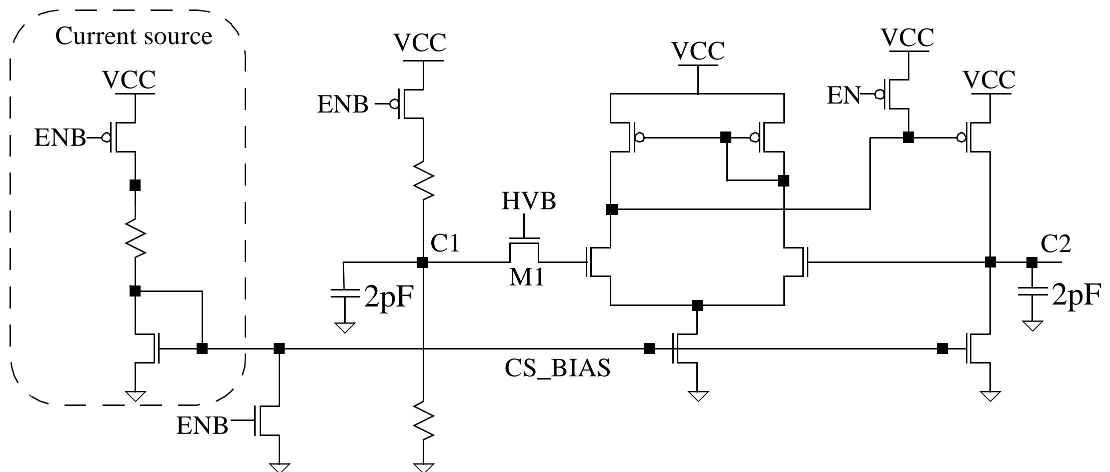


Figure 2.8: Schematic for testing the protecting feedback amplifier performance

Note that the voltage at node C1, which represents the bitline voltage, is derived directly from a resistive divider. As described in Chapter 4, deriving the bitline voltage from a resistive divider offers many advantages.

HSPICE simulation results for the protecting feedback amplifier in Figure 2.8 are shown in Figure 2.9. C2 apparently tracks C1 very well, within 5mV. In addition, the speed of C2 is acceptable, about 6ns for the node to reach 90% of the final value. Remember that without using this feedback amplifier, the voltages at nodes C1 and C2 shown in Figures 2.4 and 2.5 can be different by 60mV, which is not acceptable.

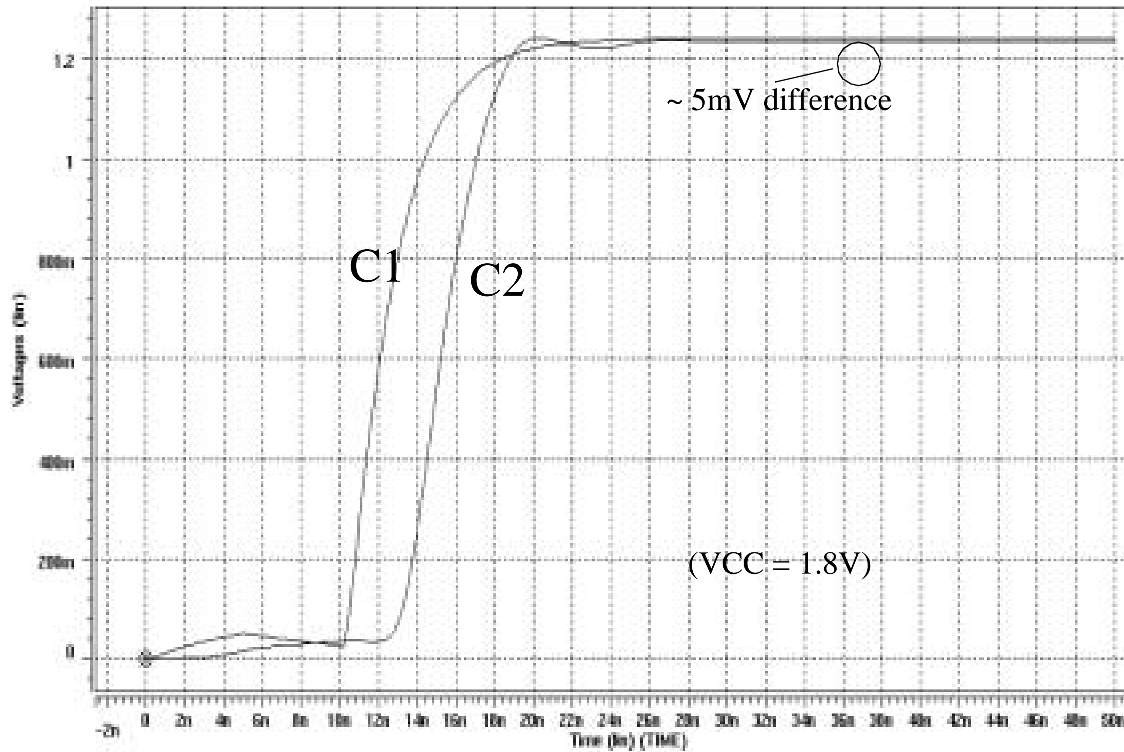


Figure 2.9: HSPICE simulation result for the protecting feedback amplifier

2.2.3 HSPICE simulation results for the Sense Current Recovery Technique

This section presents the HSPICE simulation results for the worst case “1” read margin loss as well as the worst case “0” read margin loss to show the effectiveness of the Sense Current Recovery Technique.

2.2.3.1 Simulation result for the worst case “1” read margin loss

As mentioned at the beginning of this chapter, the “1” margin loss is caused by the side-leakage current flowing from the protecting bitlines to the drain bitlines, so the worst case is when this current is largest. Figure 2.10 shows the set-up for this worst case in which Cell 6 is an over-programmed cell whereas Cell 1 to Cell 5 are over-erased cells.

Figure 2.10: Worst case “1” read margin loss

An over-programmed cell is the one with a high threshold voltage for both bits in the cell and conducts no current thus it can be represented by an open circuit or a large resistor such as 1 Giga-Ohm resistor (point D in Figure 2.11). An over-erased cell, on the other hand, has a very low threshold voltage for both bits in the cell, and the equivalent resistance for such a cell is about 16 KOhms (point A in Figure 2.11). Assume the bit being read in Figure 2.10 is Bit A, and to be a worst case for a "1" bit, Bit A should be an erased bit with the highest threshold voltage, or smallest "1" read margin (point B in Figure 2.11). The equivalent resistance of Bit A in this case is about 36 KOhms.

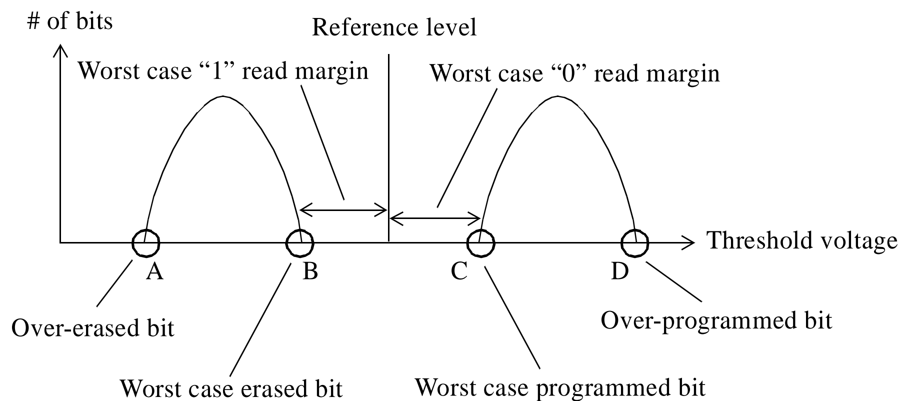


Figure 2.11: Threshold distribution of the erased and programmed bits

Because Cell 6 in Figure 2.10 is an over-programmed cell, it blocks all side-leakage current through it, thus all of the current provided by the protecting feedback amplifier is redirected toward the group of three drain bitlines, and Ileak3 is maximum in this case. Thus the set-up in Figure 2.10 is truly the worse case for the "1" read margin loss.

HSPICE simulations were conducted using a model of the circuit shown in Figure 2.10. The model is shown in Figure 2.12 in which the memory cells in Figure 2.10 are replaced by their equivalent resistances. The column decoding path for each bitline is about 4 KOhms. To simplify the simulation even more, assume that the protecting feedback

amplifier is ideal, thus both nodes C1 and C2 have the same votage, which is set to be 1.2V. The simulation results are shown in Figure 2.12.

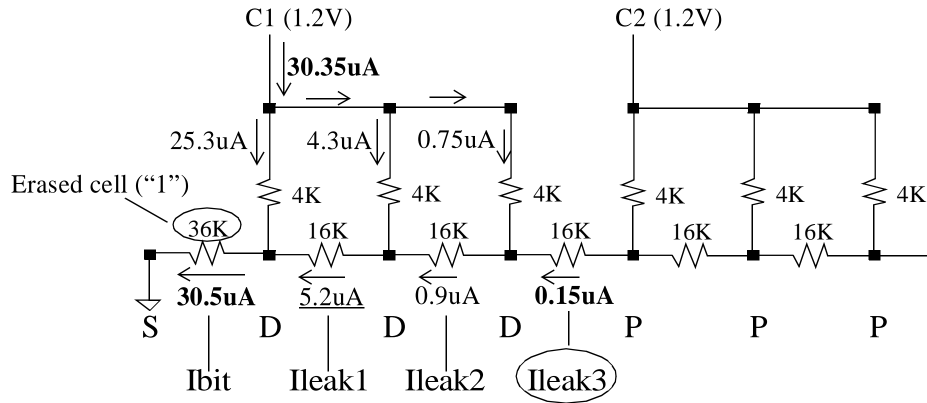


Figure 2.12: HPICE simulation for the worst case “1” read margin loss

Figure 2.12 shows that the Sense Current Recovery Technique is very effective in recovering the read margin loss for “1” bits. The remaining margin loss for “1” is only 0.15µA (Ileak3 in Figure 2.12). If there were only one drain bitline and one protecting bitline, the “1” read margin loss could have been 5.2µA (Ileak1 in Figure 2.12), or even more when there is no protecting feedback amplifier to equate the voltages at nodes C1 and C2. The Sense Current Recovery Technique, therefore, helps to reduce the “1” margin loss at least by a factor of 30.

2.2.3.2 Simulation result for the worst case “0” read margin loss

The “0” margin loss is caused by the side-leakage current flowing from the drain bitlines to the protecting bitlines, so the worst case for the “0” read margin loss is when this current is largest. Figure 2.13 shows the set-up for this worst case in which Cell 3 as well as all the cells on its right are over-erased cells. In this case, the side-leakage current

Ileak7 can be as large as 20µA, as shown in Figure 1.6 of Chapter 1. Assume the bit being read in Figure 2.13 is Bit A, and to be the worst case for a “0” bit, Bit A should be the programmed bit with the lowest threshold voltage, or smallest “0” read margin (point C in Figure 2.11). The equivalent resistance of Bit A in this case is about 50 KOhms.

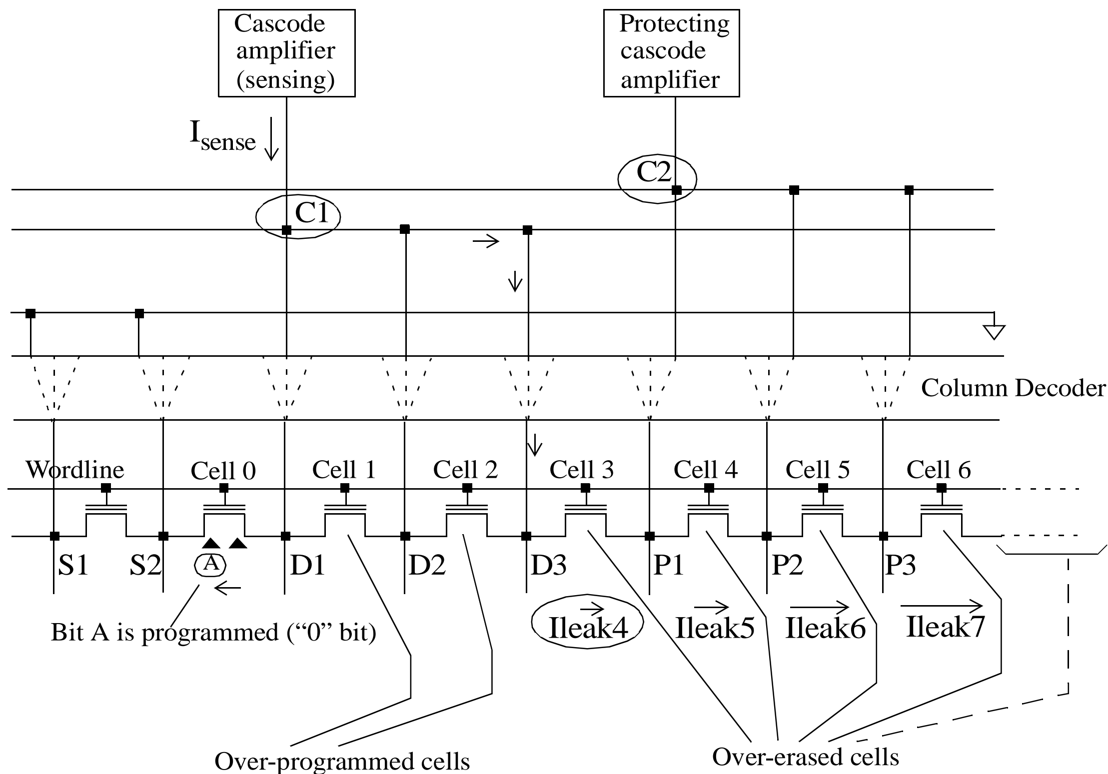


Figure 2.13: Worst case “0” read margin loss

Cell 1 and Cell 2 can be set to be over-programmed cells, which conduct no current. The set-up in Figure 2.13 is the worst case for the “0” bit, because the “0” read margin loss Ileak4 is maximum, which is because Ileak7 is maximum and Cell 3 to Cell 6 are over-erased cells.

HSPICE simulations were conducted on a model of the circuit shown in Figure 2.13. The model is shown in Figure 2.14 in which the memory cells in Figure 2.13 are replaced by

their equivalent resistances. Again, assume the voltages at nodes C1 and C2 are the same and equal to the bitline voltage of 1.2V. The simulation results are shown in Figure 2.14.

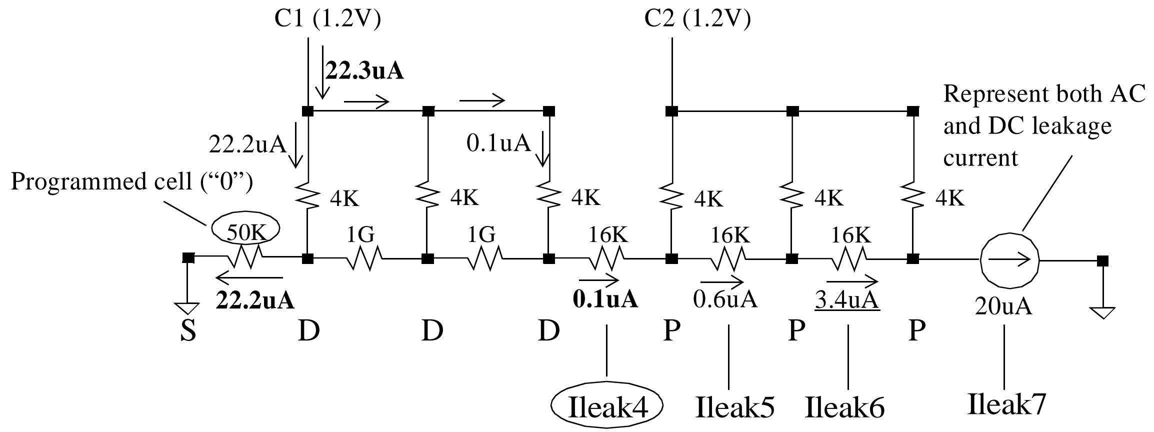


Figure 2.14: HPICE simulation for the worst case "0" read margin loss

Figure 2.14 shows that the Sense Current Recovery Technique is also very effective in recovering the read margin loss for "0" bits. The remaining margin loss for "0" is only 0.1µA (Ileak4 in Figure 2.14). If there were only one protecting bitline, the "0" read margin loss could have been 3.4µA (Ileak6 in Figure 2.14), or actually even more when there is no protecting feedback amplifier to enforce the equality in voltage at nodes C1 and C2. Again, the Sense Current Recovery Technique helps to reduce the "0" margin loss at least by a factor of 30.

2.3 Summary

By the use of multiple drain bitlines, multiple protecting bitlines and a protecting unity feedback amplifier, the Sense Current Recovery Technique has solved the margin loss problem caused by side-leakage current effects. In general, any number of drain or

protecting bitlines can be used. The discussions in this section makes it clear that if there are more drain bitlines, the read margin loss for "1" bits is recovered more and if there are more protecting bitlines, the read margin loss for "0" bits are suppressed more. A decision on the number of drain and protecting bitlines is based on the trade-off between having acceptable read margin, power consumption and column decoding size. For this 1.8V, two-bit-per-cell flash memory, to gain enough read margin for both "0" and "1" bits, three drain bitlines and three protecting bitlines must be used. The column decoding supporting the multiple drain bitlines, multiple protecting bitlines will be discussed thoroughly in the next chapter.

As a final remark, the side-leakage current problem, or sometimes called the "pattern-sensitivity" problem has been a serious problem in virtual-ground memories and difficult to solve because the side-leakage current can be present or not, depending on the data pattern stored in the array. In this chapter, the Sense Current Recovery Technique has been proved to be highly effective in dealing with this pattern-sensitivity problem without degrading the read speed. This new technique can also be applied for other modes of operation as well, such as programming and erasing verify operations; in these modes, the decoding pattern S-S-D-D-D-P-P-P as well as the protecting feedback amplifier can be easily activated, helping to achieve accurate programming and erasing threshold voltages regardless of the data pattern in the memory array.

Chapter 3

Multiple-drain-bitline and multiple-protecting-bitline column decoding

This chapter presents the overall architecture of the column decoding and the methodology for designing the multiple-drain-bitline, multiple-protecting-bitline column decoding, particularly the S-S-D-D-D-P-P-P decoding style mentioned in Chapter 2. Section 3.1 briefly introduces the chip architecture, in which the column decoding is shown as a collection of identical decoding blocks for each memory bank. Section 3.2 describes in detail the design of the column decoding for a memory bank. The section starts with the overall architecture of the bank decoding, then the design methods are presented. The methods can be applied not only for the S-S-D-D-D-P-P-P decoding style, but also for any multiple-source-drain-protecting-bitline decoding style. Logic simulation results for the decoding S-S-D-D-D-P-P-P are also presented. The column decoding for two special cases at the edges of each memory bank are discussed toward the end of the section.

3.1 Chip architecture

The chip is divided into 16 independent banks to facilitate simultaneous operation, in which one bank can be read while another bank is still busy with a programming or an erasing operation. A bank consists of 16 1-Mbit sectors; these sectors share the same 512 regular global bitlines, which run above the sectors in the vertical direction. For each sector, there are 512 poly wordlines, decoded by the row decoding circuitry. The row decoding, however, is not emphasized in this chapter. The chip has a total of 256 sectors and its density is 256 Mbits.

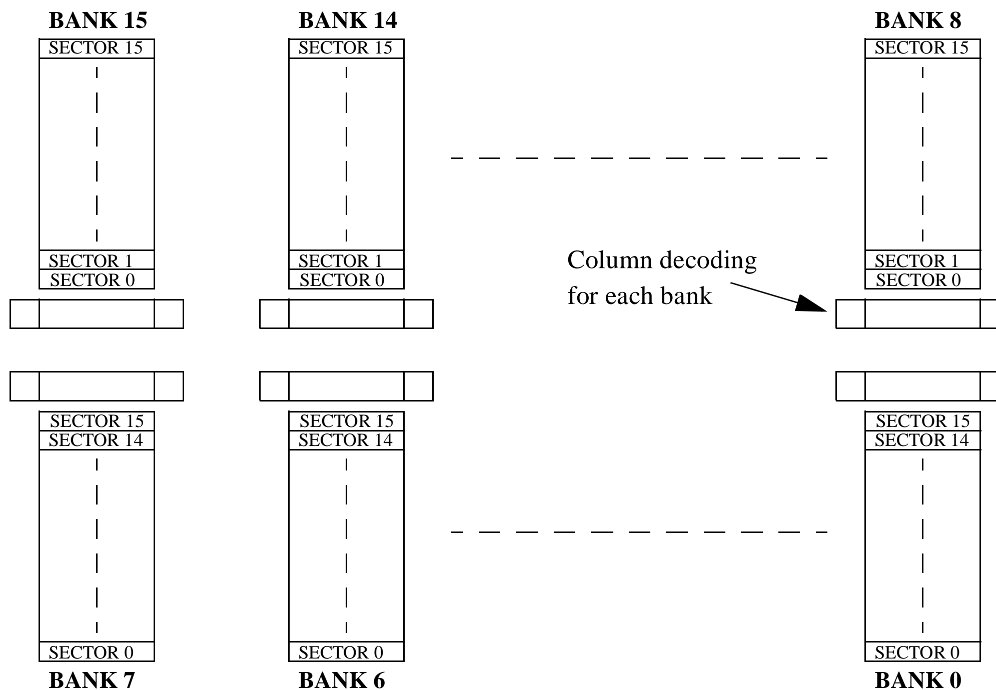


Figure 3.1: Chip architecture

3.2 Column decoding architecture

The column decoding is identical for every bank, thus the design of the column decoding can be described fully based on the decoding for one bank; Figure 3.2 shows the column decoding architecture. The local bitlines for each sector are decoded from 512 global bitlines, which are made from the top metal layer and are common for all 16 sectors in a

bank. The global bitlines can not be connected directly to the memory cells due to the wider pitch. The local bitlines use the first metal layer and can be made on a finer pitch, thus they can be connected to the memory cell drains through the diffusion layer. Two local bitlines are decoded from one global bitline, thus there are a total of 1024 regular local bitlines for each sector.

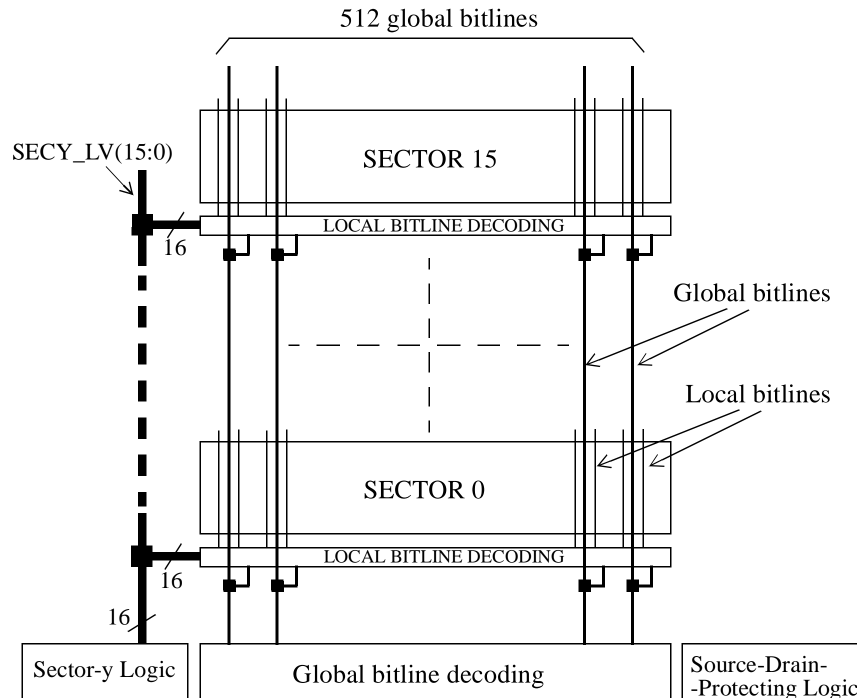


Figure 3.2: Bank column decoding architecture

The column decoding contains 4 main blocks: the local bitline decoding, the global bitline decoding, the source-drain-protecting logic and the sector-y logic blocks, respectively. Horizontally, a sector is divided into 64 identical data blocks so that 64 bits can be read out at the same time for each sensing cycle, each bit from one data block. The arrangement of these 64 data blocks is shown in Figure 3.3. Each data block is 16 memory cells wide and has 16 local bitlines (decoded from 8 global bitlines). There are 512 wordlines running horizontally through all 64 data blocks, and a wordline runs through 16 memory cells in each data block.

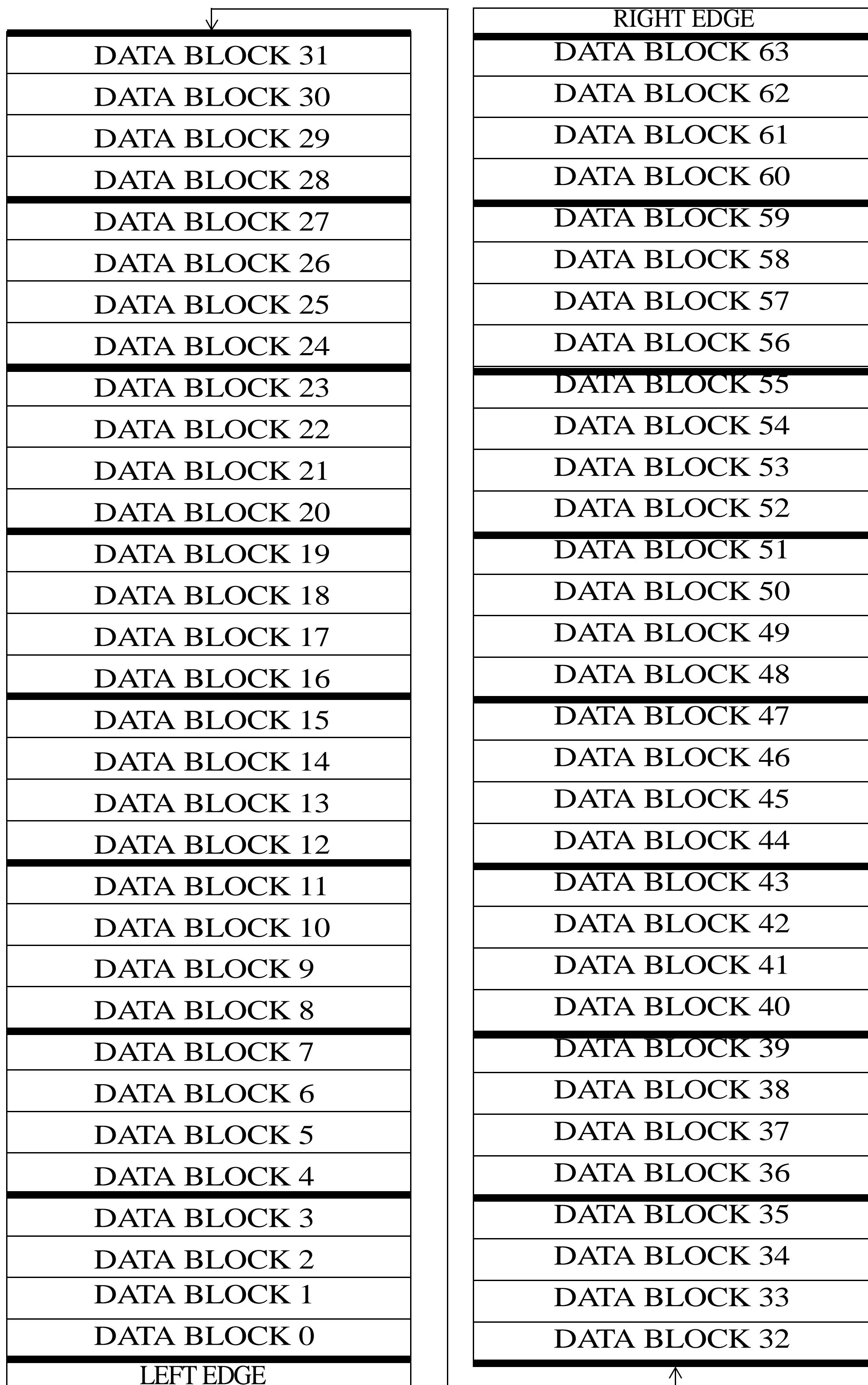


Figure 3.3: Data block arrangement

Because each cell stores 2 bits, there are 32 bits for a wordline in each data block. Note that the column decoding for each data block is identical, therefore the column decoding is basically the decoding for the 32 bits in a data block.

Five address pins A(6:2) are used to decode these 32 bits. The physical location for these 32 bits can be arranged in many ways. Figure 3.4 shows one typical arrangement, in which bits 0 and 16 reside in Cell 0, bits 1 and 17 in Cell 1, etc. All the design methodologies from now on will be based on this arrangement, but the methodologies are general and can be applied to any physical arrangement of the bits. In Figure 3.4, to read the left bit of Cell 4 (bit 4), LBLn_(3:2) are the source bitlines, LBLn_(6:4) are the drain bitlines and LBLn_(9:7) are the protecting bitlines. Similarly, to read the right bit of Cell 12 (bit 28), LBLn_(13:12) are the source bitlines, LBLn_(11:9) are the drain bitlines and LBLn_(8:6) are the protecting bitlines.

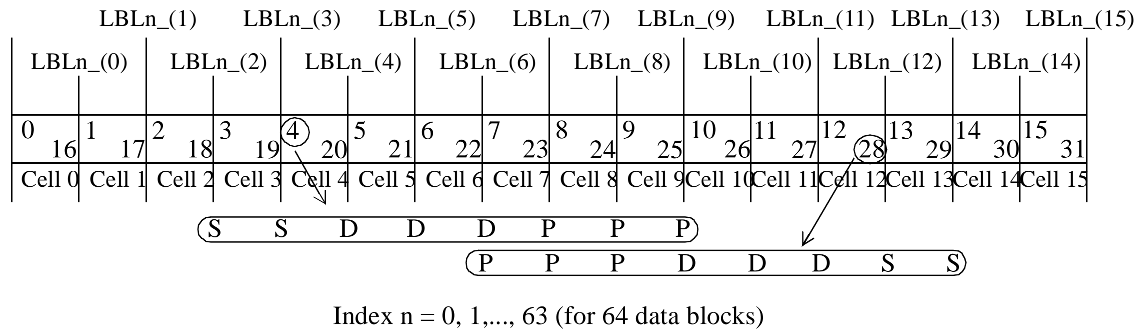


Figure 3.4: Physical address for 32 bits on a wordline for a data block

3.2.1 Local bitline decoding block

A data block has 16 local bitlines LBLn_(15:0), which are decoded from 8 global bitlines GBLn_(7:0). Two local bitlines are decoded from one global bitline, but to realize the decoding pattern S-S-D-D-D-P-P-P, there must be a special arrangement between the

global bitline and its two associated local bitlines. The column decoding discussed in [16] to [20] can not generate the pattern S-S-D-D-D-P-P-P without causing signal contention. The new column decoding described in this chapter can achieve the decoding pattern S-S-D-D-D-P-P-P without signal contention by gating two local bitlines that are eight memory cells apart to a global bitline as shown in Figure 3.6. Particularly, local bitlines LBLn_(0) and LBLn_(8) must come from global bitline GBLn_(0), local bitlines LBLn_(1) and LBLn_(9) must come from global bitline GBLn_(2), etc.

The 16 control signals SECYs_(15:0) are common for all 64 data blocks in a sector. SECYs_(15:0) are the local versions of the global signals SECY_LV(15:0) shown in Figure 3.2. The 16 lines SECYs_(15:0) are grounded if the sector is not selected. In addition, SECYs_(15:0) are level-shifted from 1.8V to about 4V to reduce the resistance of the transistors connected between the global bitlines and local bitlines (these transistors are called **sector select transistors**). Figure 3.5 shows the sector select level shifter for decoding a local control line SECYs_(k) from a global control line SECY_LV(k) and also for converting from the VCC-level signal SECY_LV(k) to the VPPI-level signal SECYs_(k). VPPI is about 4V during a read access.

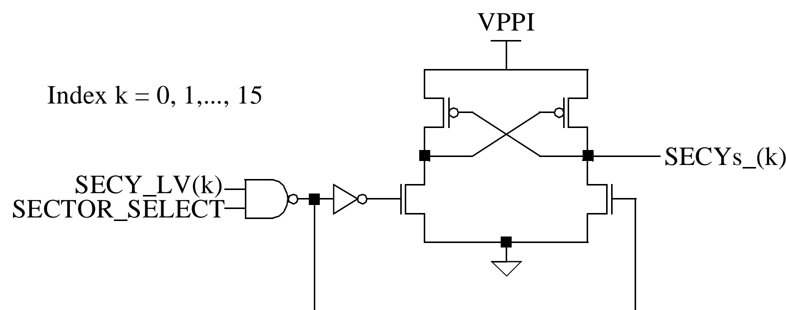


Figure 3.5: Sector select level shifter

In Figure 3.6, a local bitline decoding unit for one data block is shown. The unit is repeated 64 times for 64 data blocks.

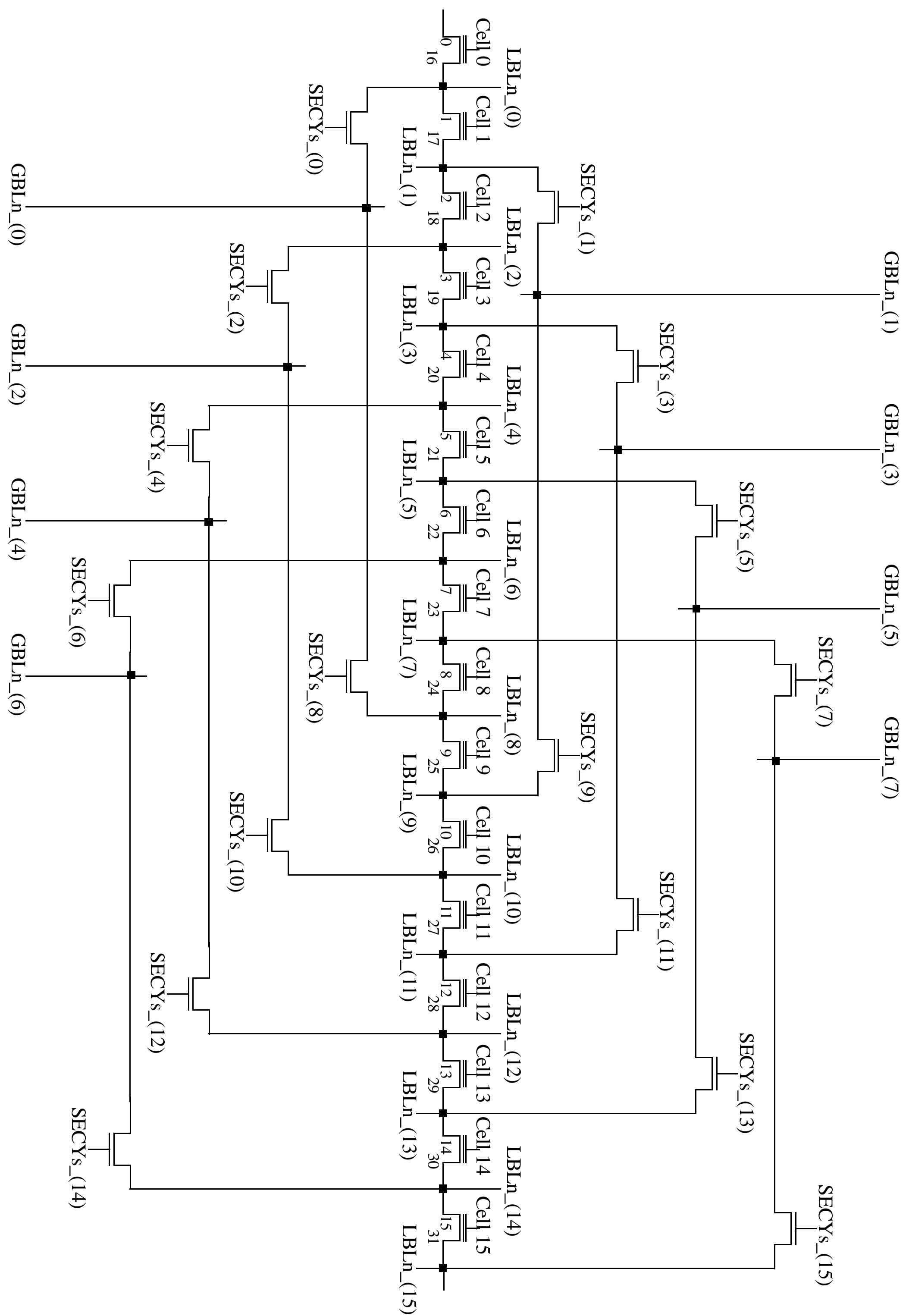


Figure 3.6: Local bitline decoding for one data block. Index n = 0, 1,..., 63 for 64 data blocks. Index s = 0, 1,..., 15 for 16 sectors in a bank.

3.2.2 Sector-y logic block

The sector-y logic block is used to generate the VCC-level signals SECY_LV(15:0) from the five address pins A(6:2). The design of this block starts with the construction of a truth table. In Figure 3.6, suppose that A(6:2) = 00100, then bit 4 (the left bit of Cell 4) will be read. The local bitlines LBLn_(9:2) must be selected to form the pattern S-S-D-D-D-P-P-P for bit 4. To select these eight local bitlines, SECYs_(9:2) must be high whereas SECYs_(15:10) and SECYs_(1:0) must be low. Similarly, to read bit 20 (the right bit of Cell 4), the address A(6:2) = 10100 should be applied and the local bitlines LBLn_(5:0) of the data block n as well as the local bitlines (15:14) of the previous data block - data block (n-1) - must be selected; for this to happen, SECYs_(15:14) and SECYs_(5:0) must be high whereas SECYs_(13:6) must be low. Note that when a bit near the edges of a data block is read, the local bitlines from the previous or the next data block may be borrowed to form the complete pattern S-S-D-D-D-P-P-P. Based on Figure 3.6, the required actions for SECY_LV(15:0) for the remaining 30 combinations of A(6:2) can be found, then a complete truth table for decoding SECY_LV(15:0) is constructed, as shown in Table 3.1. Based on this table, a logic synthesis tool can be used to synthesize the sector-y logic block. The Verilog code for SECY_LV(15:0) used by the synthesis tool is in the Appendix.

3.2.3 Global bitline decoding block

The decoding style S-S-D-D-D-P-P-P requires three different voltages to be passed onto the global bitlines (and eventually to the local bitlines): ground voltage for the two local source bitlines, the drain voltage for the three local drain bitlines and the protecting voltage for the three local protecting bitlines. Therefore, each global bitline is the output of a three-way multiplexor as shown in Figure 3.7. The control signal S(k), D(k) and P(k) can be turned on to pass the apropriate voltage to the global bitline GBLn_(k). Note that the index k assumes the values 0, 1,..., 7 for eight global bitlines in each data block.

Table 3.1: Truth table for the sector-y logic block

Bit #	A(6:2)	SECY_LV(0)	SECY_LV(1)	SECY_LV(2)	SECY_LV(3)	SECY_LV(4)	SECY_LV(5)	SECY_LV(6)	SECY_LV(7)	SECY_LV(8)	SECY_LV(9)	SECY_LV(10)	SECY_LV(11)	SECY_LV(12)	SECY_LV(13)	SECY_LV(14)	SECY_LV(15)	SECY_LV (15:0) (Hex)
0	00000	1	1	1	1	1	1									1	1	C03F
1	00001	1	1	1	1	1	1	1									1	807F
2	00010	1	1	1	1	1	1	1	1									00FF
3	00011		1	1	1	1	1	1	1	1								01FE
4	00100			1	1	1	1	1	1	1	1							03FC
5	00101				1	1	1	1	1	1	1	1						07F8
6	00110					1	1	1	1	1	1	1	1					0FF0
7	00111						1	1	1	1	1	1	1	1				1FE0
8	01000							1	1	1	1	1	1	1	1			3FC0
9	01001								1	1	1	1	1	1	1	1		7F80
10	01010									1	1	1	1	1	1	1	1	FF00
11	01011	1									1	1	1	1	1	1	1	FE01
12	01100	1	1									1	1	1	1	1	1	FC03
13	01101	1	1	1									1	1	1	1	1	F807
14	01110	1	1	1	1									1	1	1	1	F00F
15	01111	1	1	1	1	1									1	1	1	E01F
16	10000	1	1									1	1	1	1	1	1	FC03
17	10001	1	1	1									1	1	1	1	1	F807
18	10010	1	1	1	1									1	1	1	1	F00F
19	10011	1	1	1	1	1									1	1	1	E01F
20	10100	1	1	1	1	1	1									1	1	C03F
21	10101	1	1	1	1	1	1	1									1	807F
22	10110	1	1	1	1	1	1	1	1									00FF
23	10111		1	1	1	1	1	1	1	1								01FE
24	11000			1	1	1	1	1	1	1	1							03FC
25	11001				1	1	1	1	1	1	1	1						07F8
26	11010					1	1	1	1	1	1	1	1					0FF0
27	11011						1	1	1	1	1	1	1	1				1FE0
28	11100							1	1	1	1	1	1	1	1			3FC0
29	11101								1	1	1	1	1	1	1	1		7F80
30	11110									1	1	1	1	1	1	1	1	FF00
31	11111	1									1	1	1	1	1	1	1	FE01

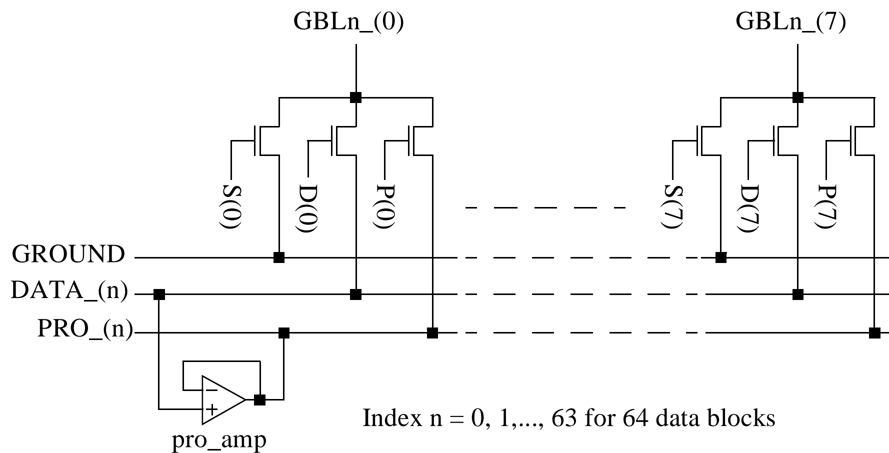


Figure 3.7: Basic global bitline decoding for one data block

The control signal S(k) can be designed easily because the ground voltage is global, meaning that it is not specific to a data block. For example, to read bit 3 (the left bit of Cell 3 in Figure 3.6), the global bitlines GBLn_(2) and GBLn_(1) must be at ground, thus S(2) and S(1) must be high. Similarly, to read bit 30 (the right bit of Cell 14 in Figure 3.6), the global bitlines GBLn_(7) and GBLn_(6) must be at ground, thus S(7) and S(6) must be high. The complete truth table to decode the control signal S(k) is shown in Table 3.2, page 52.

The designs for D(k) and P(k) are more complex. Both DATA_(n) and PRO_(n) lines are not global lines; they are specific to a data block. The sensing current for a bit in a data block should be sent to the DATA_(n) line of that data block. The complexity comes when a bit near the edges of a data block is accessed, for example bit 14 of **data block 2**, which is near the right edge of this data block. Because three drains are required, the local bitlines LBL2_(14), LBL2_(15) and LBL3_(0) are selected to be drains (Figure 3.6). Note that LBL2_(15:14) belongs to data block 2 and LBL3_(0) belongs to **data block 3**. LBL3_(0) is borrowed for the read access of bit 14 in the data block 2, and thus the current through LBL3_(0) must be sent back to DATA_(2), not DATA_(3). Figure 3.7, in which the index n is assumed to be three, clearly shows that D(0) can not be used in this case

because if D(0) is on, then LBL3_(0) will become a drain as expected, but the current through this drain will be sent to DATA_(3) through GBL3_(0), not DATA_(2). To fix this situation, a pass transistor is used to connect the global bitline GBL3_(0) of data block 3 to the DATA_(2) line of data block 2 as shown in Figure 3.8 (again assume n = 3), in case the current from GBL3_(0) needs to be sent back to data block 2. Note that the gate of this transistor is connected to the control line DL(0), not D(0).

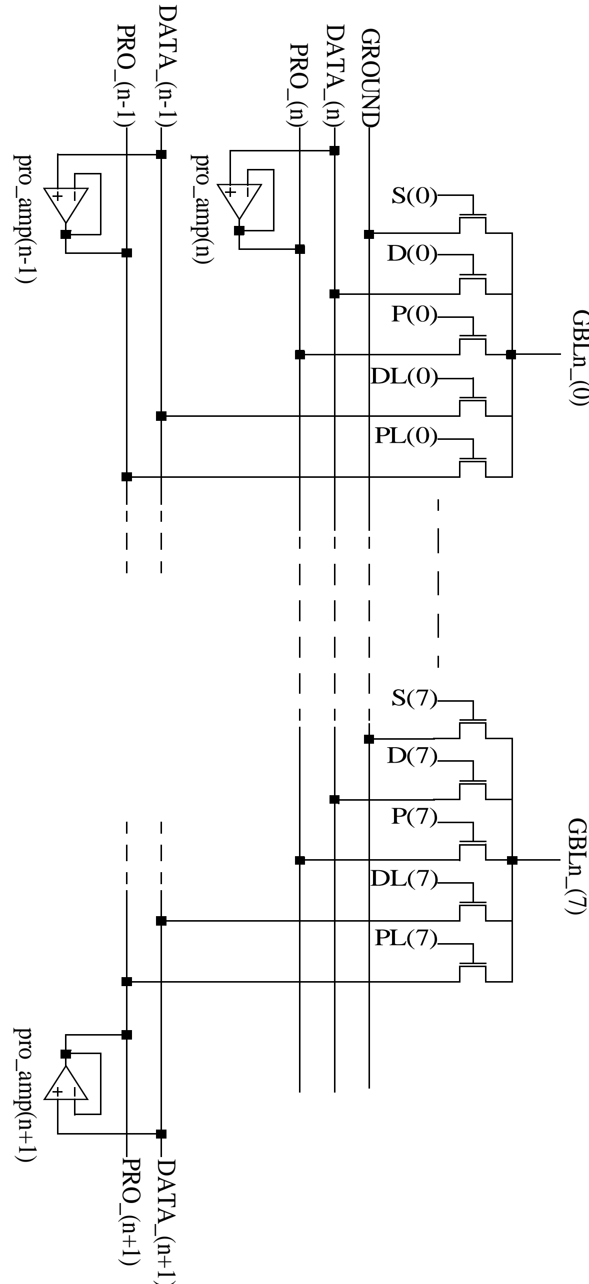


Figure 3.8: Global bitline decoding for GBLn_(0) and GBLn_(7)

Similarly, another pass transistor with the gate connected to PL(0) is used to bring the voltage of the line PRO_(2) to the line GBL3_(0) in case a bit near the right edge of the data block 2 is read. This is because the voltage at the DATA_(2) line must be tracked by the voltage at the PRO_(2) line as required by the Sense Current Recovery Technique (discussed in Chapter 2), so if DATA_(2) sends the drain voltage to GBL3_(0), then PRO_(2), not PRO_(3), should send its voltage to GBL3_(1) to minimize the side-leakage current. The transistors with the gate connected to the control signals like DL(0), PL(0), etc., are called the *interface transistors* because they are used when a bit near the interface of the two adjacent data blocks is accessed.

A similar situation occurs when a bit near the left edge of a data block is read, where the global bitlines 7, 6,... of the previous data block are borrowed; interface transistors with the gate connected to DR(7) and PR(7) in Figure 3.8 are used for this case. The complete interface architecture and interface transistors for a data block is shown in Figure 3.9. The interface control signals are DL(1:0), PL(4:0), DR(7:5) and PR(7:2). The interface control signals as well as the regular control signals S(7:0), D(7:0) and P(7:0) are high-voltage signals, level-shifted directly from the low-voltage signals S_LV(7:0), D_LV(7:0), P_LV(7:0), DL_LV(1:0), DR_LV(7:5), PL_LV(4:0) and PR_LV(7:2). The level shifters used for these control lines are similar to the one shown in Figure 3.5.

3.2.4 Source-Drain-Protecting logic block

This logic block generates the low-voltage control signals S_LV(7:0), D_LV(7:0), P_LV(7:0), DL_LV(1:0), DR_LV(7:5), PL_LV(4:0) and PR_LV(7:2). A similar design methodology as the one for designing the sector-y logic block is applied here, in which truth tables are constructed, Verilog code is written based on the tables and finally a logic synthesis tool is used to synthesize the logic block. Based on Figures 3.6, 3.7, 3.8 and 3.9, the Truth Tables 3.2, 3.3, 3.4 are constructed for the source, drain and protecting bitline

control signals, respectively. The Verilog code based on these tables are listed in the Appendix.

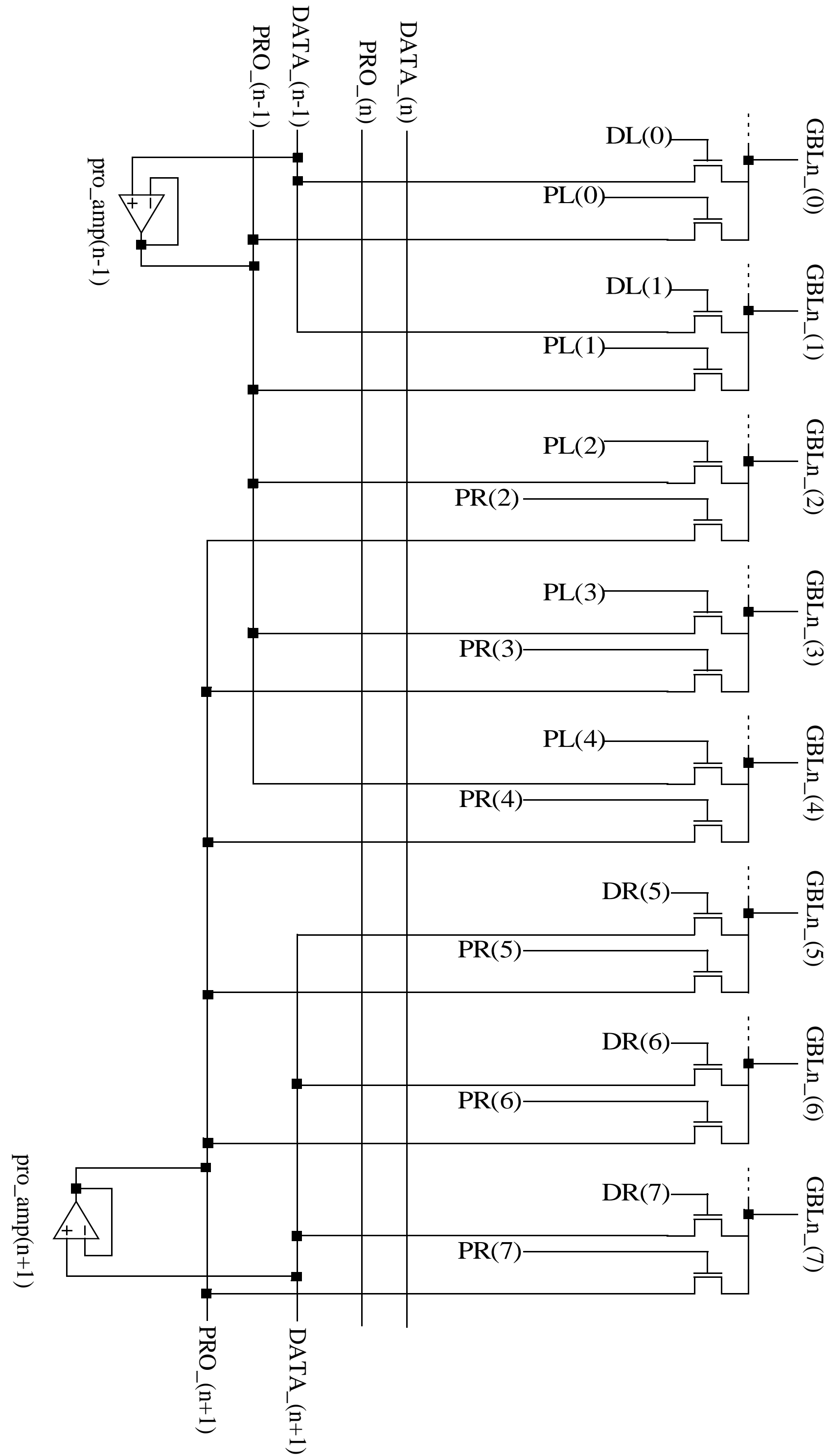


Figure 3.9: Global bitline decoding for data block interface

Table 3.2: Truth table for the source-drain-protecting logic block - Source decoding

Bit #	A(6:2)	S_LV(0)	S_LV(1)	S_LV(2)	S_LV(3)	S_LV(4)	S_LV(5)	S_LV(6)	S_LV(7)	S_LV(7:0) (Hex)
0,8	0x000							1	1	C0
1,9	0x001	1							1	81
2,10	0x010	1	1							03
3,11	0x011		1	1						06
4,12	0x100			1	1					0C
5,13	0x101				1	1				18
6,14	0x110					1	1			30
7,15	0x111						1	1		60
16,24	1x000	1	1							03
17,25	1x001		1	1						06
18,26	1x010			1	1					0C
19,27	1x011				1	1				18
20,28	1x100					1	1			30
21,29	1x101						1	1		60
22,30	1x110							1	1	C0
23,31	1x111	1							1	81

3.2.5 Simulation results and sizing for the column decoding

Figures 3.10 and 3.11 present the logic simulation results for the sector-y logic block and the source-drain-protecting logic block, respectively. These two block have a common enable signal EN. When EN is low, all the control signals SECY_LV(15:0), S_LV(7:0), D_LV(7:0), DL_LV(1:0), DR_LV(7:5), P_LV(7:0), PL_LV(4:0) and PR_LV(7:2) are grounded. The simulation results agree exactly with the Truth Tables 3.1, 3.2, 3.3 and 3.4.

HSPICE simulations have been performed to determine the best size for the pass transistors. The size for the sector select transistors (Figure 3.6) is W/L = 5.77/0.58, and the size for the pass transistors in the global bitline decoding block (Figure 3.8) is 15/0.58. The trade-off considered in sizing is between the area of the decoding and the read speed.

Table 3.3: Truth table for the source-drain-protecting logic block - Drain decoding

Bit #	**A(6:2)**	**D_LV(0)**	**D_LV(1)**	**D_LV(2)**	**D_LV(3)**	**D_LV(4)**	**D_LV(5)**	**D_LV(6)**	**D_LV(7)**	**DL_LV(0)**	**DL_LV(1)**	**DR_LV(5)**	**DR_LV(6)**	**DR_LV(7)**	**{DR_LV, DL_LV, D_LV} (12:0) (Hex)**
0	00000	1	1	1											0007
1	00001		1	1	1										000E
2	00010			1	1	1									001C
3	00011				1	1	1								0038
4	00100					1	1	1							0070
5	00101						1	1	1						00E0
6	00110	1						1	1						00C1
7	00111	1	1						1						0083
8	01000	1	1	1											0007
9	01001		1	1	1										000E
10	01010			1	1	1									001C
11	01011				1	1	1								0038
12	01100					1	1	1							0070
13	01101						1	1	1						00E0
14	01110							1	1	1					01C0
15	01111								1	1	1				0380
16	10000											1	1	1	1C00
17	10001	1											1	1	1801
18	10010	1	1											1	1003
19	10011	1	1	1											0007
20	10100		1	1	1										000E
21	10101			1	1	1									001C
22	10110				1	1	1								0038
23	10111					1	1	1							0070
24	11000						1	1	1						00E0
25	11001	1						1	1						00C1
26	11010	1	1						1						0083
27	11011	1	1	1											0007
28	11100		1	1	1										000E
29	11101			1	1	1									001C
30	11110				1	1	1								0038
31	11111					1	1	1							0070

Table 3.4: Truth table for the source-drain-protecting logic block - Protecting decoding

Bit #	A(6:2)	P_LV(0)	P_LV(1)	P_LV(2)	P_LV(3)	P_LV(4)	P_LV(5)	P_LV(6)	P_LV(7)	PL_LV(0)	PL_LV(1)	PL_LV(2)	PL_LV(3)	PL_LV(4)	PR_LV(2)	PR_LV(3)	PR_LV(4)	PR_LV(5)	PR_LV(6)	PR_LV(7)	{PR_LV, PL_LV, P_LV} (18:0) (Hex)
0	00000				1	1	1														00038
1	00001					1	1	1													00070
2	00010						1	1	1												000E0
3	00011	1						1	1												000C1
4	00100	1	1						1												00083
5	00101	1	1	1																	00007
6	00110		1	1	1																0000E
7	00111			1	1	1															0001C
8	01000				1	1	1														00038
9	01001					1	1	1													00070
10	01010						1	1	1												000E0
11	01011							1	1	1											001C0
12	01100								1	1	1										00380
13	01101									1	1	1									00700
14	01110										1	1	1								00E00
15	01111											1	1	1							01C00
16	10000														1	1	1				0E000
17	10001															1	1	1			1C000
18	10010																1	1	1		38000
19	10011																	1	1	1	70000
20	10100	1																	1	1	60001
21	10101	1	1																	1	40003
22	10110	1	1	1																	00007
23	10111		1	1	1																0000E
24	11000			1	1	1															0001C
25	11001				1	1	1														00038
26	11010					1	1	1													00070
27	11011						1	1	1												000E0
28	11100	1						1	1												000C1
29	11101	1	1						1												00083
30	11110	1	1	1																	00007
31	11111		1	1	1																0000E

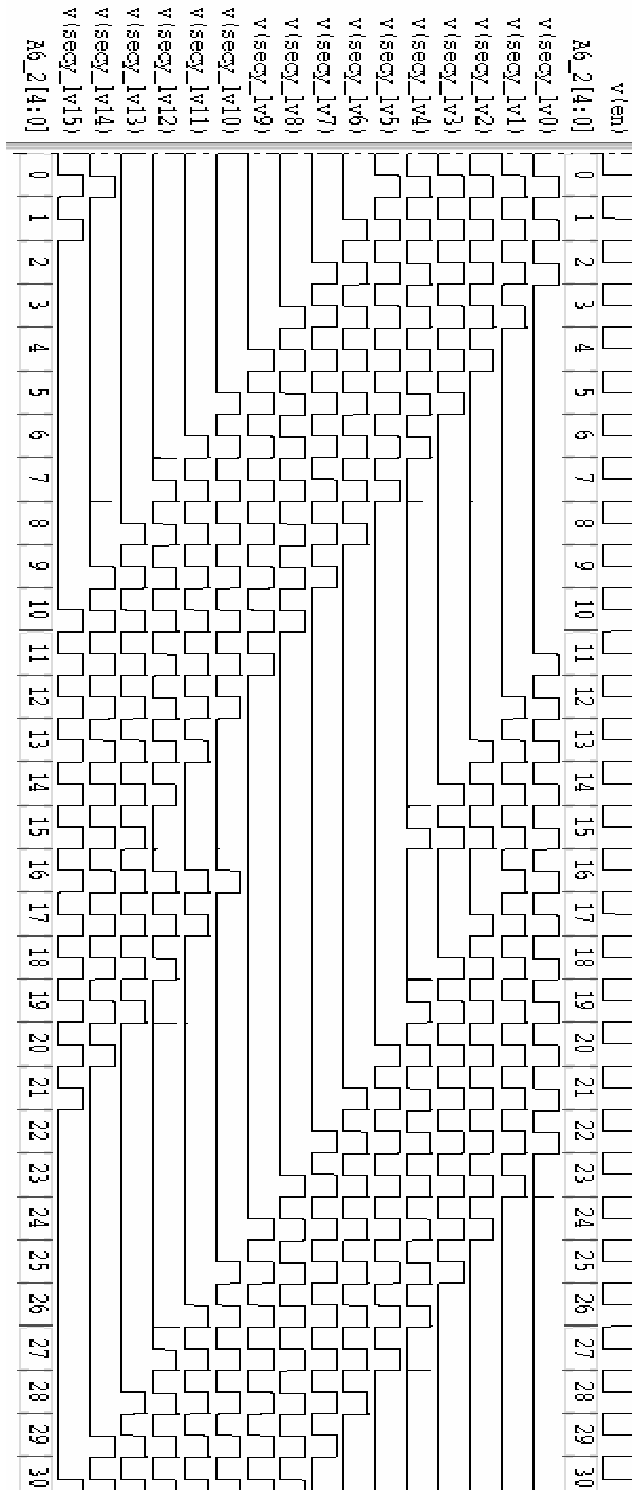


Figure 3.10: Logic simulation for the sector-y logic block

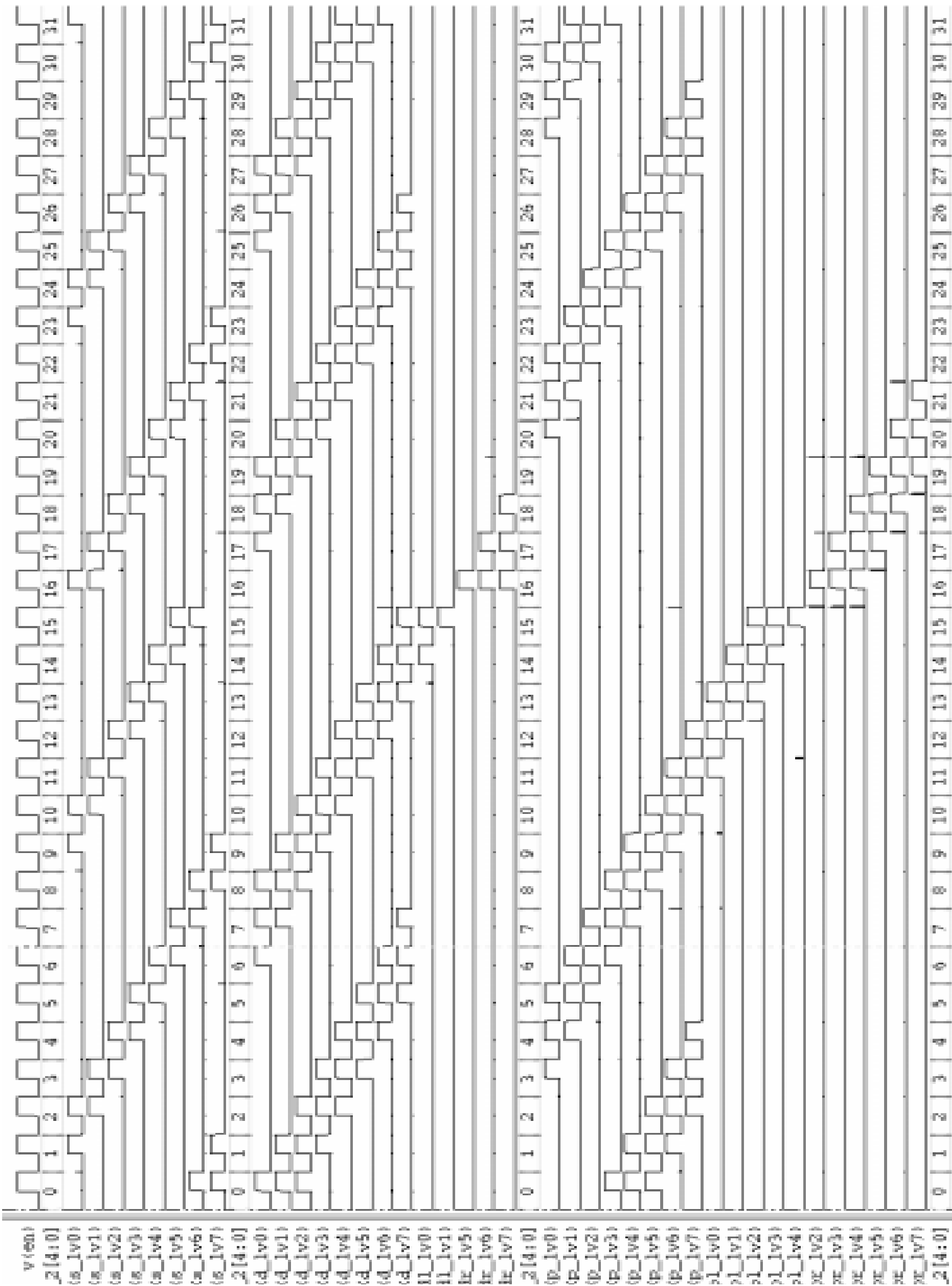


Figure 3.11: Logic simulation for the source-drain-protecting logic block

3.2.6 Bitline decoding for the edges

To avoid the non-uniform memory array due to edge effects in lithography, there are about 10 dummy cells for each wordline at each edge of a bank. Thus, the column decoding is extended into the left edge and the right edge so that when a bit is accessed near the left edge of data block 0 or near the right edge of data block 63, the decoding pattern S-S-D-D-D-P-P-P is maintained. Figures 3.10 and 3.11 shows the column decoding architecture for the right edge and the left edge, respectively.

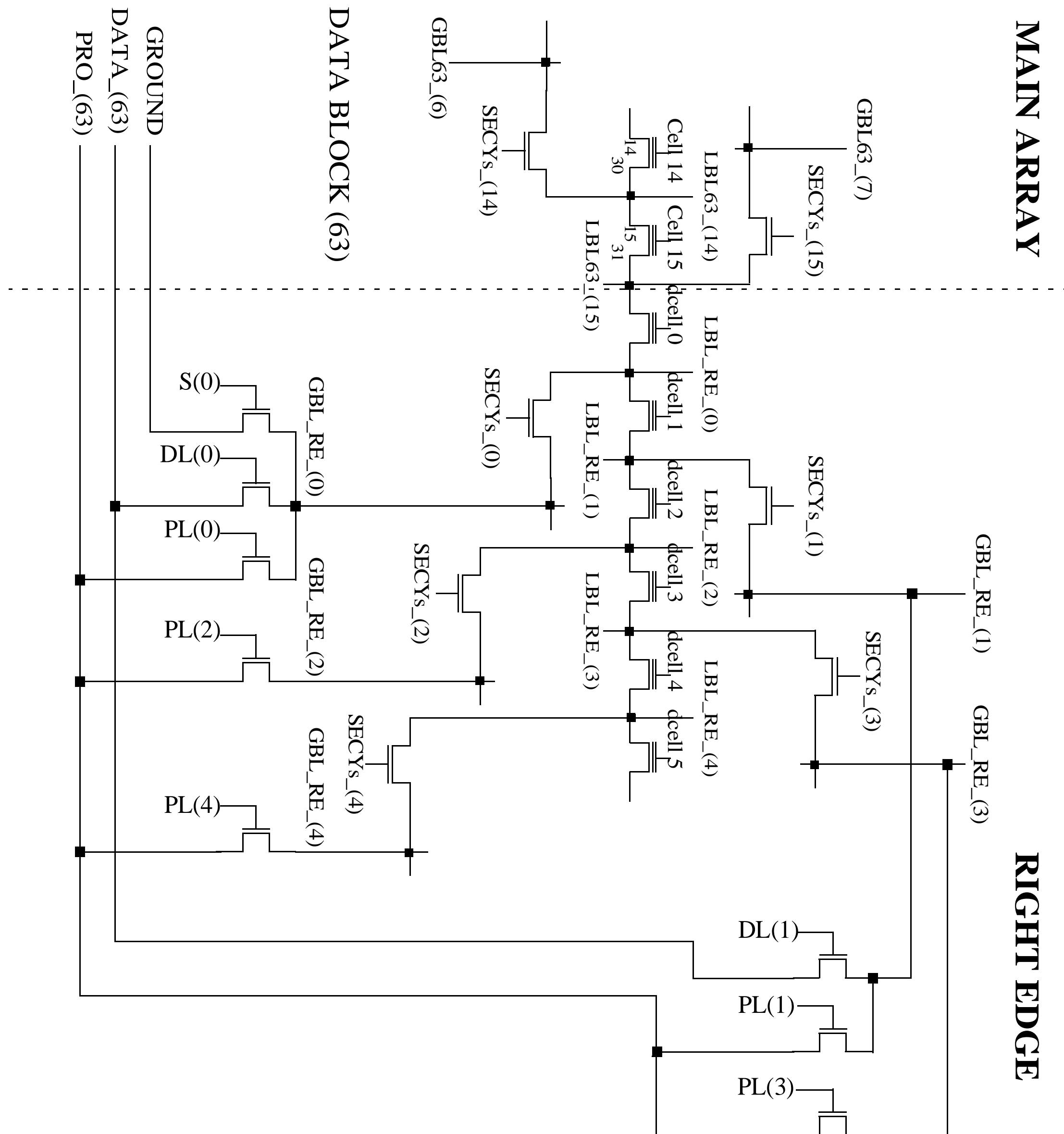


Figure 3.12: Column decoding for the right edge

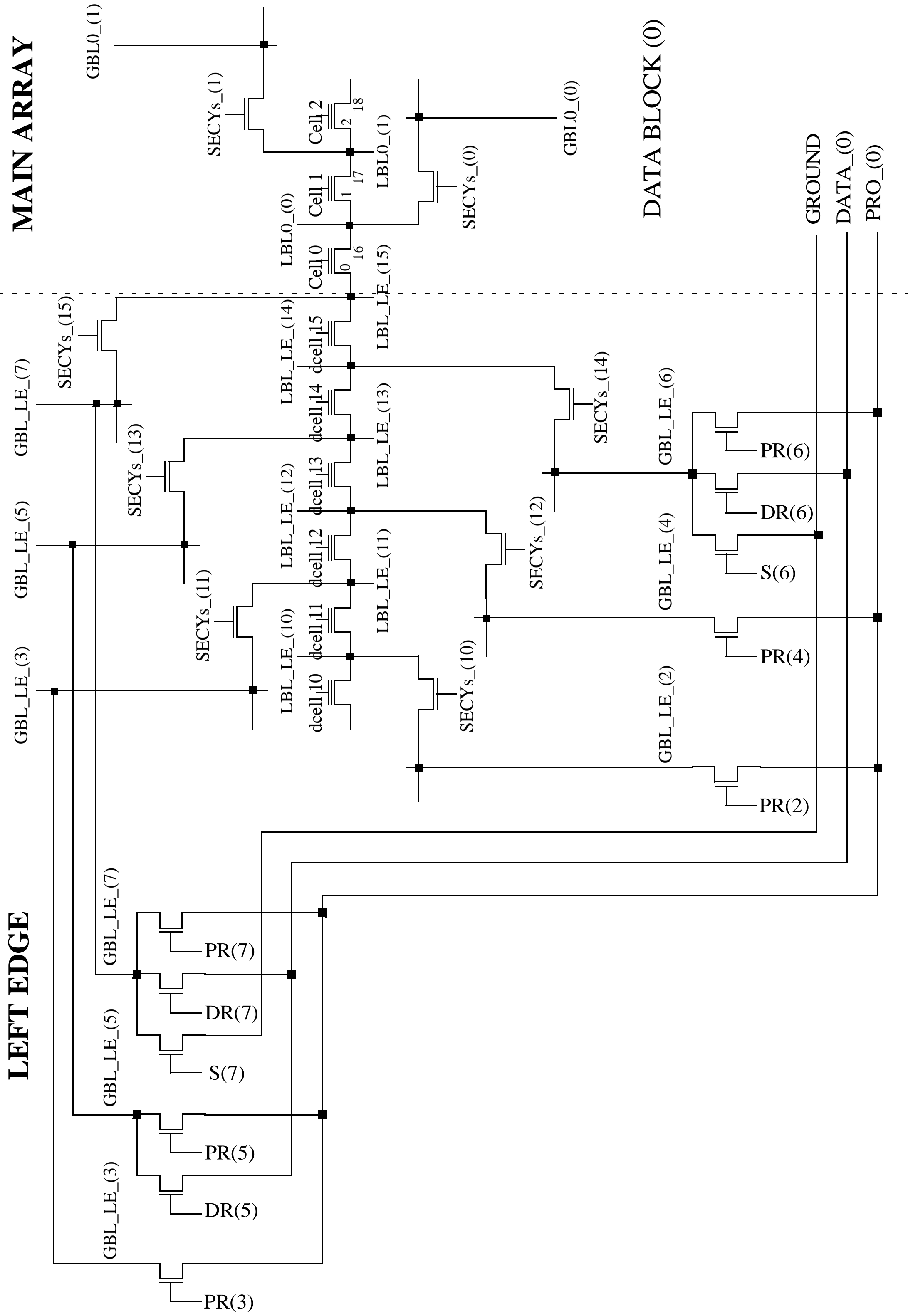


Figure 3.13: Column decoding for the left edge

3.3 Summary

In this chapter the multiple-drain-bitline and multiple-protecting bitline column decoding used to support the Sense Current Recovery Technique has been described. The chapter not only shows that the decoding style S-S-D-D-D-P-P-P is possible, but also that it can be developed rapidly with the use of logic synthesis tools. The area overhead is minimum because even though each sector needs 16 SECYs lines, these lines are routing on top of the sector select transistors, consuming no extra area. In addition, the area of the logic blocks are well optimized by the synthesis tool. The only area penalty is for the extra number of level shifters, but this area is small compared to the chip area. In fact, the decoding style S-D-D-P-P can be designed to reduce the number of level shifters, but the read margin loss is more compared to the decoding style S-S-D-D-D-P-P-P as discussed in Chapter 2.

Chapter 4

Differential feedback cascoded bitline voltage control

This chapter focuses on how to reduce the read margin loss caused by the disturbance of the other bit in the same memory cell, which is also called the Complementary-Bit Disturbance (CBD effect). As discussed in Chapter 1, the CBD is less if the memory cell is operated deep in the saturation region. This unique characteristic of the two-bit-per-cell nitride-storage flash memory requires the bitline voltage to be as high and stable as possible. These requirements are the main focus of the differential feedback cascoded bitline voltage control technique, which involves the use of a differential amplifier in a feedback loop to stabilize the bitline voltage at a flexibly-chosen level. Section 4.1 briefly analyzes the traditional cascode amplifier to show that it is difficult for this amplifier to meet the bitline control requirements stated above. Section 4.2 describes the differential feedback cascoded bitline voltage control technique which is capable of solving the CBD problem as well as providing fast read speed. Section 4.3 shows the simulation result for the differential feedback cascode amplifier. Section 4.4 summarizes the chapter.

4.1 Analysis of the traditional cascode amplifier

The cascode amplifier if often used to provide a regulated bitline voltage for the memory cells in the read mode. A traditional cascode amplifier is shown in Figure 1.10, and is placed again in Figure 4.1 for convenience.

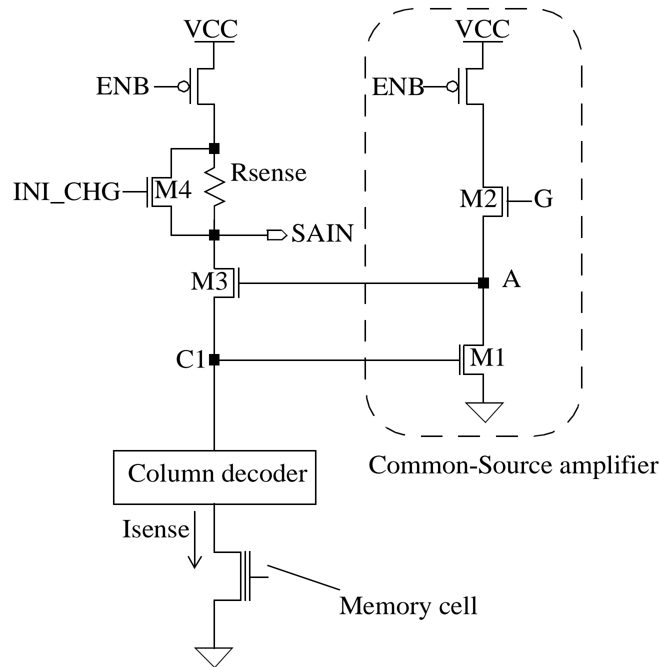


Figure 4.1: Traditional cascode amplifier

The cascode amplifier in Figure 4.1 is also used to convert the memory cell current Isense to a voltage (SAIN) for sensing, using the sense resistor Rsense. In a read operation, ENB is switched to low to turn on the cascode amplifier. Assume that all transistors are operated in the saturation region and ignore the body effect as well as the channel length modulation, then the following equations can be written directly from Figure 4.1:

$$k_3(V_A - V_{C1} - V_{T3})^2 = Isense \tag{4.1}$$

$$k_1(V_{C1} - V_{T1})^2 = k_2(V_G - V_A - V_{T2})^2 \tag{4.2}$$

Where $k_i = \frac{1}{2}\mu_n C_{ox}\left(\frac{W_i}{L_i}\right)$ and V_{Ti} (threshold voltage) are for transistor Mi, i = 1,2,3.

The following equations are obtained by solving for V_A from equations (4.1) and (4.2):

$$V_A = V_{C1} + V_{T3} - \sqrt{\frac{Isense}{k_3}} \tag{4.3}$$

$$V_A = V_G - V_{T2} - \sqrt{\frac{k_1}{k_2}}(V_{C1} - V_{T1}) \tag{4.4}$$

The bitline voltage V_{C1} can be can be calculated by solving (4.3) and (4.4) concurrently:

$$V_{C1} = \frac{V_G - V_{T2} - V_{T3} + V_{T1}\sqrt{\frac{k_1}{k_2}} - \sqrt{\frac{Isense}{k_3}}}{\left(1 + \sqrt{\frac{k_1}{k_2}}\right)} \tag{4.5}$$

The right branch of the cascode amplifier in Figure 4.1 is just a Common Source amplifier, which consists of transistors M1 and M2, and the gain of this stage [21] can be calculated as $A_v = -\sqrt{\frac{k_1}{k_2}}$, thus:

$$V_{C1} = \frac{V_G - V_{T2} - V_{T3} + |A_v| V_{T1} - \sqrt{\frac{Isense}{k_3}}}{(1 + |A_v|)} \tag{4.6}$$

The formula (4.6) shows that the bitline voltage V_{C1} varies greatly because it depends on many process parameters such as the threshold voltages of transistors M1, M2 and M3, which are not well-controlled. Even worst, these transistors are not of the same type, so in the worst case their threshold voltages can change in opposite directions due to process

variations. To reduce the variation of the bitline voltage V_{C1}, the gain A_v must be increased, but doing so will force the output common mode level (voltage at node A) to go down (because transistor M3 must be made weaker), and thus the bitline voltage V_{C1} will drop. This conflicts with the requirement that the bitline voltage must be raised as high as possible. The bitline voltage V_{C1} drops even more when the power supply voltage VCC is at the lowest level of 1.6V. This drawback prevents the traditional cascode from being used in the 1.8V, two-bit-per-cell, nitride-storage flash memory. Thus a new bitline voltage control technique called the differential feedback cascoded bitline voltage control technique is developed to overcome this problem; it also offers other advantages. This new technique is the topic of the next section.

4.2 Differential feedback cascoded bitline voltage control technique

The differential feedback cascoded bitline voltage control technique is designed to meet the two most important requirements for the 1.8V, two-bit-per-cell, nitride-storage technology: the bitline voltage must be as high as possible to reduce the CBD effect, and this voltage must be as stable as possible with respect to process variation, temperature and memory cell current. The key element of the technique is the differential feedback cascode amplifier. In Section 4.2.1, a simplified version of this new cascode amplifier is presented to show that it satifies the two most important requirements mentioned above. Section 4.2.2 shows the final version of the differential feedback cascode amplifier in which additional advantages of the new cascode will be discussed.

4.2.1 Differential feedback cascode amplifier - A simplified version

The simplified version of the differential feedback cascode amplifier is shown in Figure 4.2. Like the traditional cascode amplifier, this new cascode still uses the resistor Rsense (or transistors which act as resistors) to convert the memory cell current Isense to a voltage at node SAIN for sensing. However, the main features of the new cascode amplifier are to raise the bitline voltage as high as possible and to stabilize the bitline voltage as much as possible -- requirements that the traditional cascode amplifier fails to meet. The new cascode amplifier solves this problem by using a differential amplifier in which the bitline voltage level V_{C1} is not determined by the process parameters such as the threshold voltage, but by a reference voltage CASREF, which is assumed for now to be easy to raised. Thus the first requirement of raising the bitline voltage as high as possible is met. Note that the transconductance of the input transistors of the differential amplifier is basically not affected by the input common mode level, a characteristic the Common Source amplifier does not have, thus V_{C1} can be raised without affecting the bias level of node A.

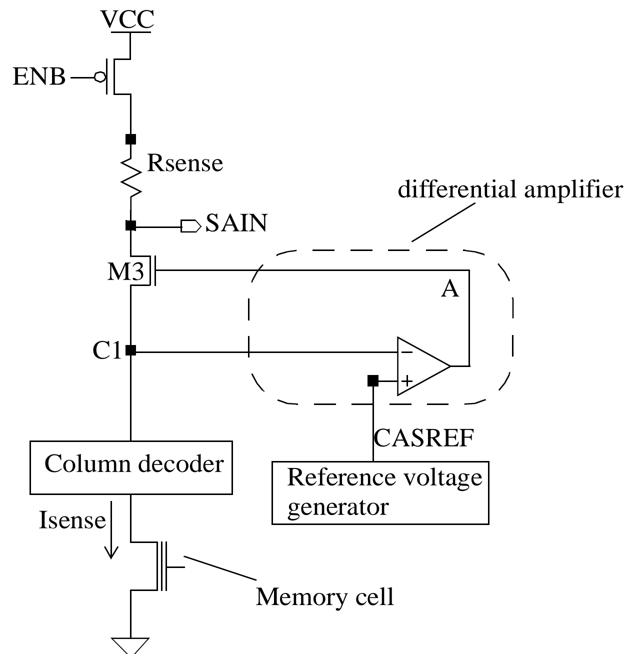


Figure 4.2: Differential feedback cascode amplifier - A simplified version

Now assume that CASREF is also stable despite the variations in process, temperature, memory cell current and power supply voltage, then the second requirement of stabilizing the bitline voltage is also met. The next two sections will describe further the differential amplifier and the reference voltage generator shown in Figure 4.2.

4.2.1.1 The differential amplifier

The differential amplifier use in the new cascode amplifier can be a traditional differential pair of any type, for example the differential pair with a current mirror load as shown in Figure 4.3a, or with a resistive load as in Figure 4.3b. The differential amplifier can have one stage or more stages to increase the gain, depending on the accuracy and speed requirements for each particular flash memory.

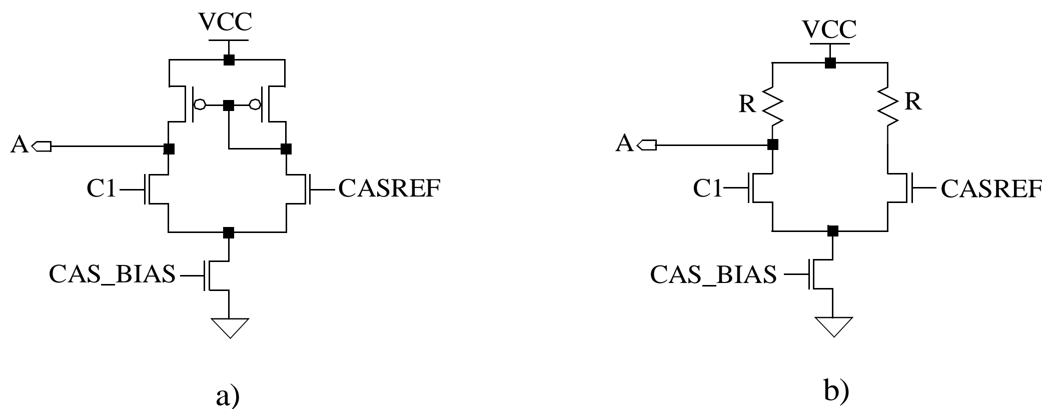


Figure 4.3: The differential amplifier a) with current mirror load b) with resistive load

However, because the differential amplifier is used in a feedback loop (going from node C1 to the input of the differential amplifier, to node A, to transistor M3 and back to node C1 in Figure 4.2), the number of stages should be limited to one or two to have a good transient response. Also, at low power supply voltage of 1.8V or less, there may not be

enough voltage headroom to build the current mirror load or some kind of transistor load, thus the differential amplifier with a resistive load as shown in Figure 4.3b is preferred. Of course, for 3V-applications, there is no such constraint and the current mirror load is preferred due to its high gain.

4.2.1.2 The reference voltage generator

The reference voltage generator can be built from a traditional bandgap reference circuit as shown in Figure 4.4. A weak resistive divider is used to divide the VREF level to the desired level, which is then buffered by a unity feedback amplifier to generate CASREF.

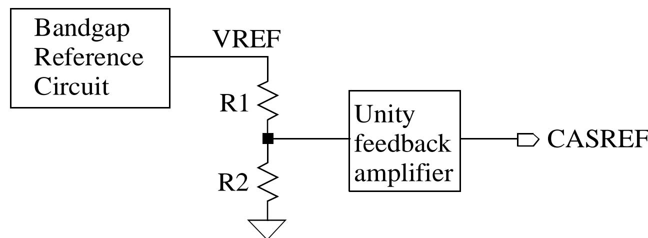


Figure 4.4: Bandgap reference voltage generator

When the bandgap reference voltage generator is used, CASREF is not only independent of process variation, temperature, memory cell current but also independent of the supply voltage VCC. However, there are many drawbacks associated with this circuit. First of all, the speed of the combination of the bandgap reference circuit, the weak resistor divider and the unity feedback buffer is slow. Even if the fast bandgap reference shown in Figure 5.5 (Chapter 5) is used, the speed can be as slow as 30ns. Secondly, the bandgap reference circuit is big (because it uses bipolar transistors) and complicated. Finally, the fact that the bandgap reference circuit makes CASREF (and thus the bitline voltage) independent of the supply voltage is unnecessary. The bitline voltage should be stable as much as possible with process variation, temperature and memory cell current, but it need not be independent of the supply voltage, as long as it varies in the same direction with the

supply voltage. Recall that if the bitline voltage increases, it helps to reduce the CBD effect, therefore there would be no problem if CASREF increases with the supply voltage VCC. Thus, CASREF is designed to be at least 1.15V when the supply voltage is at the lowest level of 1.6V, to make sure that the memory cell is in the saturation region.

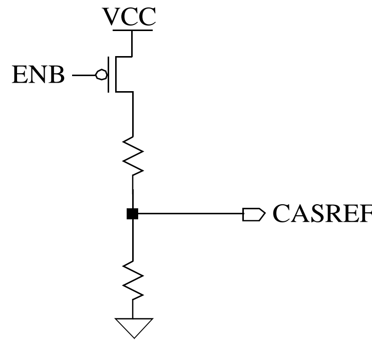


Figure 4.5: Resistive reference voltage generator

The discussion above leads to the design of a very simple, yet very effective, reference voltage generator shown in Figure 4.5. It is just a resistive divider, which is very fast, very simple, independent of process, temperature and varies in the same direction with the power supply voltage VCC. This reference voltage generator is the one actually built for the 1.8V, two-bit-per-cell flash memory described in this dissertation.

4.2.1.3 Power consumption of the new cascode amplifier.

In a read operation, 64 bits are read out at the same time, thus 64 cascode amplifiers must be activated concurently. To have a low-power flash memory, the maximum current consumption for the cascode amplifier must be small. Besides the advantages discussed in the previous sections, another advantage of the differential feedback cascode amplifier is that its maximum current consumption is much less than that of the traditional cascode

amplifier. This is because the bias current of the Common Source amplifier in the traditional cascode amplifier varies significantly with process and the power supply voltage, while the bias current of the differential amplifier in the new cascode stage is from a better controlled reference current. The reference current circuit for the new cascode is shown in Figure 4.6.

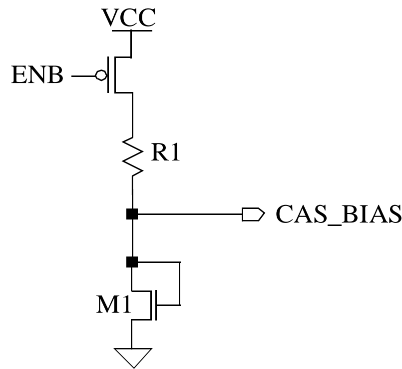


Figure 4.6: Current reference for the new cascode amplifier

The variation of the reference current in Figure 4.5 is less because the resistor sheet resistance is usually controlled well enough. Secondly, the resistor R1 is an unsilicided polysilicon resistor, which has a negative temperature coefficient, thus at lower temperatures, when the transistor M1 is stronger, the resistor R1 is weaker (more resistive), compensating somewhat for the decreasing in the equivalent resistance of transistor M1.

4.2.2 Differential feedback cascode amplifier - the complete version

The full version of the differential feedback cascode amplifier is shown in Figure 4.7. The focus of this section, however, is on other aspects of the new cascode amplifier that help it

to achieve a fast sensing time. These aspects are the elimination of the disturbance to the sensing node and the elimination of the bitline voltage overshoot.

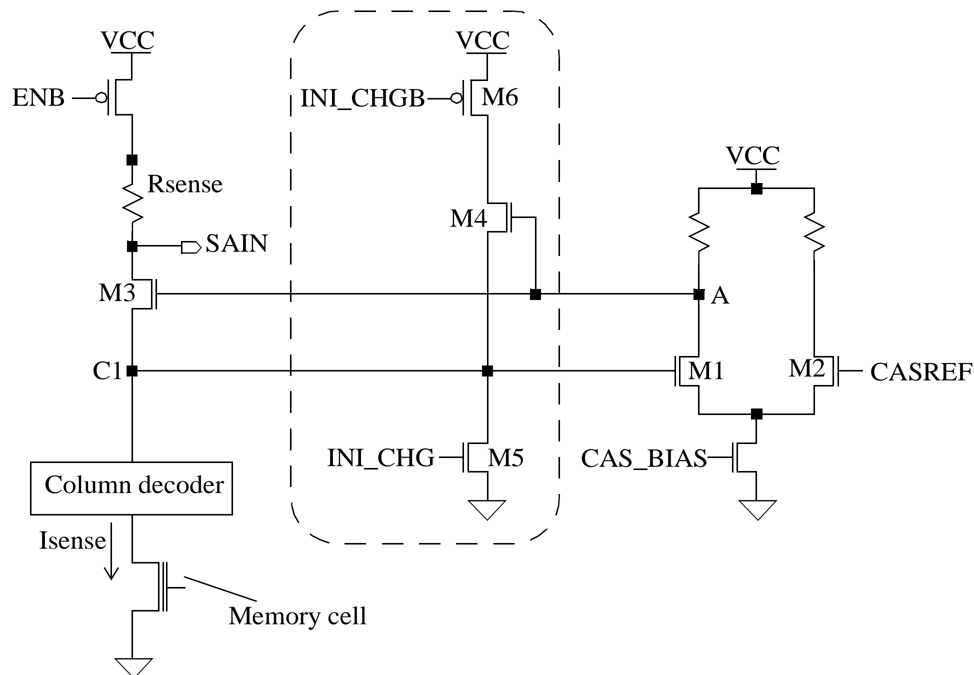


Figure 4.7: The full version of the differential feedback cascode amplifier

4.2.2.1 Eliminating the disturbance to the sensing node

To have a short read access time, the bitline of the selected memory cell needs to be pulled up to the target voltage quickly. The capacitance of the bitline, however, is very significant (~0.5p) and thus it takes time to bring the bitline up. In Figure 4.1, if transistor M4 did not exist, then the current used to charge the bitline, which is coupled to node C1, would have to run through the very resistive resistor Rsense, thus the bitline would be brought up very slowly. To fix this problem, transistor M4 is used to short out the resistor Rsense in about 10ns to 15ns, which is the pulse width of the signal INI_CHG. When the bitline voltage reaches the target value, INI_CHG is deasserted, and the cascode amplifier returns to its sensing mode, meaning that the Rsense is not shorted out, but is used to

develop the voltage SAIN. Transistor M4, however, introduces large disturbance to the sensing node SAIN, causing slow read access because it takes longer for SAIN to settle to the final level. Moreover, the control signal level at the gate of M4 (INI_CHG) is just the supply voltage VCC, which varies significantly (1.6V to 2.0V), and the threshold voltage of M4 also varies with process, therefore the disturbance at node SAIN caused by M4 is difficult to compensate; correctly compensating at a particular VCC and process may not work at different VCC and different process corner.

The differential feedback cascode amplifier avoids the disturbance at node SAIN by using another path to charge up the bitline. This path is through transistors M6 and M4 in Figure 4.7. The path is activated when INI_CHG is high (or INI_CHGB is low). This additional charging path is not connected to the sensing node SAIN, thus it does not cause any disturbance to this node, leading to faster read access. A minor detail that needs to be mentioned is that for low supply voltages, due to the voltage headroom constraint, transistors M3 and M4 in Figure 4.7 have to be intrinsic transistors, which have threshold voltages close to 0V.

4.2.2.2 Eliminating the bitline voltage overshoot

As mentioned in the previous section, there is a strong charging path for the bitline through transistors M6 and M4. When the differential feedback cascode amplifier is activated, this path can overcharge the bitline to some extent, although the gate of transistor M4 is indirectly controlled by CASREF through the feedback loop. This happens because the differential amplifier needs some time to reponse to a sudden event such as turning on of the cascode. The bitline voltage overshoot is severe because it takes considerable time to settle down, affecting directly the sensing operation. As shown in Figure 4.8, the overshoot comes down slowly to the final voltage of about 1.25V. This is because the cascode amplifier is good at pulling the bitline up, not pulling it down, thus

when the bitline overshoots, the only current that can bring down the bitline is the memory cell current Isense, which is usually small.

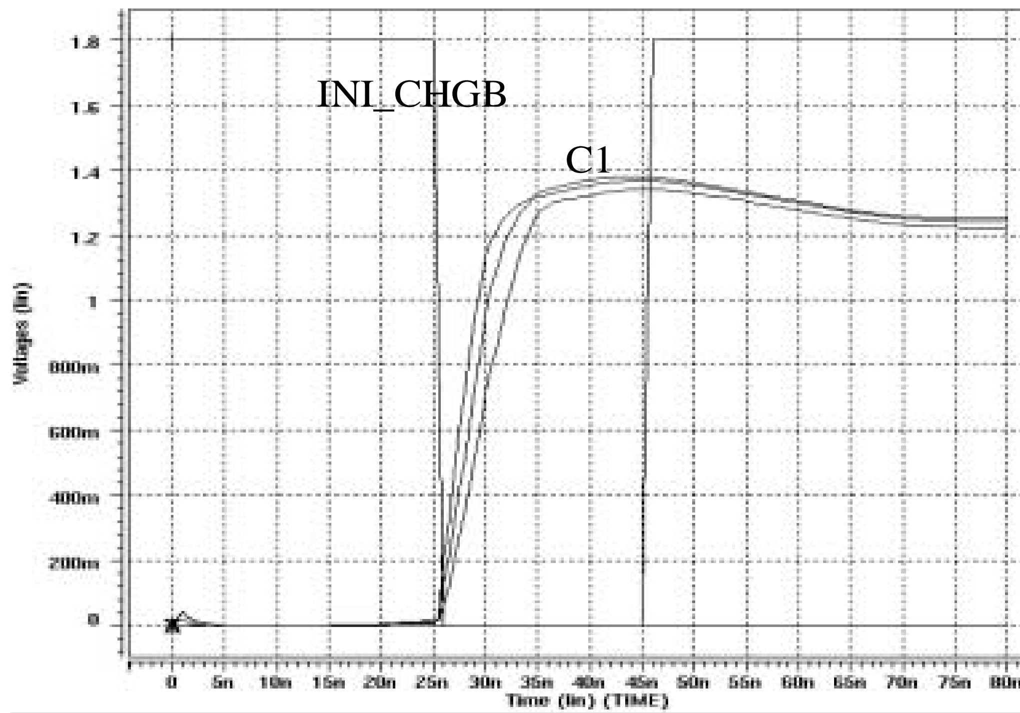


Figure 4.8: Bitline voltage overshoot

The new differential feedback cascode amplifier solves this problem by using a weak transistor M5 shown in Figure 4.7, its gate connected to the INI_CHG signal. When the new cascode amplifier is activated, INI_CHG switches to high and the charging path through transistor M4 starts to pull node C1 up. At the same time, the weak transistor M5 is also turned on to prevent the voltage overshoot at node C1. The size of transistor M5 can be chosen appropriately to eliminate the voltage overshoot. M5 acts as a clamping device, temporarily holding node C1 when INI_CHG is high. The threshold voltage of M5, however, varies significantly with process, thus at some process corners the clamping action may be too strong, holding C1 at lower voltages. Therefore, it takes more time for C1 to climb to the final value when M5 is turned off, prolonging the read access time. To fix this problem, M5 and M4 are chosen to be of the same transistor type, forming a good “resistive” divider, which is insensitive to the process variation.

4.3 HSPICE simulation result for the differential feedback cascode amplifier

The simulation results for the traditional cascode amplifier are shown in Figure 4.9 for comparison with the results for the differential feedback cascode amplifier shown in Figure 4.10. The simulations are conducted with VCC=1.8V, 25C, but *with all process corners*.

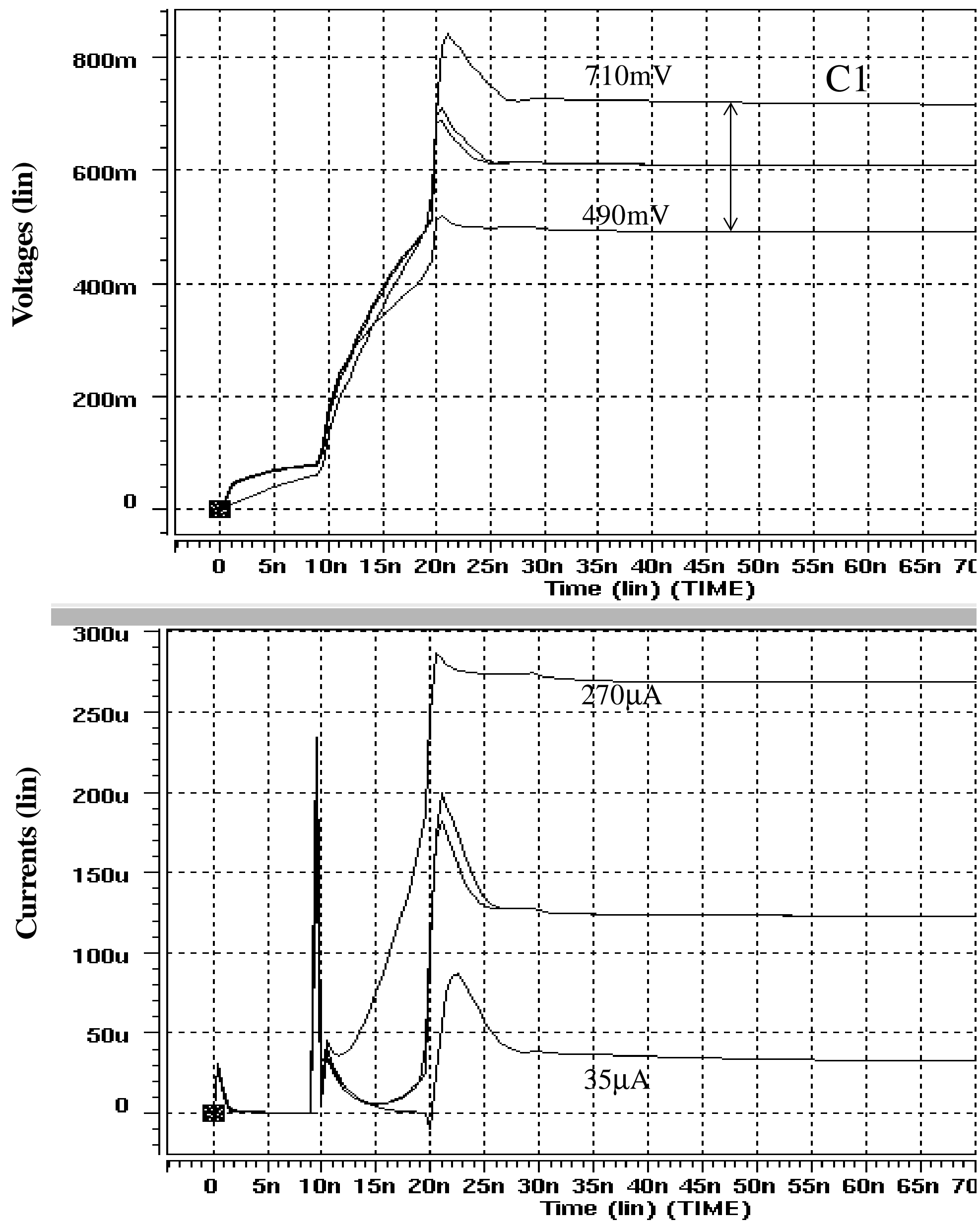


Figure 4.9: Simulation result for the traditional cascode amplifier

The simulation result in the top panel of Figure 4.10 shows that the differential feedback cascode amplifier satifies the two important requirements. First of all, the bitline voltage (voltage at node C1) can be raised easily to 1.25V, while for the traditional cascode amplifier, it is difficult to raise the bitline voltage to even 0.9V; in Figure 4.9, this bitline voltage is about 0.6V.

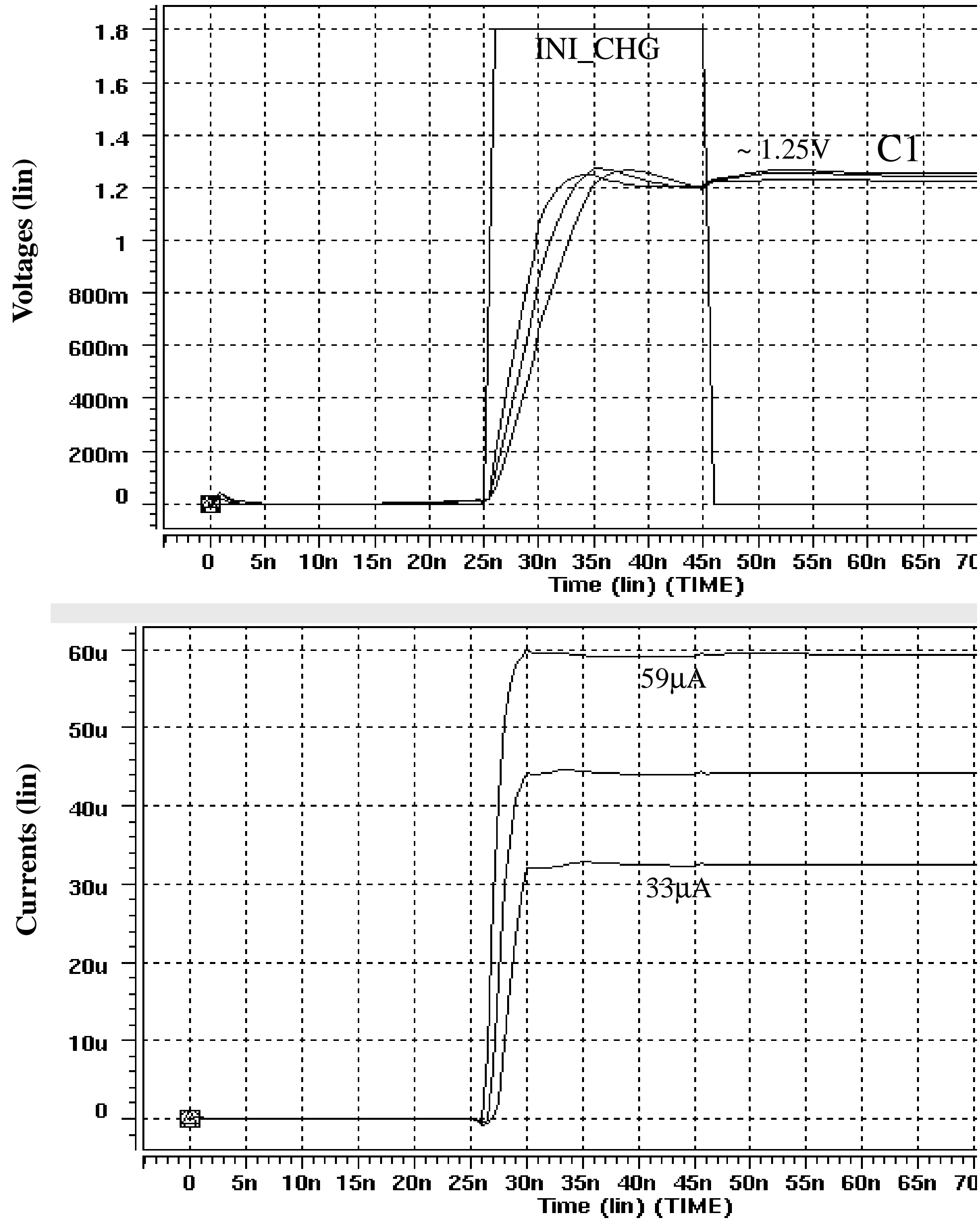


Figure 4.10: Simulation result for the differential feedback cascode amplifier

Secondly, the new cascode amplifier stabilizes the bitline voltage very well despite the variation in process; the top panel of Figure 4.10 shows that the bitline voltage varies only about 20mV across all process corners, while for the traditional cascode amplifier this variation is about 210mV, as shown in the top panel of Figure 4.9. Also, the simulation result in the top panel of Figure 4.10 shows that there is no overshoot in the bitline voltage. C1 is clamped at an appropriate voltage when INI_CHG is high and is released for it to go to the final level when INI_CHG is deactivated. Note that the pulse width of INI_CHG needs not to be 20ns as shown in the figure; it can be only 10ns or 15ns.

The bottom panel of Figure 4.10 confirms that the maximum current consumption of the new cascode amplifier is much smaller than the tradition cascode amplifier. This maximum current is only 59µA, while the maximum current of the traditional cascode is 270µA, as shown in the bottom panel of Figure 4.9.

4.4 Summary

The differential feedback cascoded bitline voltage control technique discussed in this chapter is very effective in reducing the read margin loss caused by the Complementary-Bit disturbance (CBD effect), which is the disturbance from the other bit in the same memory cell. The differential feedback cascode amplifier, which is the core of the differential feedback cascoded bitline voltage control technique, not only fulfills the two main requirements of raising the bitline voltage as high as possible and stabilizing this voltage as much as possible, but also offers many other advantages such as fast speed, no disturbance at the sensing node and much less current consumption. Controlling well the current consumption for the new cascode amplifier and the protecting amplifiers (discussed in Chapter 2) makes this 1.8V, two-bit-per-cell flash memory a low power

device. The maximum bias current for 64 new cascode amplifiers, 64 protecting amplifiers and 64 sense amplifiers activated in a read access is only 14mA.

Finally, while the traditional cascode amplifier can not be used for the 1.8V, two-bit-per-cell, nitride-storage flash memory, the new cascode can be used for any type of flash memory. The differential feedback cascode amplifier is versatile and can be applied for different memory technologies with different power supply voltages.

Chapter 5

Auto-calibrated wordline voltage control

This chapter describes the design and analysis of the auto-calibrated wordline booster. The new booster, also called the A/D wordline booster, is used to generate quickly and accurately the boosted wordline voltage during a fast read access. As mention in Chapter 1, controlling the wordline voltage accurately is essential in minimizing the read margin loss due to the cycling induced mobility degradation, in minimizing the memory cell current variation to ease the design of the sensing circuitry. Figure 5.1 shows the overall architecture of the booster. The new aspect of the booster is that an A/D converter, which is in the A/D converter block, is used to measure the supply voltage, whose variation is mainly responsible for the variation in the boosted wordline voltage. The Vboost block uses the measured digital code from the A/D converter to adjust the amount of boosting capacitance, thus the boosted wordline voltage can be controlled accurately. Fast speed is achieved because no feedback regulating operation is involved after the selected wordline has been boosted to a higher voltage than the power supply voltage; note that the digital code used by the Vboost block has been measured before the boosting operation. Section

5.1 describes the design and the analysis of the A/D converter block and the design trade off between speed and accuracy. Section 5.2 focuses on the design methodology for the Vboost block.

Figure 5.1: A/D booster

5.1 A/D Converter block

The A/D converter block is used to measure the power supply voltage VCC before a boosting operation occurs, producing a digital thermometer code AD(3:0). A "1" in the digital code will activate a booster cell in the Vboost block. The more bits the A/D converter block has, the more accurate the booster is. The number of bits, however, is limited by the accuracy of the reference voltage FVREF as well as the speed of the comparators and the fast reference circuit. For 1.8V flash memories, the power supply voltage VCC can vary from 1.6V to 2.0V. With this range of variation, 4 bits in a thermometer code is a reasonable number, and thus the target thermometer codes at different supply voltages can be designed to be those shown in Table 5.1.

Table 5.1: Target thermometer codes for the A/D converter block at different supply voltages

VDD (V)	AD(3:0)
1.6	1111
1.7	1110
1.8	1100
1.9	1000
2.0	0000

Figure 5.2 shows the architecture of the A/D converter block. It consists of three main components: the resistive divider chain, the comparators and the fast reference circuit.

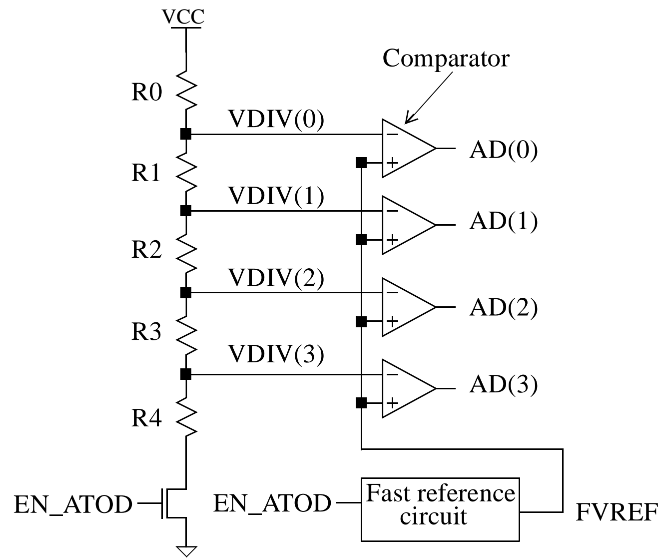


Figure 5.2: A/D converter block

The timing of the A/D converter block in relative with that of a read access is depicted in Figure 5.3. The typical read access time is about 45ns. For a new read access, an address transition pulse (ATD) is generated when a new address is applied. The A/D converter block is enabled (by the signal EN_ATOD shown in Figure 5.3) at the rising edge of the ATD pulse to measure the supply voltage VCC.

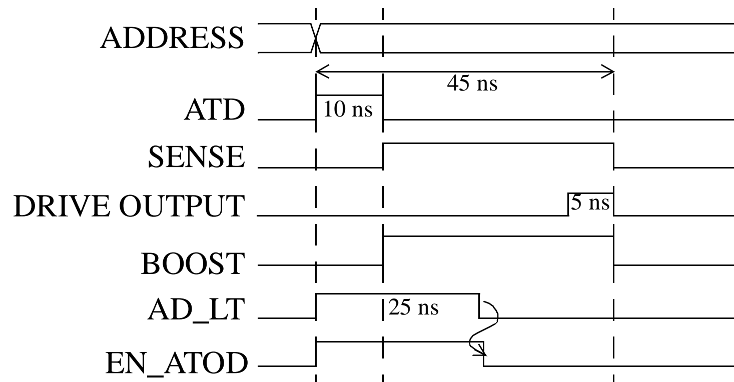


Figure 5.3: Typical booster timing in an initial read access

The speed (the settling time of FVREF in response to the signal EN_ATOD) of the reference circuit must be fast so that the A/D thermometer code can become valid sooner and thus the read access time can be shortened. To reduce the read access time further, at the end of the ATD pulse, the wordline is boosted to some voltage close to but not exactly the final value (because the A/D thermometer code is not valid yet). About 15ns later when the thermometer code from the A/D converter block becomes valid and is latched, the boosted level will be adjusted to the correct level. After latching, the A/D converter block can be disabled to save power.

5.1.1 Designing the resistor chain

A separate resistive divider could be used for each comparator in Figure 5.3, but to save area a common resistor chain with four output taps is used. Suppose that the maximum allowed current for the resistor chain is 100µA, then the total resistance $R_{total} = R0 + R1 + R2 + R3 + R4$ is $\frac{2V}{100\mu A} = 20k\Omega$. Assume that the reference voltage FVREF is 1.17V (will be clarified in Section 5.1.2) and also assume that the comparator will switch when VDIV equals to FVREF (in reality, the comparator needs a small differential input to switch), then the value of R4 can be chosen such that when VCC=1.9V, VDIV(3) equals to 1.17V for the bit AD(3) to switch from 0 to 1. Using the resistor divider formula, the value of all resistors in the resistor chain can be calculated from the following five equations:

$$R4 = \frac{(1.17)R_{total}}{1.9} \quad (5.1)$$

$$R3 = \frac{(1.17)R_{total}}{1.8} - R4 \quad (5.2)$$

$$R2 = \frac{(1.17)R_{total}}{1.7} - (R4 + R3) \quad (5.3)$$

$$R1 = \frac{(1.17)R_{total}}{1.6} - (R4 + R3 + R2) \quad (5.4)$$

$$R0 = R_{total} - (R4 + R3 + R2 + R1) \quad (5.5)$$

Using the resistor values calculated from the above equations, the voltages of VDIV(3:0) at different VCCs are calculated and shown in Table 5.2:

Table 5.2: Voltage levels of VDIV nodes at different VCCs

	VDD=1.6V	VDD=1.7V	VDD=1.8V	VDD=1.9V	VDD=2.0V
VDIV(3) (mV)	985	1,047	1,108	**1,170**	1,232
VDIV(2) (mV)	1,040	1,105	**1,170**	1,235	1,300
VDIV(1) (mV)	1,101	**1,170**	1,239	1,308	1,376
VDIV(0) (mV)	**1,170**	1,243	1,316	1,389	1,462

Theoretically, one of the 4 A/D bits can be wrong no matter how the resistive chain is designed because at some VCC between 1.6V and 2.0V, one of the signal VDIV will be equal to FVREF and the corresponding comparator can not switch in time; this one-bit error defines the accuracy of the booster. However, a 2-bit error can certainly be avoided. Table 5.2 reveals that the margin between every other VDIV nodes is about 120 mV, thus the variation of the reference voltage FVREF with respect to temperature (from -40C to 100C for a typical specification), power supply voltage VCC and process variation must not be larger than 120 mV. Designing a fast reference circuit that can generate FVREF in about 25ns is a challenging task but it can be done.

5.1.2 Designing the fast reference voltage circuit

The fast reference voltage circuit can be constructed from a traditional bandgap reference circuit [22] shown in Figure 5.4. The bipolar current mirror insures equal currents through Q1 and Q2. Note that Q1 is n times bigger than Q2, thus the voltage drop across the resistor R1 is $\Delta V_{BE} = V_{BE,Q2} - V_{BE,Q1} = V_T \ln\frac{I_C}{I_S} - V_T \ln\frac{I_C}{nI_S} = \frac{kT}{q}\ln(n)$.

ΔV_{BE} exhibits a positive temperature coefficient, and if it is combined with $V_{BE,Q3}$, which has a negative temperature coefficient, VREF will be basically independent with temperature.

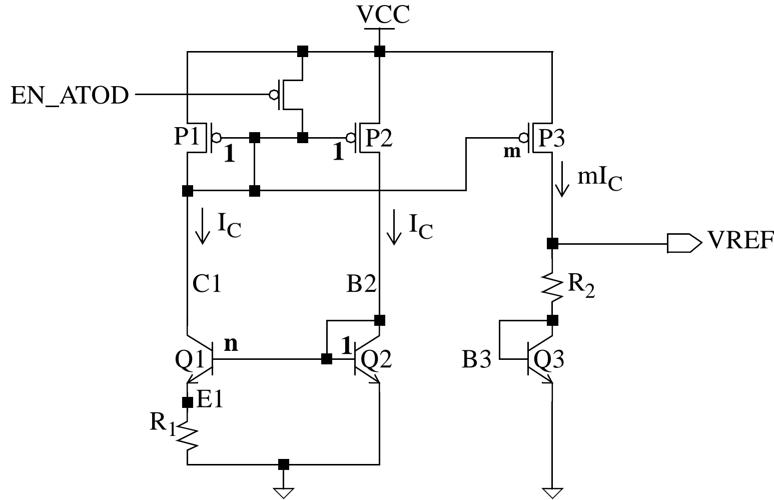


Figure 5.4: Traditional bandgap reference circuit

From Figure 5.4, VREF can be calculated as:

$$VREF = V_{BE,Q3} + R_2(mI_C) = V_{BE,Q3} + R_2 m\frac{\Delta V_{BE}}{R_1} = V_{BE,Q3} + m\frac{R_2}{R_1}\ln(n)\frac{kT}{q} \quad (5.6)$$

To have good temperature compensation, $\frac{\partial}{\partial T}VREF$ should be equal to 0:

$$\frac{\partial}{\partial T}VREF = \frac{\partial V_{BE}}{\partial T} + m\frac{R_2}{R_1}\ln(n)\frac{k}{q} = 0 \quad (5.7)$$

Assume $\frac{\partial V_{BE}}{\partial T} = -1.5mV/^{o}K$, VREF with zero temperature coefficient can be achieved at 1.25V. Note that the current mirror in Figure 5.4 is usually in the cascode form to reduce the variation of VREF with the supply voltage VCC.

Unfortunately, the circuit in Figure 5.4 can not be used directly for the A/D booster due to its slow settling time in response to EN_ATOD. Moreover, it is not easy to design a reference voltage of 1.25V when the supply voltage VCC is at 1.6v; at extreme process corners and temperatures, the PMOS transistor P3 can go out of the saturation region. Thus there is a need to lower the FVREF level for the circuit to work even when VCC is 1.6V. To lower the FVREF level, according to Equation (5.7), $\frac{\partial V_{BE}}{\partial T}$ needs to be less negative. As discussed in [23], $\frac{\partial V_{BE}}{\partial T}$ can be less negative if the current running through Q3 is PTAT (proportional to absolute temperature). This makes sense because if the current through Q3 increases while the temperature is kept unchanged, V_{BE} will increase, and therefore when the temperature increases, the reduction in V_{BE} due to the increasing in temperature is partly compensated by the increasing in V_{BE} due to the increasing in the current. In short, to lower the level of FVREF, the current running through Q3 should be PTAT. As shown in Figure 5.4, the current I_C can be calculated as $\frac{\Delta V_{BE}}{R_1} = \frac{\frac{kT}{q}\ln(n)}{R_1}$. To have a large positive temperature coefficient for I_C, R1 is chosen to be an unsilicided polysilicon resistor, which has a negative temperature coefficient. With this choice, the level of FVREF with zero temperature coefficient is about 1.17V. At this FVREF level, the circuit can be biased such that transistor P3 is in saturation region at VCC=1.6V, as long as the gate drive of transistor P1 (diode-connected) is less than (1.6V - 1.17V); this condition can be satisfied by choosing the size of transistors P1 and P2 to be large enough.

Figure 5.5 presents the new fast reference circuit in which R1 is made of unsilicided polysilicon. The word “fast” here means the fast settling time for FVREF in response to the signal EN_ATOD. There are further details about this circuit that need to be discussed. First of all, the cascode topology is not used for the PMOS current mirror because there would be no head room when VCC=1.6V. This decision affects the accuracy of the fast reference circuit, but in fact a very accurate FVREF for this A/D converter is not needed; the variation in FVREF of about 70mV is still acceptable in this context.

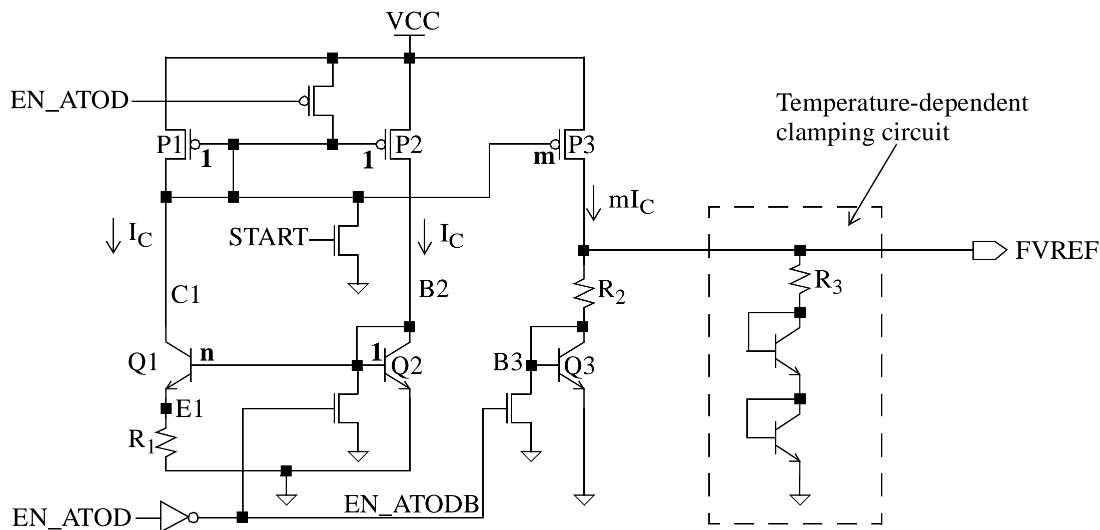


Figure 5.5: Fast bandgap reference circuit

Secondly, to shorten the setling time, FVREF is initialized to VCC first; if FVREF starts from ground and goes to 1.17V, the variation in FVREF at a particular time, for example 25ns after being enabled, is large. This is because the ac performance of FVREF is very different at extreme process variations, temperatures and power supply voltages; at some condition, FVREF has overshoot, at other it has undershoot. Starting FVREF at VCC has the advantage that at all extreme conditions FVREF behaves the same way: going down from VCC to about 1.17V. The signal START in Figure 5.5, which is a pulse of about 2 to 3ns, is used as a start-up circuitry and also used to initialize FVREF to VCC.

The third detail need to be mentioned is that, to lower the level of FVREF, the current I_C is designed to be PTAT, but this also leads to an undesirable effect; at high temperature the current through P3 increases, keeping FVREF from going quickly to the correct level. To minimize this effect, a temperature-dependent clamping circuit is added as shown in Figure 5.5. The clamp is used to bring FVREF quickly to the final level, especially at high temperature because V_{BE} of the two bipolar transistors decreases when temperature increases. Resistor R3 is used to fine-tune the effect of the clamp. The present of the

clamp, of course, will affect the final value of FVREF, but by adjusting R3 appropriately, the error is less than 20mV.

A few last details about the circuit in Figure 5.5 are discussed next. To adjust FVREF with the process variation, trimming is used. Resistor R_2 can be trimmed using two or three CAM (Content Addressable Memory) cells. NMOS transistors with the gate connected to EN_ATODB are used to discharge B2 and B3 to make sure that subsequent FVREF pulses will have the same level as the first pulse. The simulation result for the fast bandgap reference is shown in Figure 5.6. The 81 curves cover all extreme and typical combinations of supply voltages, temperatures and process variation. At about 25ns to 30ns after being enabled, FVREF is fairly stable. Shortly after that when A/D digital code has been determined, the fast reference circuit is turned off to save power.

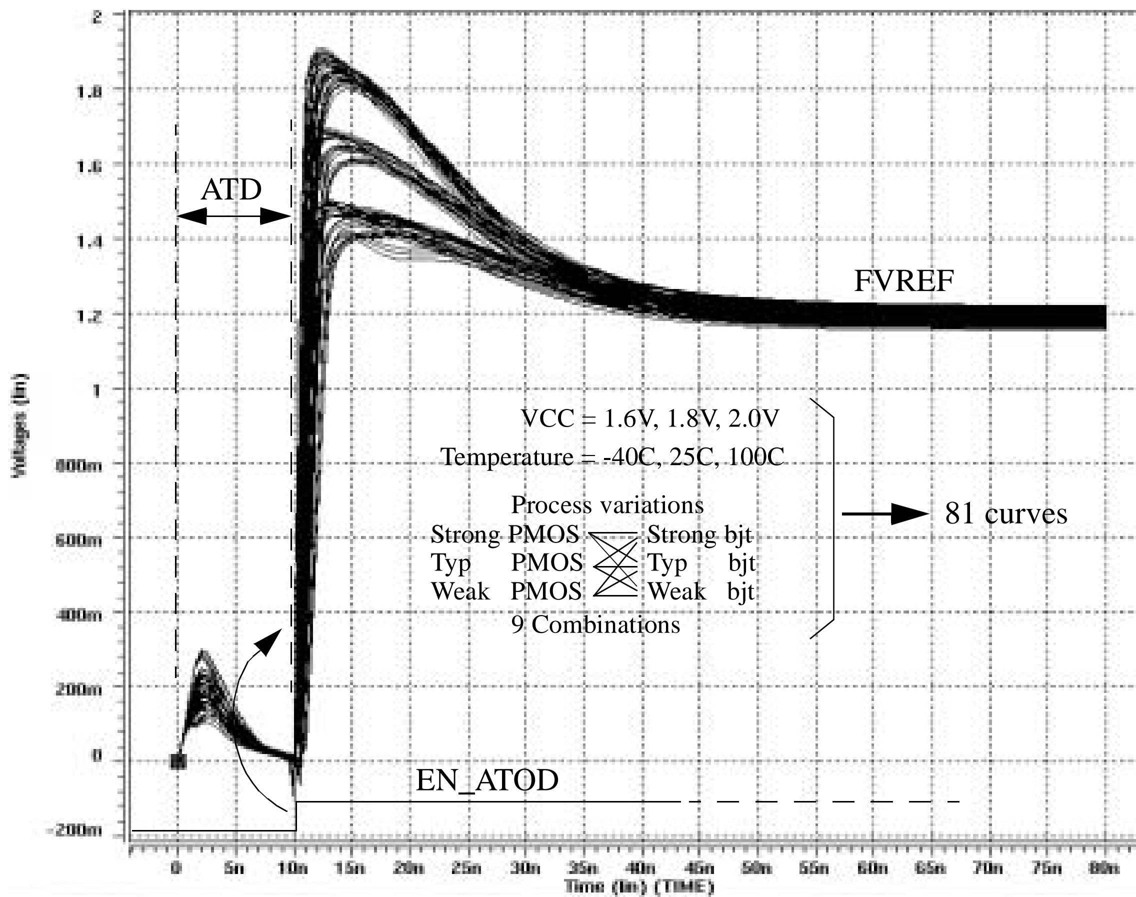


Figure 5.6: HSPICE simulation result for fast bandgap reference circuit

5.1.3 Designing the comparator

The comparator is a 3-stage differential amplifier followed by a latch as shown in Figure 5.7. The number of stages is chosen to be three to minimize the delay and area, according to [24]. Because the levels of FVREF and VDIV is not very high compared to VCC, the load for the first stage need not to be resistors. The comparator here is actually a copy of the sense amplifier, in which the input level can be close to VCC, to save the layout effort.

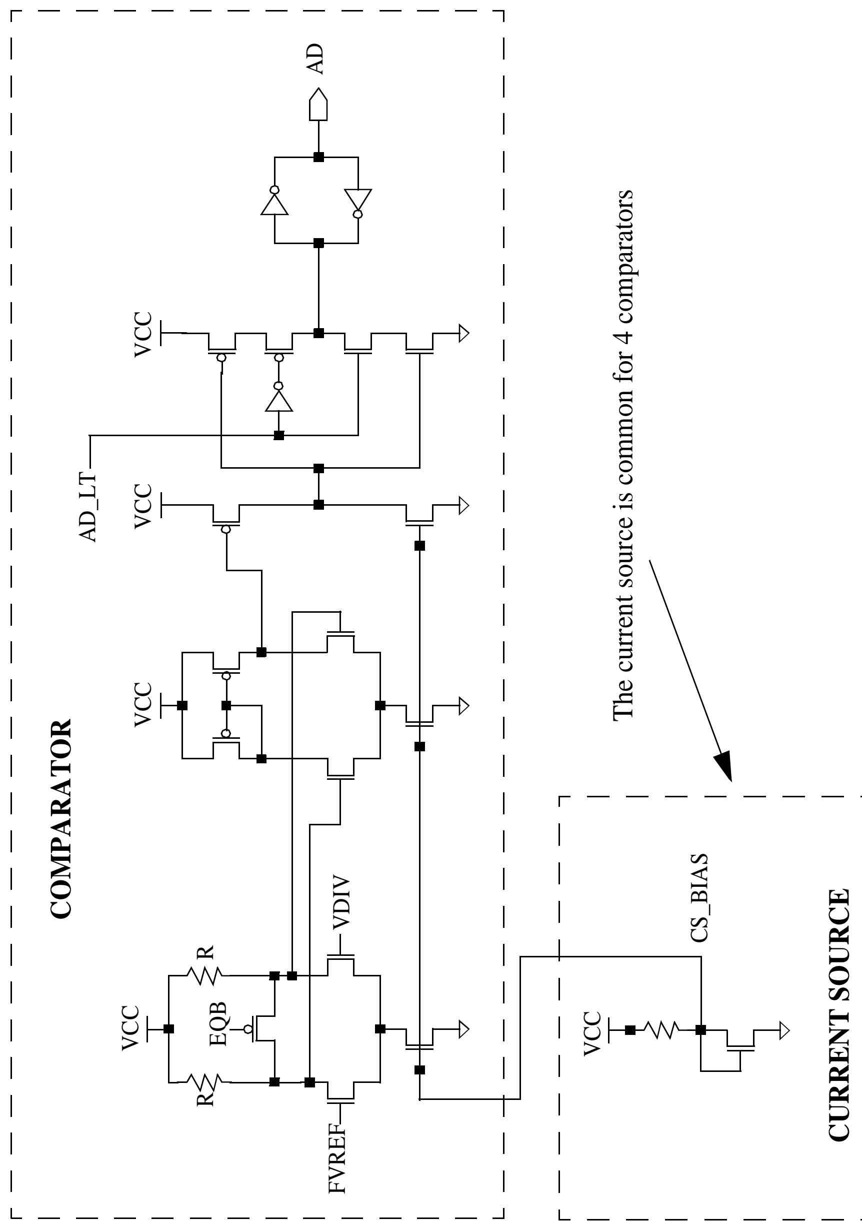


Figure 5.7: Comparator

5.1.4 Simulation result for the A/D converter block

The HSPICE simulation result for the whole A/D converter block at all conditions is shown in Table 5.3. The good results in Table 5.3 show that for all combinations, there are only 1-bit errors at different VCCs, compared to the ideal digital codes in Table 5.1. As will be shown later in Section 5.2, 1-bit error causes only 200mV to 300mV error in the boosted wordline voltage (~ 4.8V).

Table 5.3: Simulation result for the A/D converter block

Process combinations	VCC=1.6 AD(0:3)	VCC=1.7 AD(0:3)	VCC=1.8 AD(0:3)	VCC=1.9 AD(0:3)	VCC=2.0 AD(0:3)
TEMPERATURE = 100C					
Weak_Pmos-Weak_bjt	1111	0111	0011	0001	0000
Weak_Pmos-Typical_bjt	1111	1111	0111	0011	0001
Weak_Pmos-Strong_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Weak_bjt	1111	0111	0011	0001	0000
Typical_Pmos-Typical_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Strong_bjt	1111	1111	0111	0011	0001
Strong_Pmos-Weak_bjt	1111	0111	0011	0001	0000
Strong_Pmos-Typical_bjt	1111	0111	0111	0011	0001
Strong_Pmos-Strong_bjt	1111	1111	0111	0011	0001
TEMPERATURE = 25C					
Weak_Pmos-Weak_bjt	1111	0111	0111	0011	0001
Weak_Pmos-Typical_bjt	1111	1111	0111	0011	0001
Weak_Pmos-Strong_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Weak_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Typical_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Strong_bjt	1111	1111	0111	0011	0001
Strong_Pmos-Weak_bjt	1111	0111	0011	0001	0000
Strong_Pmos-Typical_bjt	1111	0111	0011	0001	0001
Strong_Pmos-Strong_bjt	1111	0111	0111	0011	0001
TEMPERATURE = -40C					
Weak_Pmos-Weak_bjt	1111	0111	0011	0011	0000
Weak_Pmos-Typical_bjt	1111	0111	0111	0011	0001
Weak_Pmos-Strong_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Weak_bjt	1111	0111	0011	0011	0000
Typical_Pmos-Typical_bjt	1111	1111	0111	0011	0001
Typical_Pmos-Strong_bjt	1111	1111	0111	0011	0001
Strong_Pmos-Weak_bjt	1111	0111	0011	0001	0000
Strong_Pmos-Typical_bjt	1111	0111	0011	0001	0000
Strong_Pmos-Strong_bjt	1111	0111	0011	0001	0000

Figure 5.6 shows that at about 25ns after being enabled, FVREF is still changing, although slowly, thus the digital code obtained can be affected depending on when the signal AD_LT (Figure 5.3) goes down to latch the A/D code. To test the robustness of the A/D converter block, another HSPICE simulation is performed in which the falling edge of the signal AD_LT is shifted by 3ns. The new result shows that at different VCCs, there are only 1-bit errors, which is very similar to the result shown in Table 5.3.

5.1.5 Alternative FVREF design - Speed/Accuracy trade-off

Figure 5.3 reveals that the speed of the A/D converter block is limited by the speed of the fast reference circuit, which needs additional 15ns to 20ns after the falling edge of the ATD pulse to settle. This additional time takes up a large fraction of the 45ns read access time. It is very difficult to further improve the speed of the fast bandgap reference circuit shown in Figure 5.5. To substantially cut down the delay time for the fast reference circuit, different and simple topologies must be used. Simpler topologies are faster but the accuracy may be less. For the wordline booster, the speed is very important because it directly affects the read access time, thus the accuracy should be sacrificed somewhat in favor of read speed. This means that the 2-bit errors may be acceptable.

Figure 5.8 shows the simple fast reference circuit which is actually used in the 256Mbit memory chip described in this dissertation. Resistors R1 and R2 are made of unsilicided polysilicon and N-well, respectively. Trimming is used to minimized the variation in FVREF caused by process variation of transistor M1, resistor R1 and resistor R2. Extensive simulations have been conducted on this circuit and the result (with trimming applied) is shown in Figure 5.9. The speed of the reference circuit is very fast; FVREF starts from VCC (when EN_ATOD is low) then goes down to the final level (when EN_ATOD becomes high) in just about 2ns. This is a significant improvement from the circuit shown in Figure 5.5 where the speed is about 25ns to 30ns. The variation of FVREF is about 235mV (from 1.207V to 1.422V), but most of this variation is due to the

power supply VCC. Thus, by redesigning the resistor chain to track the variation of FVREF with VCC, the new reference circuit can make the A/D converter block work.

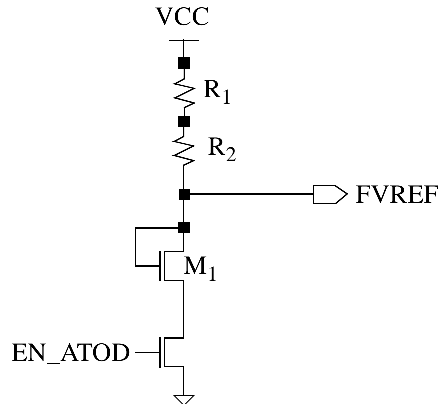


Figure 5.8: Simple fast reference circuit

In Table 5.2, which is used to design the resistor chain for the fast reference circuit in Figure 5.5, VDIV(3) is designed to be 1.17V when VCC=1.9V. Note that FVREF is assumed to be 1.17V when Table 5.2 is constructed. Similarly, VDIV(2), VDIV(1) and VDIV(0) are designed to be 1.17V when VCC=1.8V, 1.7V and 1.6V, respectively. For the simple fast reference circuit, the design for the resistor chain is slightly modified since FVREF varies with VCC. From the simulation results in Figure 5.9, FVREF can be determined as 1.24V, 1.26V, 1.29V, 1.33V and 1.37V when VCC=1.6V, 1.7V, 1.8V, 1.9V and 2.0V, respectively. Thus VDIV(0), VDIV(1), VDIV(2) and VDIV(3) are designed to be 1.24V, 1.26V, 1.29V and 1.33V when VCC=1.6V, 1.7V, 1.8V and 1.9V, respectively. With this tracking design for the resistor chain, the A/D converter block will work if the change in FVREF due to 100mV change in VCC, not 400mV change (from 1.6V to 2.0V), is smaller than the change in VDIV and this is actually true for this type of topology. Temperature also contributes to the variation of FVREF, which can not be fixed by trimming. Because R1 (unsilicided polysilicon) has a negative temperature coefficient, R2 (N-well) and transistor M1 have positive temperature coefficients, a careful trimming for these three components can minimize the variation of FVREF with temperature.

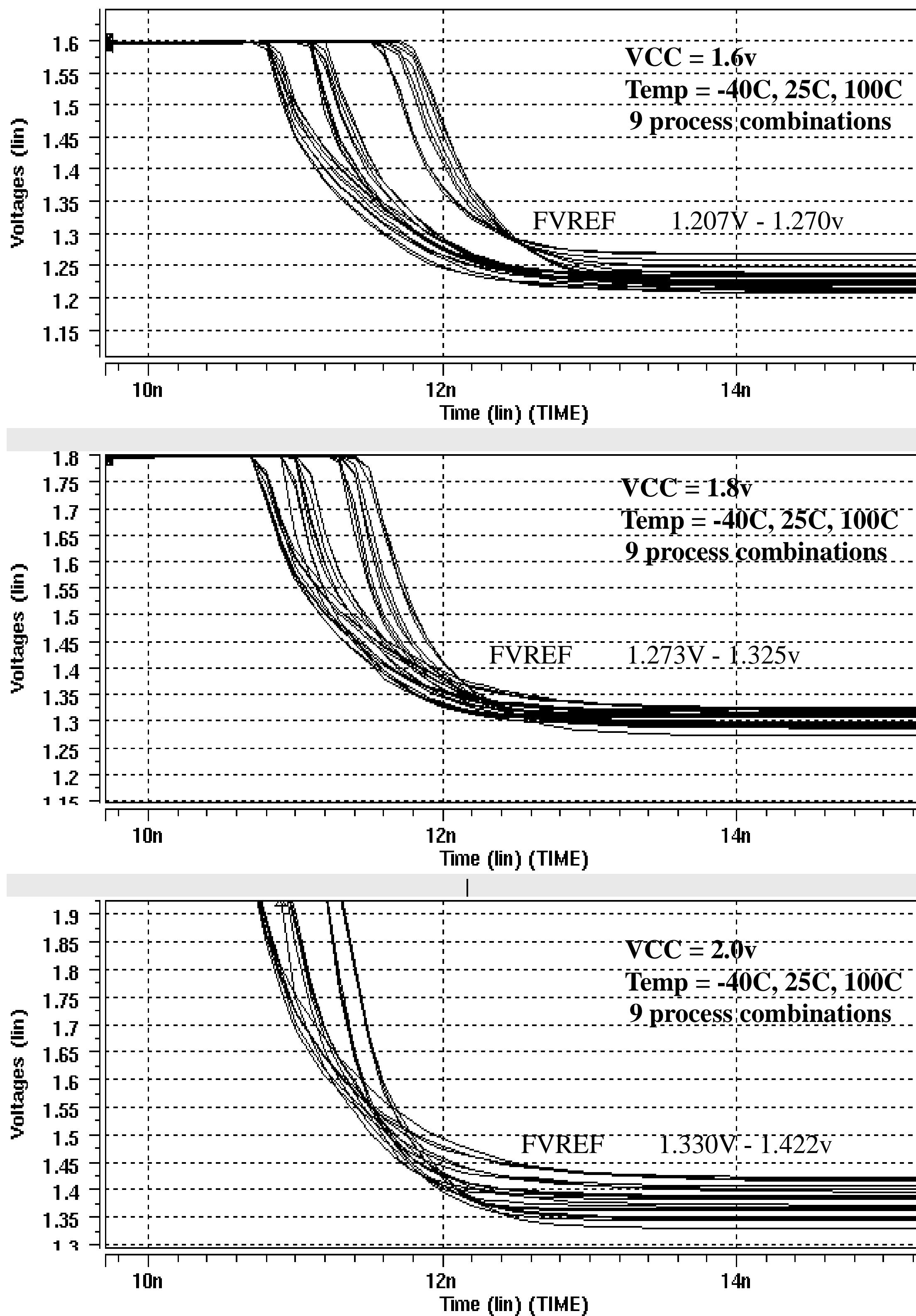


Figure 5.9: HSPICE simulation result for the simple fast reference circuit

Table 5.4 shows the result for one set of 135 simulations, with different combinations of temperature, process and supply voltage VCC. Most of the errors, compared to Table 5.1, are 1-bit errors, which is very good. There are only a few 2-bit errors when VCC=2.0V, but this is fine because VCC=2.0V is usually the best case for read, thus this inaccuracy just makes the best case a little bit worse, and does not make it the worst case. The result for another set of 135 simulations when maximum polysilicon resistance is combined with minimum N-well resistance gives similar result. Other process combinations can be fixed by trimming in similar manner to obtain similar result shown in Table 5.4.

Table 5.4: A/D converter block simulation result, using the simple fast reference circuit

TEMPERATURE = 100C					
Process combinations	VCC=1.6 AD(0:3)	VCC=1.7 AD(0:3)	VCC=1.8 AD(0:3)	VCC=1.9 AD(0:3)	VCC=2.0 AD(0:3)
Weak_Nmos, Max poly, Max nwell	1111	0111	0011	0001	0000
Weak_Nmos-Typ poly, Typ nwell	1111	0111	0011	0001	0000
Weak_Nmos-Min poly, Min nwell	1111	0111	0011	0001	0000
Typical_Nmos-Max poly, Max nwell	1111	0111	0011	0011	0001
Typical_Nmos-Typ poly, Typ nwell	1111	1111	0011	0011	0001
Typical_Nmos-Min poly, Min nwell	1111	1111	0011	0011	0001
Strong_Nmos-Max poly, Max nwell	1111	1111	0111	0011	**0011**
Strong_Nmos-Typ poly, Typ nwell	1111	1111	0111	0011	**0011**
Strong_Nmos-Min poly, Min nwell	1111	1111	0111	0011	**0011**
TEMPERATURE = 25C					
Weak_Nmos, Max poly, Max nwell	1111	1111	0011	0001	0000
Weak_Nmos-Typ poly, Typ nwell	1111	1111	0011	0001	0000
Weak_Nmos-Min poly, Min nwell	1111	1111	0011	0001	0000
Typical_Nmos-Max poly, Max nwell	1111	1111	0011	0011	0001
Typical_Nmos-Typ poly, Typ nwell	1111	1111	0011	0011	0001
Typical_Nmos-Min poly, Min nwell	1111	1111	0011	0011	0001
Strong_Nmos-Max poly, Max nwell	1111	1111	0111	0011	0001
Strong_Nmos-Typ poly, Typ nwell	1111	1111	0111	0011	0001
Strong_Nmos-Min poly, Min nwell	1111	1111	0111	0011	0001
TEMPERATURE = -40C					
Weak_Nmos, Max poly, Max nwell	1111	1111	0011	0001	0000
Weak_Nmos-Typ poly, Typ nwell	1111	1111	0111	0001	0000
Weak_Nmos-Min poly, Min nwell	1111	1111	0111	0001	0000
Typical_Nmos-Max poly, Max nwell	1111	0111	0011	0001	0000
Typical_Nmos-Typ poly, Typ nwell	1111	1111	0011	0001	0000
Typical_Nmos-Min poly, Min nwell	1111	1111	0011	0001	0000
Strong_Nmos-Max poly, Max nwell	1111	0111	0011	0001	0001
Strong_Nmos-Typ poly, Typ nwell	1111	0111	0011	0001	0000
Strong_Nmos-Min poly, Min nwell	1111	0111	0011	0001	0001

5.2 Vboost block

Vboost block contains the boosting capacitors used to boost the wordline path (VBOOST) to a voltage higher than the supply voltage VCC in a read operation. A simple view of the Vboost block is presented in Figure 5.10, in which the wordline path capacitive loading of about 17pF is represented by the capacitor CLOAD.

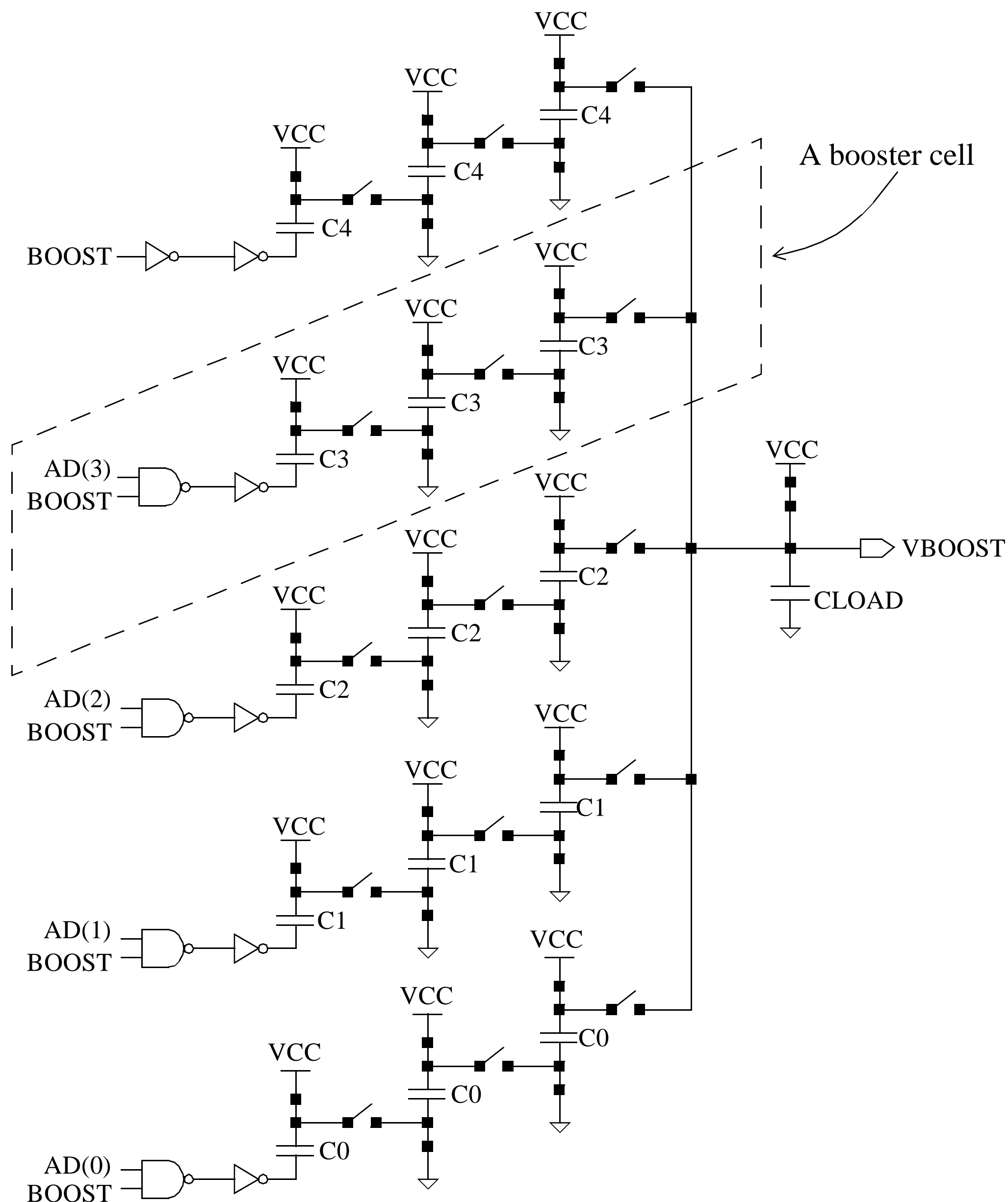


Figure 5.10: A simple view of the Vboost block. The block is in the precharged condition

The Vboost block is divided into five booster cells. Each cell, except cell 4 at the top of Figure 5.10, is controlled by an A/D bit from the A/D converter block. The A/D bits, which contains the information about the supply voltage level, dictate the number of active booster cells, thus the boosting level VBOOST is controlled accurately despite the variation in the supply voltage VCC. Booster cell 4 is not controlled by any A/D bit and is always activated when the BOOST signal arrives; when the power supply voltage VCC is at its maximum (i.e. 2.0V), only this booster cell is used. At lower power supply voltages, additional booster cells will be activated, based on the status of the A/D bits. The actual schematic of the Vboost block is shown in Figure 5.11. The non-overlap clocks N_RESET, BOOST_BH and P_ONEB, using the cross-couple nand gates, are used to prevent the charge leakage when connecting and opening the control switches in Figure 5.10. BOOST_BH is used to activate the selected booster cells.

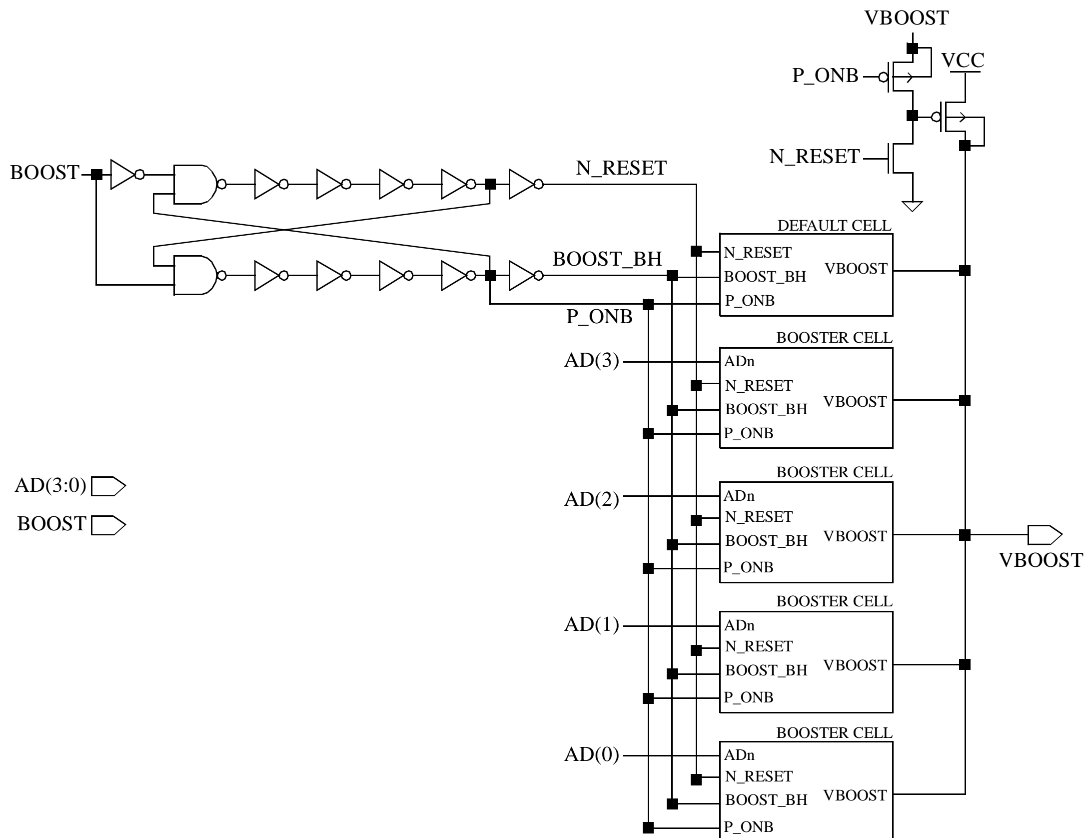


Figure 5.11: Actual Vboost block schematic

The actual booster cell is shown in Figure 5.12. The capacitors used in the booster cells are poly-nwell capacitors.

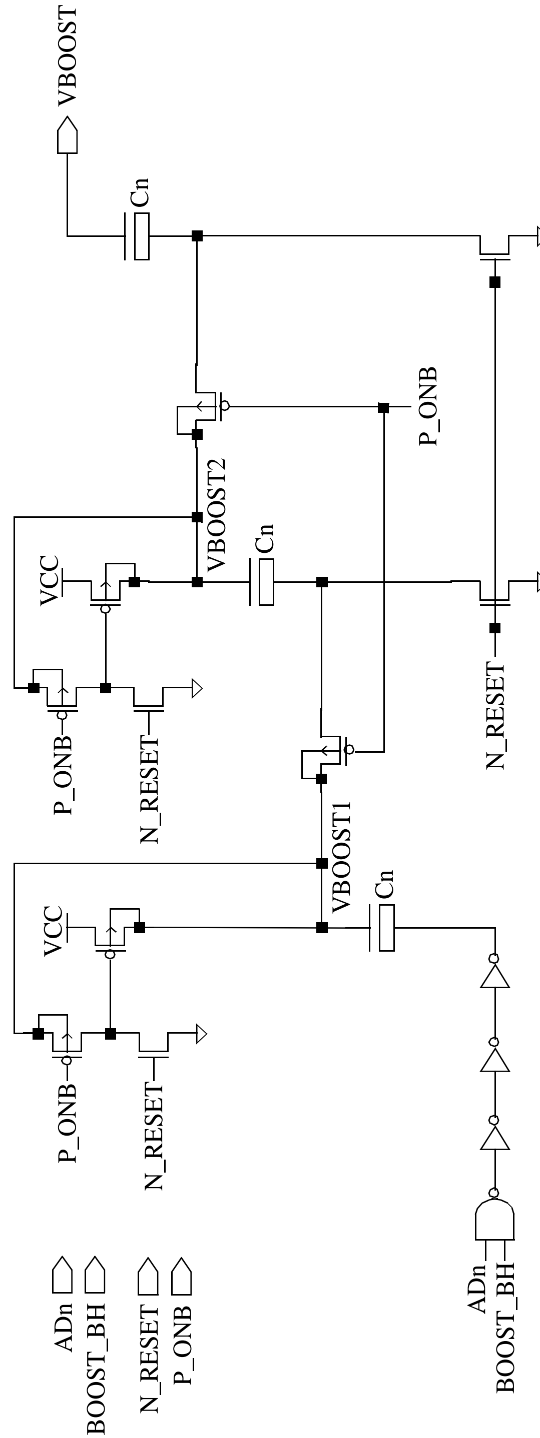


Figure 5.12: Booster cell. Index n = 0, 1, 2, 3.

The target voltage of VBOOST is about 4.8V. The minimum power supply voltage VCC can be as low as 1.6V, thus each booster cell needs three capacitors to boost the output from 1.6V to 4.8V. These three capacitors are precharged to VCC during the ATD time shown in Figure 5.3, then when the boosting time comes, they are connected in series; the bottom of this series combination is also switched from ground to VCC, thus the output VBOOST is about four times of VCC level, if there is no capacitive loading at the output. When the output capacitive loading is present, due to the charge sharing between the boosting capacitors and the load capacitor, the output voltage VBOOST is lower. Note that CLOAD is also precharged to VCC before the boosting operation begins.

In Figure 5.10, there is a switch at the output of each booster cell to connect the cell to the output of the booster VBOOST, if the A/D bit for that cell is asserted. This switch is controlled by a high voltage switch shown in Figure 5.13.

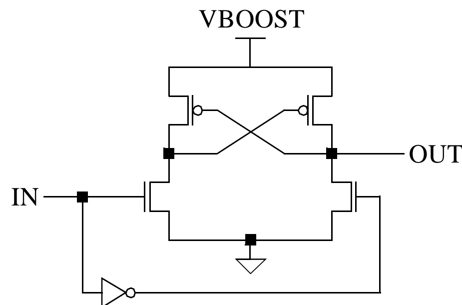


Figure 5.13: High voltage switch

Due to the speed requirements, however, all the high voltage switches at the outputs of the booster cells are eliminated. The outputs of the booster cells are tied directly together, thus if some cells are off and the others are on, then the off-cells, along with the wordline path, will act as the load of the on-cells. Therefore, to find the size of the capacitors in the booster cells (C4, C3, C2, C1, C0) we have to solve concurrently five equations with five unknowns, equations (5.11), (5.12), (5.13), (5.14) and (5.15). These equations are written for 5 VCCs: 1.6, 1.7, 1.8, 1.9 and 2.0V. For each equation, the initial charge on the capacitors is equated with the final charge. Note that because three capacitors are in series, their equivalent capacitance is only C/3.

$$V_{BOOST}\left(\frac{C_3}{3}+\frac{C_2}{3}+\frac{C_1}{3}+\frac{C_0}{3}+C_{load}\right)+(V_{BOOST}-2.0)\left(\frac{C_4}{3}\right) = (2.0)C_{total} \quad (5.11)$$

$$V_{BOOST}\left(\frac{C_2}{3}+\frac{C_1}{3}+\frac{C_0}{3}+C_{load}\right)+(V_{BOOST}-1.9)\left(\frac{C_4}{3}+\frac{C_3}{3}\right) = (1.9)C_{total} \quad (5.12)$$

$$V_{BOOST}\left(\frac{C_1}{3}+\frac{C_0}{3}+C_{load}\right)+(V_{BOOST}-1.8)\left(\frac{C_4}{3}+\frac{C_3}{3}+\frac{C_2}{3}\right) = (1.8)C_{total} \quad (5.12)$$

$$V_{BOOST}\left(\frac{C_0}{3}+C_{load}\right)+(V_{BOOST}-1.7)\left(\frac{C_4}{3}+\frac{C_3}{3}+\frac{C_2}{3}+\frac{C_1}{3}\right) = (1.7)C_{total} \quad (5.13)$$

$$V_{BOOST}(C_{load})+(V_{BOOST}-1.6)\left(\frac{C_4}{3}+\frac{C_3}{3}+\frac{C_2}{3}+\frac{C_1}{3}+\frac{C_0}{3}\right) = (1.6)C_{total} \quad (5.14)$$

where $C_{total} = C_4+C_3+C_2+C_1+C_0+C_{load}$

HSPICE simulations are performed on the Vboost block to show the effectiveness of the A/D wordline booster in controlling accurately the wordline voltage. If the A/D converter block is not used, VBOOST variation with VCC is very large, from 4.8V to 6.0V, as shown in Figure 5.14.

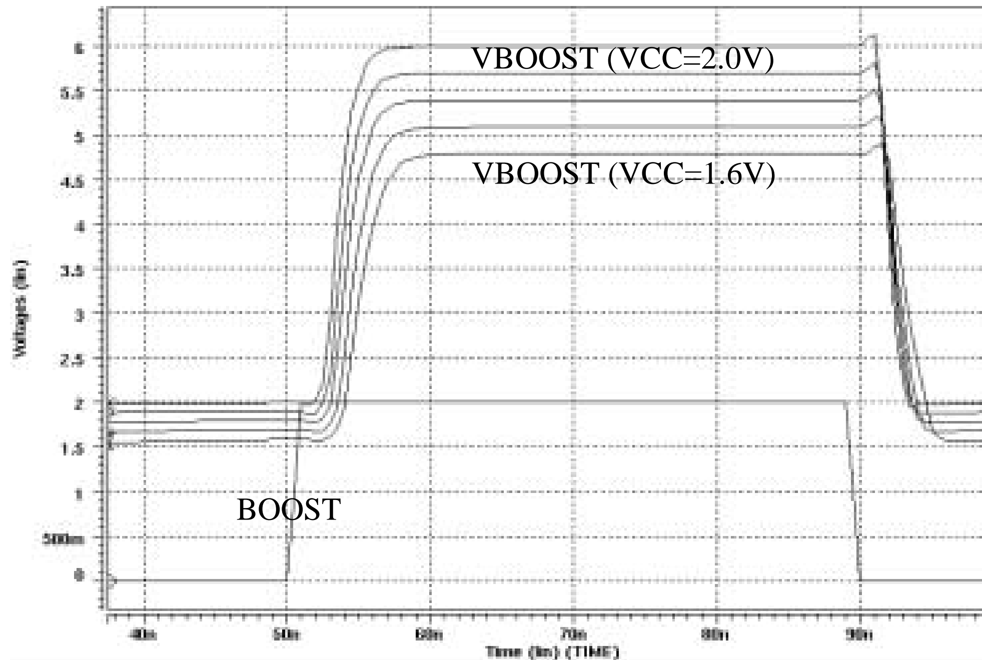


Figure 5.14: Simulation result for the wordline booster without using the A/D converter

When the A/D bits are used, the VBOOST variation are much less. Figure 5.15 shows the small variation of VBOOST when no bit in the A/D code is wrong. Also note that the speed of the A/D wordline booster is very fast. It takes about 10ns for VBOOST to go from VCC to the final value of 4.8V, assuming that CLOAD = 17pF.

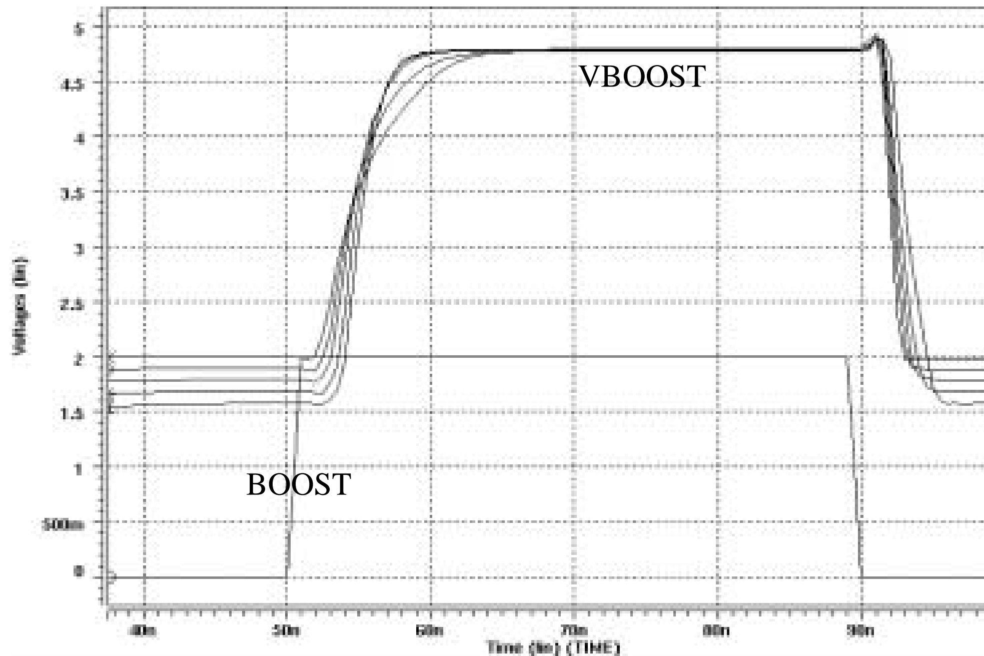


Figure 5.15: Simulation result for the A/D wordline booster - 0-bit error

However, as discussed in Section 5.1.1, no matter how the A/D converter block is designed, the A/D code must have a wrong bit at some power supply voltage VCC. If there is one wrong bit, VBOOST variation is still very good, about 250mV as shown in Figure 5.16. In case the fast reference circuit in Figure 5.5 is used, VBOOST can be boosted to a level near the final level right after the ATD pulse, without waiting 15ns or 20ns when the FVREF settles; in this case two of the A/D bits are forced to be high, then when the A/D code becomes valid, VBOOST level will be adjusted to the correct level. In case the simple fast reference circuit in Figure 5.8 is used, the adjusting action is not required because the simple fast reference circuit is very fast and the valid A/D code can be generated at the end of the ATD pulse. When this simple fast reference circuit is used,

there may be 2-bit errors, but according to the simulation in Figure 5.16, VBOOST variation is about 500mV, still much better than 1.2V if the A/D converter block is not used. This speed and accuracy trade-off has been discussed in Section 5.1.5.

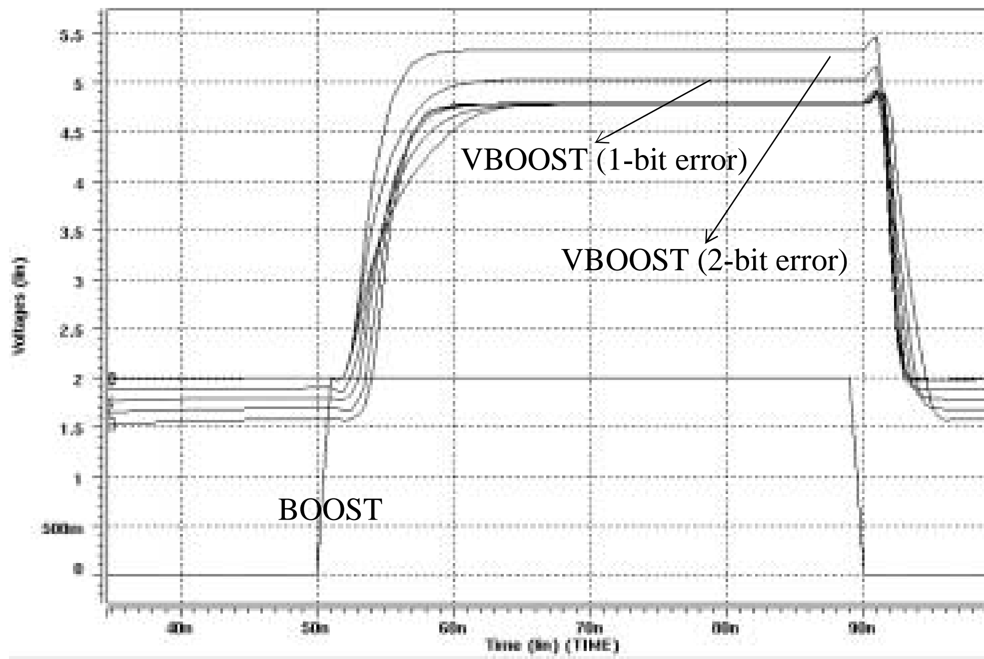


Figure 5.16: Simulation result for the A/D wordline booster - 1-bit and 2-bit errors

5.3 Summary

The auto-calibrated wordline booster described in this chapter can offer both speed and accuracy that other types of boosters can not. In addition, the maximum power consumption of the A/D booster is smaller because at higher supply voltage, some of the booster cells are deactivated. The area for the entire booster is only 0.7% of the die size. The accuracy of the booster can be traded off with its speed when the simple fast reference circuit is used. Also, the fast reference circuit is very small because it does not contain large bipolar transistors. Sacrificing some accuracy can make the booster both fast and small.

Chapter 6

Read path simulation and measured results

The new sensing techniques developed in Chapters 2, 3, 4, 5, which are the sense current recovery technique, the SSDDDPPP column decoding, the differential feedback cascoded bitline voltage control and the auto-calibrated wordline voltage control have been incorporated into the sensing path (read path) of the 1.8V, 256Mb, two-bit-per-cell nitride-storage flash memory. Section 6.1 shows the overall simulation results of the read path, which also includes the effects of parasitic resistive and capacitive loading of all elements in the speed path such as the wordline decoding, bitline decoding, read control signal and data-out paths. This 1.8V, 256Mb, two-bit-per-cell nitride-storage flash memory has been fabricated in the 0.13μm, triple-metal, nitride-storage CMOS process technology. Section 6.2 presents the most important read path measured results related to the new sensing techniques developed for the memory; the section also describes the set-up and the conditions in which these measured results are obtained.

6.1 Read path simulation results

The read path simulation set-up diagram is shown in Figure 6.1. The diagram shows clearly where the new sensing techniques are applied, relative to other components of the whole read path (sensing path). The read speed-path is shown to aid the discussions about the read access time.

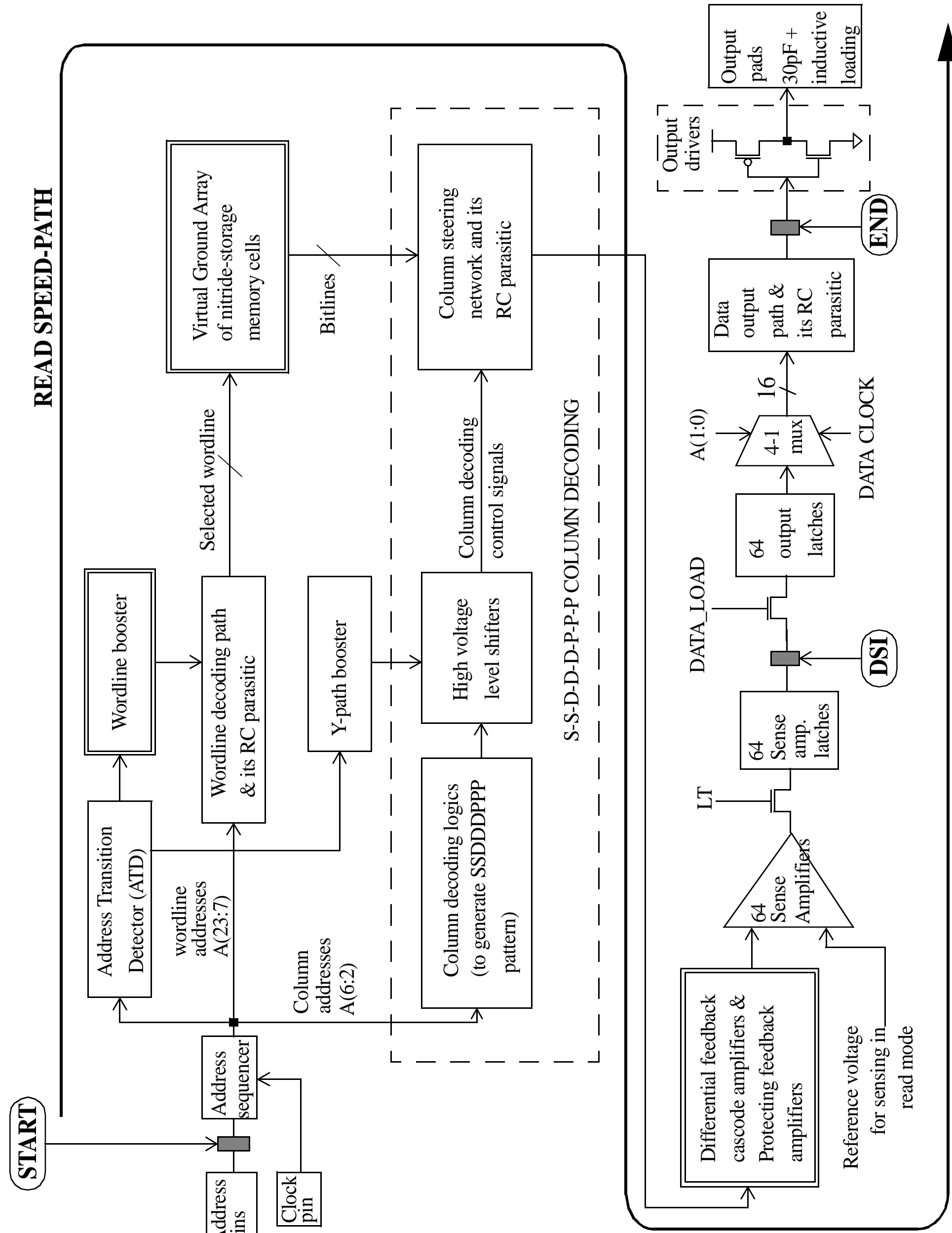


Figure 6.1: Read path simulation set-up diagram

The speed for the burst-mode read operation is usually characterized by two numbers: the *initial read access time*, measured in ns, and the *burst-mode read access speed*, measured in MHz. To start a burst-mode read operation, an initial address is applied, then after an initial delay, the clock pin can be used to clock out data sequentially. The internal address is incremented automatically every four external clock cycles. This delay time is called the *initial read access time* and it is measured from the time the new address is applied to the time the output pad (I/O pad) reaches 50% of the output rail-to-rail swing.

During the initial latency the wordline is boosted to a high voltage of about 4.8V, then the *initial internal sensing operation* begins. Note that, for each internal sensing operation, 64 bits (4 words) are sensed at the same time. When the sensing results are available, they will be latched into 64 sense amplifier latches (shown in Figure 6.1). The results are latched to free the sense amplifiers for the sensing operation of the next 64 bits. The data in the sense amplifier latches is then transferred to 64 output latches by the control signal DATA_LOAD (which can happen concurrently when LT is high). At this point the data is clocked out at each rising edge of the clock, 16 bits (1 word) at a time. The sensing operation for the next 64 bits, as well as for subsequent 64-bit groups, will be faster because the wordline is already at the required high voltage. These internal sensing operations are called *burst-mode internal sensing operations* to distinguish them from the initial internal sensing operation, in which the wordline needs to be first boosted. The architecture of the data path is a *pipeline*, meaning that at the end of an internal sensing operation, data is stored into 64 latches and the next internal sensing operation can start right away. At the same time, the words in the latches are brought to the output pads sequentially, one word for each clock cycle. After 4 cycles for 4 words, the data in the latches is exhausted and the data for the next 64 bits needs to be loaded in. This means that the burst internal sensing time for 64 bits must fit into 4 clock cycles. Thus, the faster the burst-mode internal sensing speed, the faster the overall clocking. The *burst read access speed* is the frequency of the clock such that a burst-mode internal sensing operation can finish in four cycles of that clock speed. Therefore, the burst read access speed can be computed as follows: 1) Measure the time T_{burst_sense} from the rising edge of the clock which triggers the new burst internal sensing cycle, to the time when the DSI

node (shown in Figure 6.1) reaches 50% of its final voltage. 2) Divide T_{burst_sense} by four, and take the inverse of the quotient.

Figure 6.2 presents the simulation results for the entire read path shown in Figure 6.1. These simulation results are intended to show the read access speed that is achieved when all the new sensing techniques are applied.

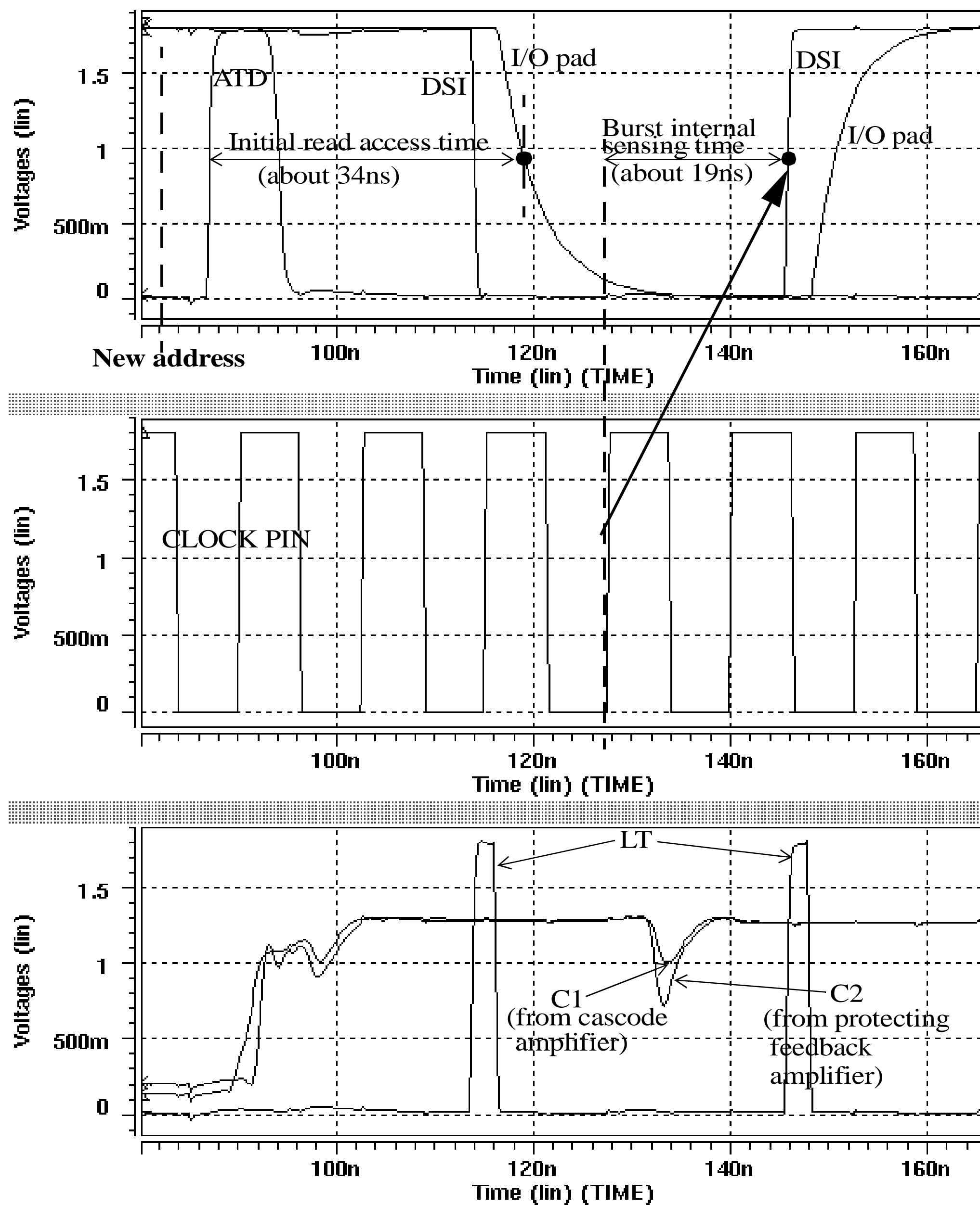


Figure 6.2: Read path simulation result

The simulations are run at VCC = 1.8V, temperature = 25°C and with typical process transistor parameters. Figure 6.2 shows one initial access, which reads a “0”, and one burst access, which reads a “1”. The new address is applied at t = 85ns, triggering the address transition detection pulse (ATD pulse) shortly after that. The initial internal sensing operation starts from the rising edge of the ATD pulse. At t = 119ns, the I/O pad crosses the 50% point (0.9V), thus the initial read access time is 34ns. For the burst-mode read access, the rising edge which triggers the internal burst sensing operation occurs at t = 127ns. The sensing data arrives at the DSI node at t = 136ns, thus the burst read access speed is:

$$1 \Big/ \left[\frac{(136ns - 127ns)}{4} \right] \sim 210MHz$$

Figure 6.2 also shows the performance of the differential feedback cascode amplifier and the protecting feedback amplifier. As shown in Figure 2.10 or Figure 2.13, the new cascode amplifier drives node C1, which is coupled to the group of three drain bitlines and the protecting feedback amplifier drives node C2, which is coupled to the group of three protecting bitlines. The bottom panel of Figure 6.2 presents the simulation waveforms for the nodes C1 and C2. The new cascode amplifier quickly raises C1 to about 1.25V when VCC=1.8V. Node C2 follows node C1 very well, an indication that the performance of the protecting feedback amplifier is good.

In Figure 6.3, the waveforms for the sensing node SAIN and the read reference are shown. Note that SAIN is the output of the differential feedback cascode amplifier and the read reference is the output of another differential feedback cascode amplifier connected to a reference memory cell. At the time LT goes up, the sensing result is transferred to the sense amplifier latch; the read voltage margin (the voltage difference between SAIN and the read reference) at this time is about 100mV, which is a very good margin, an indication that the margin loss mechanisms discussed in Chapters 2, 4 and 5 have been suppressed significantly.

Simulation results for the new wordline booster are not presented here since they have already been presented at the end of Chapter 5 (Figure 5.16).

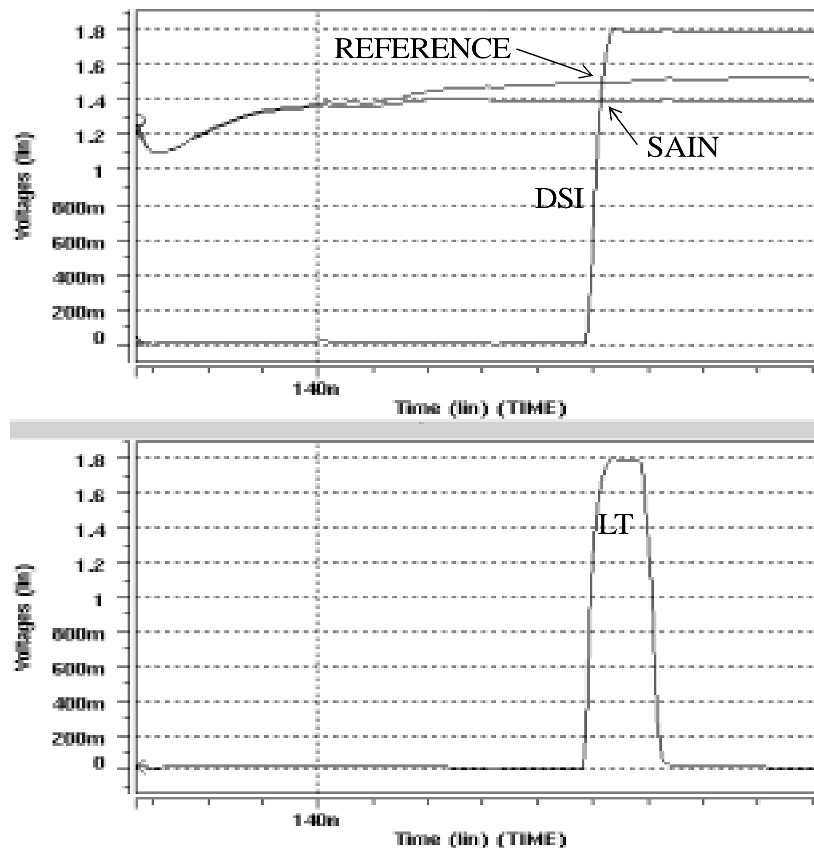


Figure 6.3: Simulation waveforms for the sensing node SAIN and the reference node

6.2 Read path measured results

A microphotograph of the 1.8V, two-bit-per-cell, 256 Mb, nitride-storage flash memory which employs all the new sensing techniques is shown in Figure 6.4. The figure shows 16 independent memory banks; the sensing and column decoding logic blocks for one bank are clearly marked. The wordline booster is common for all 16 memory banks, and is placed on the left of the chip. The area of the chip is $52mm^2$. The main input-output pins are: 24 address pins (A0 to A23), 16 I/O pins and 7 control pins.

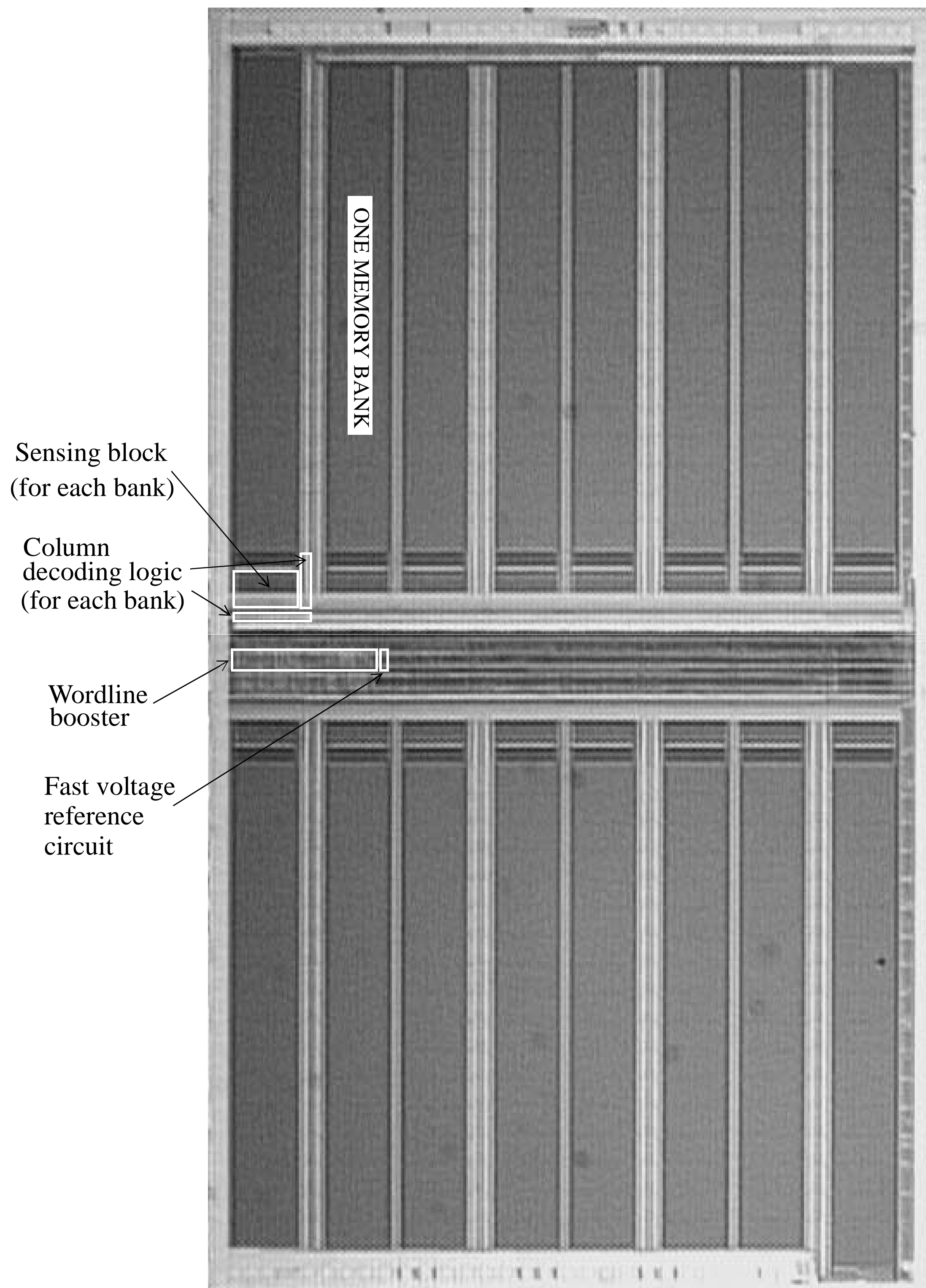


Figure 6.4: Microphotograph of the 1.8V, two-bit-per-cell, 256Mb flash memory

The set-up for measuring the performance of the memory, especially its initial read access time and burst-mode read access speed, are shown in Figure 6.5. For a burst read, the tester provides the initial address and control signals to the chip through a cable and a probe card. The specification requires a capacitive load at each I/O pad of 30pF. However, due to the parasitic capacitive loading of the cable and the probe card pins, the capacitive loading may actually exceed 30pF. Therefore, to accurate measure the read speed, the node right before the I/O pad is probed, thus the parasitic capacitance at the I/O pads does not affect the measurement result.. The probed node is node END shown in Figure 6.1; it is the input of the output driver.

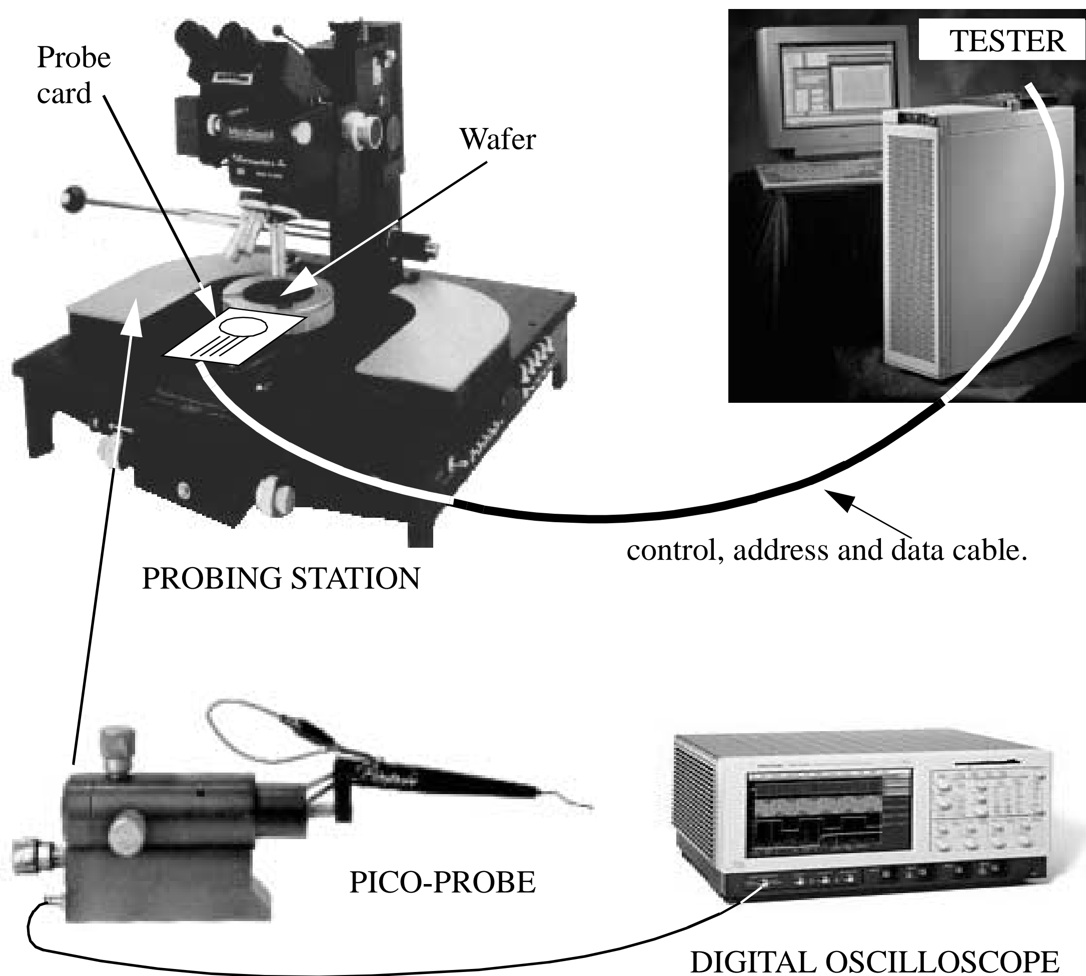


Figure 6.5: Measurement set-up

The actually initial read access time is the sum of the delay from node START to node END (Figure 6.1), and the delay of the output driver. Simulation can predict the delay of the output driver very well because this driver is just an inverter driving a 30pF load. The simulation result for the output driver is shown in Figure 6.6. The delay of the output driver is about **2.8ns**.

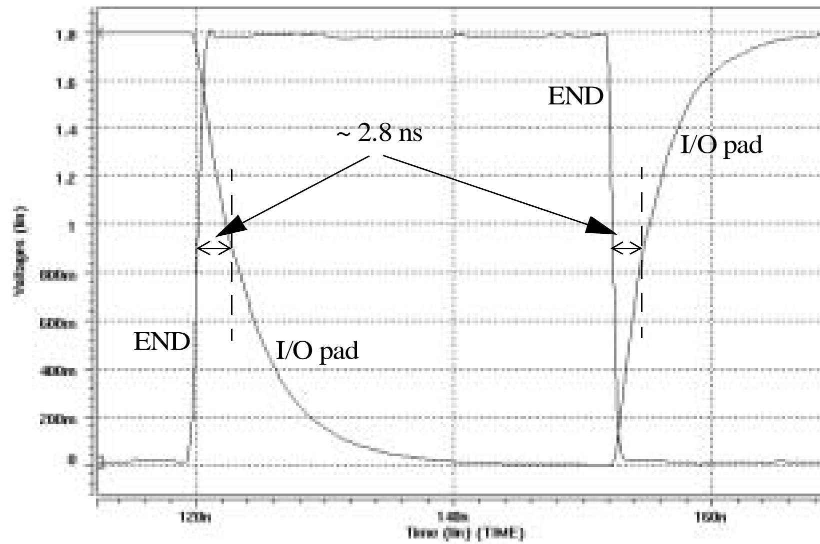
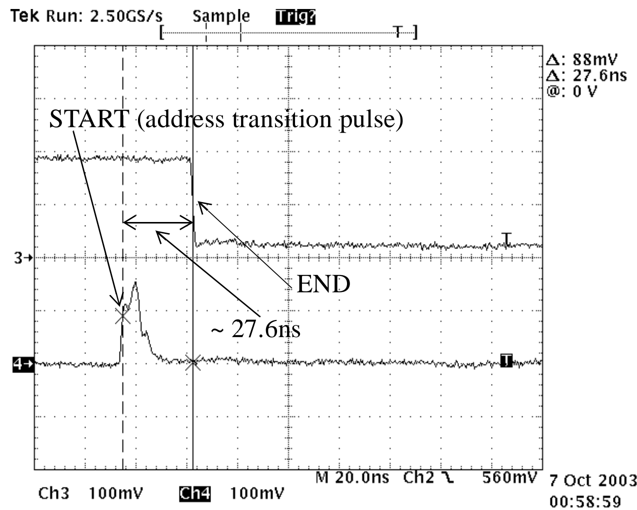


Figure 6.6: Simulation result for the output buffer

Probing is performed at room temperature, with the power supply voltage VCC=1.8V. Pico-probes are used because the parasitic capacitance of a pico-probe is very small (~ 0.1pF), thus it does not significantly affect the measurements, especially if the probed nodes are strong (low impedance nodes).

Figure 6.7 shows the delay from node START (address pin) to the node END in an initial read access. This time is about 27.6ns, thus the initial read access time is 27.6ns + 2.8ns = 30.4ns. Figure 6.8 is the measured delay from the internal clock rising edge, which initiates the burst internal sensing operation, to the time DSI (shown in Figure 6.1) switches. This time is about 19ns, thus the burst read access speed is (note that there is a 1ns-delay from the external clock to the internal clock):

$$1/\left[\frac{20ns}{4}\right] \sim 200MHz$$

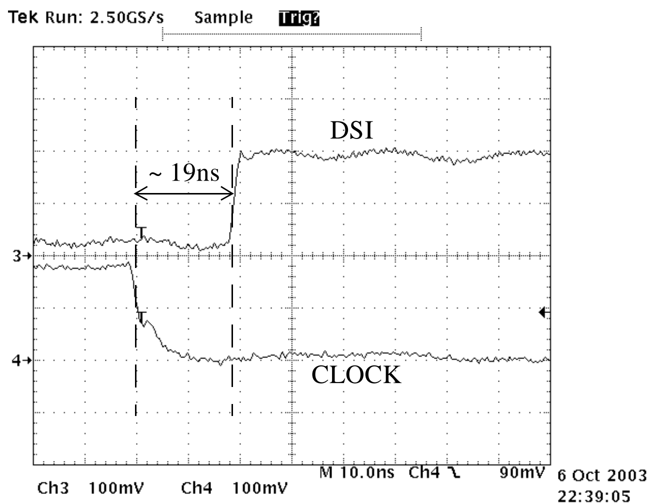


Note 1: The vertical scale is 1V/division, not 100mV/division. This is because the signals are scaled down by 10 times when pico probes are used.

Note 2: Measured speed for bit "0" of a cell, in which the other side is a "1" bit. This represent the worst case for the Complementary Bit Disturbance.

Figure 6.7: Initial read access - delay from node START to node END

As mentioned in Section 6.1, due to the data path pipeline architecture, the delay from node START to node DSI determines the burst read access speed, not the delay from node DSI to the I/O pads, as long as the delay from node DSI to the I/O pads is less than the one clock period.



Note: Measured speed for bit "1" of a cell, in which the other side is a "0" bit. This represent the worst case for the Complementary Bit Disturbance.

Figure 6.8: Burst read access - delay from clock to node DSI

As an example, if the clock period is 8ns (burst read access speed is 125MHz), then the delay from node DSI to the I/O pads must be smaller than 8ns, assuming that the system data setup time is 0ns. Many techniques can be used to shorten this delay as much as possible, such as the techniques discussed in [25], [26], and [27].

The speed measurements have shown that the memory can read out the data correctly and quickly, an indication that the new sensing techniques have been successful in recovering the read margin loss. The discrepancy between the simulation results and the measured results occur mainly due to differences between the real parasitic capacitance and resistance of the read path and the estimated parasitic values used in simulations.

Probing is also performed for other internal nodes to show the performance of the new sensing techniques. Figure 6.9 shows the waveform of node C1, which is driven by the new cascode amplifier. Recall that C1 is coupled to the three selected drain bitlines in the read mode. The measured level at C1 is about 1.25V, which is the expected value. The waveform of node C2 is also shown; this node is driven by the protecting feedback amplifier, discussed in Chapter 2. C2 is coupled to the three protecting bitlines in the read mode. The fact that C2 tracks C1 well indicates good performance of the protecting feedback amplifier.

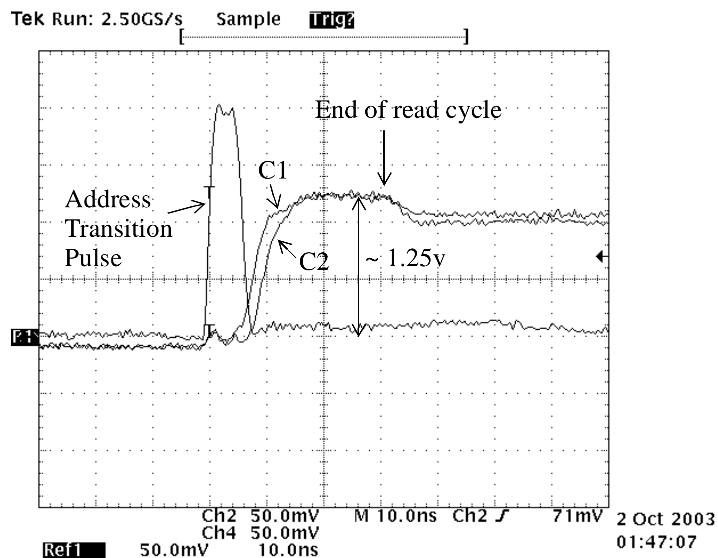


Figure 6.9: Measured performance of the differential feedback cascode amplifier and the protecting feedback amplifier

The measured read voltage margin is shown in Figure 6.10. This margin is about 100mV, which is a very healthy margin. Note that SAIN and REFERENCE nodes in Figure 6.10 are the outputs of the new cascode amplifiers.

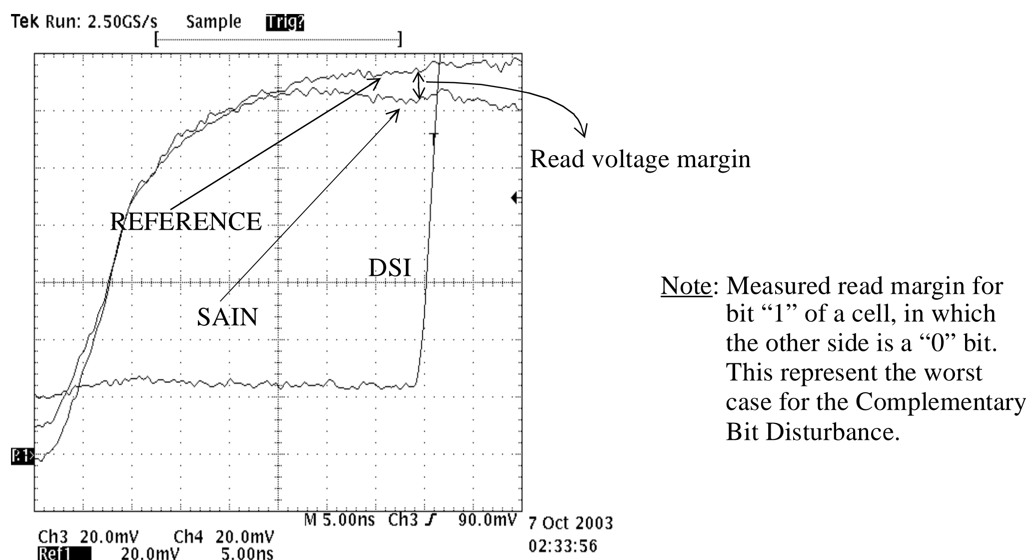


Figure 6.10: Read voltage margin

Figure 6.11 shows the measured output waveforms of the new wordline booster when the power supply voltage varies from 1.6V to 2.0V. The wordline booster output variation is about 0.3V, which is very good. Note that the fast reference voltage circuit used in the wordline booster is the one with greater speed but less accuracy; this fast reference circuit is discussed in Section 5.1.5, Chapter 5. Another detail needs to be mentioned. Due to the need to fit the layout of the wordline booster into the intended area, the default booster cell (shown in Figure 5.11, Chapter 5) has been removed. Although the boosting capacitance of this booster cell is small compared to that of the other booster cells, this causes some loss in accuracy.

The measured performance of the chip is summarized in Table 6.1. The average burst-mode read current consumption of 26mA, which includes the current consumption of the

cascode amplifiers, protecting amplifiers, sense amplifiers and other circuits such as the wordline booster, the booster for the column decoding path, etc. It also includes the average of all switching currents, such as the charging and discharging currents for long buses and other signal lines.

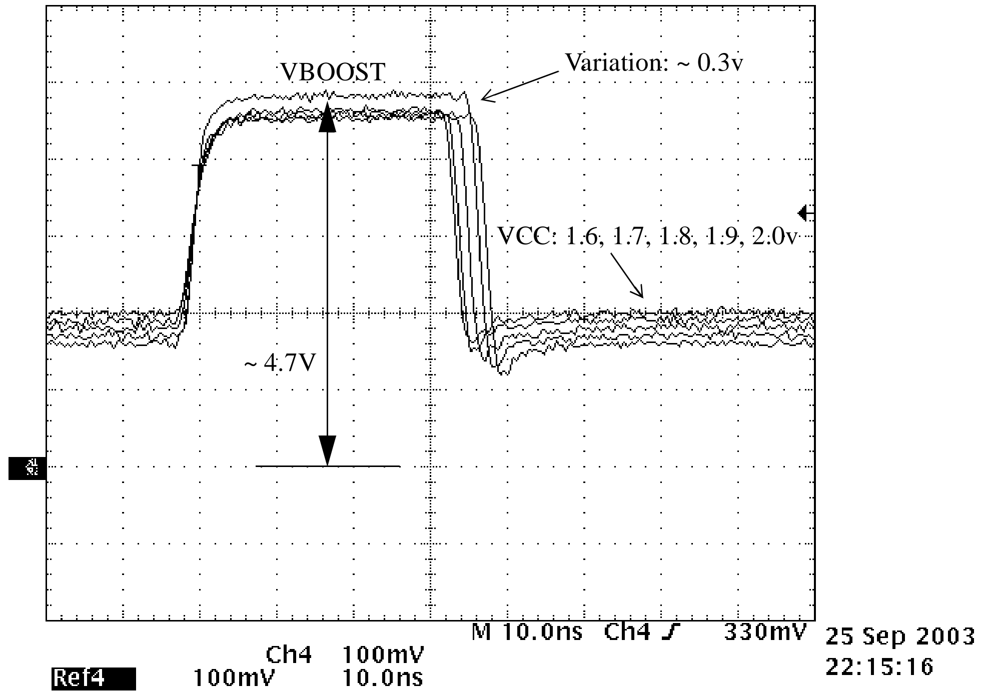


Figure 6.11: Wordline booster output level at different supply voltages

Table 6.1: Measured performance

Parameter	Value
Supply voltage	1.8 V
Active burst read current (66MHz)	26 mA
Random (initial) read access time	30.4 ns
Burst read access speed (internal)	200 MHz
Density	256 Mbit
Area	52 mm^2
Technology	0.13 µm

6.3 Summary

The simulation and measured results presented in this chapter have proved that the new sensing techniques discussed in Chapter 2 through Chapter 5 help to regain enough sensing margin for 1.8V, two-bit-per-cell nitride-storage flash memories -- the most important goal of this research. Although the results show the reading performance of the memory for a few bits, the reading operation have actually been conducted successfully on an entire sector of about 1 million bits. The data pattern stored in this sector is the checker-board pattern, in which every memory cells in the sector stores a “1” on one side and a “0” on the other side of the cell; as mentioned in Chapter 4, this data pattern represents the worst-case for the Complementary-Bit Disturbance.

The chapter also shows that the memory not only reads correctly but also read fast, exceeding the Specifications of 45ns for the initial read access and 133MHz for the burst-mode internal sensing speed.

Chapter 7

Conclusion

High density, low voltage and low cost flash memory has become very important recently due to trends of using more and more flash memory in a wide range of applications such as cellular phones and digital camcorders. Two-bit-per-cell, nitride-storage flash memory is one of the top candidates for this usage trend due to its simpler process, which leads to low cost and its capability of storing two or more bits per cell, which leads to high density. However, nitride-storage flash memory suffers from a number of read margin loss mechanisms, which sometimes are serious enough to make the read operation fail. While the 3V, two-bit-per-cell, nitride-storage flash memory works using the more traditional sensing techniques, these techniques prove to be inadequate or even stop functioning for 1.8V, two-bit-per-cell, nitride-storage flash memory operation.

This research has introduced new sensing techniques to make the first 1.8V, two-bit-per-cell, nitride-storage flash memory work despite the presence of read margin loss mechanisms, which tend to be more serious at low supply voltages and smaller memory cell dimensions.

7.1 Contributions

The research has contributed four important sensing techniques, which not only make the read operation of the 1.8V, two-bit-per-cell, nitride-storage possible, but also make its sensing operation fast; additionally it consumes much less power. Some of the new sensing techniques introduced can be applied for other types of flash memory, not only for two-bit-per-cell nitride-storage technology.

Chapter 2 introduces the first contribution, which is the Sense Current Recovery Technique. This technique is very effective in eliminating significant read margin losses caused by the side-leakage current in the virtual-ground memory configuration. In a read operation, the technique uses multiple drains to recover the read margin loss for “1” bits, multiple protecting bitlines to recover the read margin loss for “0” bits, and a protecting feedback amplifier to eliminate the voltage mismatch between the drain bitlines and the protecting bitlines, making the read margin loss even smaller. Fast read speed is achieved because all decoded bitlines are driven either from the cascode amplifiers or the protecting feedback amplifiers.

The second contribution - the multiple-drain-bitline, multiple-protecting-bitline column decoding technique discussed in Chapter 3 - is essential in realizing the Sense Current Recovery Technique in which multiple drain bitlines and multiple protecting bitlines must be used to recover the read margin loss.

Chapter 4 describes the third contribution, the differential feedback cascoded bitline voltage control technique. The technique is important in reducing the read margin loss due to the disturbance of the other bit in the same memory cell. To reduce this disturbance, the bitline voltage needs to be raised as high as possible, and kept as stable as possible with process and temperature variations. The new differential feedback cascode amplifier, which is the centerpiece of this new technique, has satisfied these two difficult bitline voltage control requirements even when the power supply voltage is 1.6V, the task

that the traditional cascode amplifier fails. Additionally, the new cascode amplifier performs the tasks in a fast manner with much less power consumption. This new differential feedback cascode amplifier can certainly be applied for other flash memory configurations, especially for low voltage, low power flash memory.

The final contribution of the research is the auto-calibrated wordline voltage control, described in Chapter 5. By controlling accurately the wordline voltage, the technique helps to reduce the read margin loss caused by the cycle-induced mobility degradation in the memory cells, and also helps to ease the design of the sensing circuitry such as the cascode amplifier. This task is accomplished by using an A/D converter to measure the supply voltage and using this information to adjust the wordline voltage level. The technique also helps to generate the wordline voltage quickly, accurately with less power consumption.

7.2 Recommendations for future work

As stated at the beginning of this chapter, the dominant trend in designing flash memory is for the high density, low voltage operation, supporting growing needs for low cost memory. Thus as this trend continues to evolve, there are many directions for future research.

Scaling down the dimension of the nitride-storage memory cell and increasing the chip area are the primary ways to achieve high density and low cost. Therefore, an area for future work can be the affect of scaling the memory dimension down and increasing the chip area to the new sensing techniques presented in this dissertation. For example, when the memory cell dimensions are reduced, the side-leakage current will definitely be increased, causing more margin loss. Moreover, if the chip dimensions become larger, the global bitline resistance will go up, leading to larger side-leakage current. Thus, future

research can explore what is needed for the Sense Current Recovery Technique to preserve the read margin; this could include increasing the number of drains and protecting bitlines or the number of stages of the protecting feedback amplifier to make the side-leakage current smaller.

Another direction for future work to increase the density and lower the cost can be exploring the possibilities to store more than two bits per nitride-storage memory cell. The interaction of all these bits is complicated and special sensing techniques would be needed.

Scaling down the power supply voltage and exploring new circuit topologies to achieve low power consumption are other major areas for future work. The differential feedback cascode amplifier may continue to work at the power supply voltage of 1.5V, but beyond that, voltage headroom limitations will become a major concern and the new cascode amplifier discussed in Section 4.2 may require significant modifications.

Finally, for the auto-calibrated wordline voltage control, this is certainly an area that needs more work. Future research may focus on finding novel circuit topologies to achieve both accuracy and fast reference circuits without trading off these two aspects of performance as has been done in this dissertation.

Appendix

Verilog codes

A.1 Verilog Code for Local bitline decoding

```
module yd_secy(A, EN_SECY, SECY_LV);
    input      [6:2]   A;
    input              EN_SECY;

    output     [15:0]  SECY_LV;
    reg        [15:0]  SECY_LV;
    //------------------------------------------//

    always @ (A or EN_SECY)
    if (!EN_SECY)
        SECY_LV = 16'h0000;
```

```
else
    case (A)
        5'b00000: SECY_LV = 16'hC03F;
        5'b00001: SECY_LV = 16'h807F;
        5'b00010: SECY_LV = 16'h00FF;
        5'b00011: SECY_LV = 16'h01FE;

        5'b00100: SECY_LV = 16'h03FC;
        5'b00101: SECY_LV = 16'h07F8;
        5'b00110: SECY_LV = 16'h0FF0;
        5'b00111: SECY_LV = 16'h1FE0;

        5'b01000: SECY_LV = 16'h3FC0;
        5'b01001: SECY_LV = 16'h7F80;
        5'b01010: SECY_LV = 16'hFF00;
        5'b01011: SECY_LV = 16'hFE01;

        5'b01100: SECY_LV = 16'hFC03;
        5'b01101: SECY_LV = 16'hF807;
        5'b01110: SECY_LV = 16'hF00F;
        5'b01111: SECY_LV = 16'hE01F;
        //----------------------------------------//

        5'b10000: SECY_LV = 16'hFC03;
        5'b10001: SECY_LV = 16'hF807;
        5'b10010: SECY_LV = 16'hF00F;
        5'b10011: SECY_LV = 16'hE01F;

        5'b10100: SECY_LV = 16'hC03F;
        5'b10101: SECY_LV = 16'h807F;
        5'b10110: SECY_LV = 16'h00FF;
```

```
            5’b10111: SECY_LV = 16’h01FE;

            5’b11000: SECY_LV = 16’h03FC;
            5’b11001: SECY_LV = 16’h07F8;
            5’b11010: SECY_LV = 16’h0FF0;
            5’b11011: SECY_LV = 16’h1FE0;

            5’b11100: SECY_LV = 16’h3FC0;
            5’b11101: SECY_LV = 16’h7F80;
            5’b11110: SECY_LV = 16’hFF00;
            5’b11111: SECY_LV = 16’hFE01;
        endcase
endmodule
```

A.2 Verilog Code for Global source bitline decoding

```
module yd_s(A, EN_S, S_LV);
    input       [6:2]   A;
    input               EN_S;

    output      [7:0]   S_LV;
    reg         [7:0]   S_LV;
    //--------------------------------//

    always @ (A or EN_S)
    if (!EN_S)
        S = 8’h00;
    else
```

```
    casex (A)
        5'b0x000: S_LV = 8'hC0;
        5'b0x001: S_LV = 8'h81;
        5'b0x010: S_LV = 8'h03;
        5'b0x011: S_LV = 8'h06;

        5'b0x100: S_LV = 8'h0C;
        5'b0x101: S_LV = 8'h18;
        5'b0x110: S_LV = 8'h30;
        5'b0x111: S_LV = 8'h60;

        5'b1x000: S_LV = 8'h03;
        5'b1x001: S_LV = 8'h06;
        5'b1x010: S_LV = 8'h0C;
        5'b1x011: S_LV = 8'h18;

        5'b1x100: S_LV = 8'h30;
        5'b1x101: S_LV = 8'h60;
        5'b1x110: S_LV = 8'hC0;
        5'b1x111: S_LV = 8'h81;
    endcase
endmodule
```

A.3 Verilog Code for Global drain bitline decoding

```
module yd_d(A, EN_D, DR_LV, DL_LV, D_LV);
  input    [6:2]    A;
  input             EN_D;
```

```
output      [7:5]    DR_LV;
output      [1:0]    DL_LV;
output      [7:0]    D_LV;

reg         [7:5]    DR_LV;
reg         [1:0]    DL_LV;
reg         [7:0]    D_LV;
//-------------------------------//

always @ (A or EN_D)
    if (!EN_D)
        {DR_LV,DL_LV,D_LV}= 13'h0000;
    else
        case (A)
            5'b00000: {DR_LV,DL_LV,D_LV}= 13'h0007;
            5'b00001: {DR_LV,DL_LV,D_LV}= 13'h000E;
            5'b00010: {DR_LV,DL_LV,D_LV}= 13'h001C;
            5'b00011: {DR_LV,DL_LV,D_LV}= 13'h0038;

            5'b00100: {DR_LV,DL_LV,D_LV}= 13'h0070;
            5'b00101: {DR_LV,DL_LV,D_LV}= 13'h00E0;
            5'b00110: {DR_LV,DL_LV,D_LV}= 13'h00C1;
            5'b00111: {DR_LV,DL_LV,D_LV}= 13'h0083;

            5'b01000: {DR_LV,DL_LV,D_LV}= 13'h0007;
            5'b01001: {DR_LV,DL_LV,D_LV}= 13'h000E;
            5'b01010: {DR_LV,DL_LV,D_LV}= 13'h001C;
            5'b01011: {DR_LV,DL_LV,D_LV}= 13'h0038;

            5'b01100: {DR_LV,DL_LV,D_LV}= 13'h0070;
```

```
            5'b01101: {DR_LV,DL_LV,D_LV}= 13'h00E0;
            5'b01110: {DR_LV,DL_LV,D_LV}= 13'h01C0;
            5'b01111: {DR_LV,DL_LV,D_LV}= 13'h0380;
            //------------------------------------------//

            5'b10000: {DR_LV,DL_LV,D_LV}= 13'h1C00;
            5'b10001: {DR_LV,DL_LV,D_LV}= 13'h1801;
            5'b10010: {DR_LV,DL_LV,D_LV}= 13'h1003;
            5'b10011: {DR_LV,DL_LV,D_LV}= 13'h0007;

            5'b10100: {DR_LV,DL_LV,D_LV}= 13'h000E;
            5'b10101: {DR_LV,DL_LV,D_LV}= 13'h001C;
            5'b10110: {DR_LV,DL_LV,D_LV}= 13'h0038;
            5'b10111: {DR_LV,DL_LV,D_LV}= 13'h0070;

            5'b11000: {DR_LV,DL_LV,D_LV}= 13'h00E0;
            5'b11001: {DR_LV,DL_LV,D_LV}= 13'h00C1;
            5'b11010: {DR_LV,DL_LV,D_LV}= 13'h0083;
            5'b11011: {DR_LV,DL_LV,D_LV}= 13'h0007;

            5'b11100: {DR_LV,DL_LV,D_LV}= 13'h000E;
            5'b11101: {DR_LV,DL_LV,D_LV}= 13'h001C;
            5'b11110: {DR_LV,DL_LV,D_LV}= 13'h0038;
            5'b11111: {DR_LV,DL_LV,D_LV}= 13'h0070;
        endcase
endmodule
```

A.4 Verilog Code for Global protecting bitline decoding

```
module yd_p(A, EN_P, PR_LV, PL_LV, P_LV);
    input       [6:2]   A;
    input               EN_P;

    output      [7:2]   PR_LV;
    output      [4:0]   PL_LV;
    output      [7:0]   P_LV;

    reg         [7:2]   PR_LV;
    reg         [4:0]   PL_LV;
    reg         [7:0]   P_LV;
    //---------------------------------//

    always @ (A or EN_P)
        if (!EN_P)
            {PR_LV,PL_LV,P_LV} = 19'h00000;
        else
            case (A)
                5'b00000: {PR_LV,PL_LV,P_LV} = 19'h00038;
                5'b00001: {PR_LV,PL_LV,P_LV} = 19'h00070;
                5'b00010: {PR_LV,PL_LV,P_LV} = 19'h000E0;
                5'b00011: {PR_LV,PL_LV,P_LV} = 19'h000C1;

                5'b00100: {PR_LV,PL_LV,P_LV} = 19'h00083;
                5'b00101: {PR_LV,PL_LV,P_LV} = 19'h00007;
                5'b00110: {PR_LV,PL_LV,P_LV} = 19'h0000E;
                5'b00111: {PR_LV,PL_LV,P_LV} = 19'h0001C;

                5'b01000: {PR_LV,PL_LV,P_LV} = 19'h00038;
```

```
            5'b01001: {PR_LV,PL_LV,P_LV} = 19'h00070;
            5'b01010: {PR_LV,PL_LV,P_LV} = 19'h000E0;
            5'b01011: {PR_LV,PL_LV,P_LV} = 19'h001C0;

            5'b01100: {PR_LV,PL_LV,P_LV} = 19'h00380;
            5'b01101: {PR_LV,PL_LV,P_LV} = 19'h00700;
            5'b01110: {PR_LV,PL_LV,P_LV} = 19'h00E00;
            5'b01111: {PR_LV,PL_LV,P_LV} = 19'h01C00;
            //-------------------------------------------//

            5'b10000: {PR_LV,PL_LV,P_LV} = 19'h0E000;
            5'b10001: {PR_LV,PL_LV,P_LV} = 19'h1C000;
            5'b10010: {PR_LV,PL_LV,P_LV} = 19'h38000;
            5'b10011: {PR_LV,PL_LV,P_LV} = 19'h70000;

            5'b10100: {PR_LV,PL_LV,P_LV} = 19'h60001;
            5'b10101: {PR_LV,PL_LV,P_LV} = 19'h40003;
            5'b10110: {PR_LV,PL_LV,P_LV} = 19'h00007;
            5'b10111: {PR_LV,PL_LV,P_LV} = 19'h0000E;

            5'b11000: {PR_LV,PL_LV,P_LV} = 19'h0001C;
            5'b11001: {PR_LV,PL_LV,P_LV} = 19'h00038;
            5'b11010: {PR_LV,PL_LV,P_LV} = 19'h00070;
            5'b11011: {PR_LV,PL_LV,P_LV} = 19'h000E0;

            5'b11100: {PR_LV,PL_LV,P_LV} = 19'h000C1;
            5'b11101: {PR_LV,PL_LV,P_LV} = 19'h00083;
            5'b11110: {PR_LV,PL_LV,P_LV} = 19'h00007;
            5'b11111: {PR_LV,PL_LV,P_LV} = 19'h0000E;
        endcase
endmodule
```

Bibliography

[1] M. Bauer, et al., “A Multilevel-cell 32Mb flash memory,” ISSCC Digest of Technical papers, pp. 132-133, Feb. 1995.

[2] Tae-Sung Jung, et al., “A 3.3V 128Mb Multi-Level NAND Flash Memory for Mass Storage Applications,” ISSCC Digest of Technical papers, pp. 32-33, Feb. 1996.

[3] D. Elmhurst, et al., “A 1.8V 128Mb 125MHz Multi-level cell Flash Memory with Flexible Read While Write,” ISSCC Digest of Technical papers, pp. 286-287, Feb. 2003.

[4] Eduardo Maayan, et al., “A 512Mb NROM Flash Data Storage Memory with 8MB/s Data Rate,” ISSCC Digest of Technical papers, pp. 75-76, Feb. 2002.

[5] Binh Quang Le, et al., “Drain side sensing scheme for virtual ground flash eprom array with adjacent bit charge and hold,” US Patent No. 6510082, Jan. 2003.

[6] Nobuhiko Ishizuka, “Semiconductor Memory,” US Patent No. 5875128, Jun. 1997.

[7] B. G. Streetman and S. Banerjee, “Solid State Electronic Devices,” 5th ed., Prentice Hall, 2000, pp. 245-246.

[8] Richard S. Muller and Theodore I. Kamins, “Device Electronics for Integrated Circuits,” 2nd ed., Wiley, 1986, pp. 584-586.

[9] D. Mills, et al., “A 3.3V 50MHz Synchronous 16Mb Flash Memory,” ISSCC Digest of Technical papers, pp. 120-121, Feb. 1995.

[10] P. R. Gray and R. G. Meyer, “Analysis and Design of Analog Integrated Circuits,” 3rd ed., Wiley, 1993, pp. 584-586.

[11] Behzad Razavi, “Design of Analog CMOS Integrated Circuits,” McGraw-Hill, 2001, pp. 307-309.

[12] P. R. Gray and R. G. Meyer, “MOS operational amplifier design: A tutorial overview ,” IEEE Journal of Solid-State Circuits, vol. 17, pp. 969-982, Dec. 1982.

[13] B. A. Wooley, “MOS Operational Amplifiers,” EE315 course notes, Stanford University, CA, 1998.

[14] Behzad Razavi, “Design of Analog CMOS Integrated Circuits,” McGraw-Hill, 2001, pp. 355-361.

[15] G. F. Franklin, J. D. Powell and A. Emami-Naeini, “Feedback Control of Dynamic Systems,” 3rd ed., Addison Wesley, 1994, pp. 243-440.

[16] S. Ali, et al., “A new staggered Virtual Ground array architecture implimented in a 4Mb CMOS EPROM,” VLSI Circuits Digest of Technical papers, pp. 35-36, May 1989.

[17] Boaz Eitan, et al., “Alternate Metal Virtual Ground (AMG) - A New Scaling Concept for Very High-Density EPROM’s,” IEEE Electron Devices Letters, vol. 12, no. 8, pp. 450-452, Aug. 1991.

[18] Loc Hoang, et al., “A 65ns 1Mb CMOS alternate metal virtual ground eprom with dual reference sensing scheme and word line voltage regulator,” 1993 International Symposium on VLSI Technology, Systems, and Applications, Proceedings of Technical Papers, pp. 336-338, May 1993.

[19] Yoshimitsu Yamauchi, et al., “A New Cell Structure for Sub-quarter Micron High Density Flash Memory,” International Electron Devices Meeting Tech. Dig., pp. 267-270, Dec. 1995.

[20] Kurihara, et al., “Decoder apparatus and methods for pre-charging bit lines,” US Patent No. 6525969, Feb. 2003.

[21] Behzad Razavi, “Design of Analog CMOS Integrated Circuits,” McGraw-Hill, 2001, pp. 53-55.

[22] J. C. Chen, et al., “A 2.7V only 8Mbx16 NOR Flash Memory,” 1996 Symposium on VLSI Circuits, pp. 172-173, 1996.

[23] Behzad Razavi, “Design of Analog CMOS Integrated Circuits,” McGraw-Hill, 2001, pp. 385-386.

[24] B. A. Wooley, “Voltage comparators,” EE315 course notes, Stanford University, CA, 1998.

[25] J. Rabaey, “Digital Integrated Circuits, A Design Perspective,” Prentice-Hall, 1996.

[26] I. Sutherland, B. Sproull, and D. Harris, “Logical Effort: Designing Fast CMOS Circuits,” Morgan Kaufmann, 1999.

[27] Srividya Srinivasaraghavan and Wayne Burleson, “Interconnect Effort - A Unification of Repeater Insertion and Logical Effort,” Proceedings of the IEEE Computer Society Annual Symposium on VLSI, pp. 55-61, Feb. 2003.