University of Alberta

Design, Implementation and Testing of a Multilevel DRAM with Adjustable Cell Capacity

by

Yunan Xiang ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science

Department of Electrical & Computer Engineering

Edmonton, Alberta
Spring 2002

0-612-69779-7

University of Alberta

Library Release Form

Name of Author: Yunan Xiang

Title of Thesis: Design, Implementation and Testing of a Multilevel DRAM with Adjustable Cell Capacity

Degree: Master of Science

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Yunan Xiang
Apt. 507, 11147-82 Ave.
Edmonton, AB   T6G 0T5

Apr. 15, 2002

Date

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Design, Implementation and Testing of a Multilevel DRAM with Adjustable Cell Capacity* by *Yunan Xiang* in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor: Dr. Bruce F. Cockburn

D G Elliott

Supervisor: Dr. Duncan G. Elliott

Professor: Dr. Xiaoling Sun

Doug Gingrich

Professor: Dr. Douglas M. Gingrich

Apr. 16, 2002

Date

ABSTRACT

By storing more than one bit per memory cell, MultiLevel Dynamic Random-Access Memory (MLDRAM) explores an additional dimension to increase the per-cell storage capacity over conventional two-level DRAM. A well-balanced and robust MLDRAM scheme was proposed previously by Birk, Elliott and Cockburn. We designed and implemented a test chip for this MLDRAM in TSMC's 0.18-micron CMOS technology. The test chip has an adjustable cell capacity that can be selected from 2, 3, 4 and 6 levels per cell, corresponding to 1, 1.5, 2 and 2.5 bits per cell. Prototypes of the test chip were verified using an Agilent 81200 digital IC tester. Most of the cells in operational chips were found to work. However, small voltage offsets affecting the signal and reference cells cause read errors for some cells. A follow-up project would be to characterize the offset problem in greater detail and to design an improved test chip.

ACKNOWLEDGEMENTS

This research was supported by Micronet R&D, MOSAID Technologies Inc., ATMOS Corporation, the Canadian Microelectronics Corporation, the Natural Sciences and Engineering Research Council of Canada, and the University of Alberta. I would like to thank everybody who offered technical support and helpful suggestions for this project. Especially, I would like to express my great gratitude to my supervisors Bruce Cockburn and Duncan Elliott for providing me with this opportunity to study memories and for all they have taught me. Also, I would like to thank John Koob and Tyler Brandon for helping me with the workstations and the CAD tools. The test chip layout design and the evaluation of the silicon prototypes were both teamwork in the VLSI lab at the University of Alberta. Participants in the chip design and layout were Tyler Brandon, Dan Leder, and Julien Lamoureux. Participants in the chip prototype testing were Michael Hume, Tyler Brandon, Dan Leder, Craig Joly and Michael Redeker.

Table of Contents

Chapter 1 Introduction 1

1.1 Semiconductor Memory 1

1.2 DRAM Technology 3

1.3 MLDRAM Technology 6

1.4 Thesis Outline 9

Chapter 2 Previous Work 11

2.1 Basic DRAM Circuits 11

2.1.1 The DRAM Cell Array 12

2.1.1.1 Rows and Columns 12

2.1.1.2 Folded Bitlines and Twisted Bitlines 13

2.1.1.3 Friendly Cells and Dummy Cells 15

2.1.2 The DRAM Core 16

2.1.2.1 The Sense Amplifier 17

2.1.2.2 Data Buses 20

2.1.3 The DRAM Periphery 21

2.1.4 DRAM Floorplanning 22

2.2 DRAM Operations 23

2.3 Embedded DRAMs 26

2.4 MLDRAM Circuits and Techniques 27

2.4.1 Furuyama's MLDRAM 28

2.4.2 Gillingham's MLDRAM 31

2.4.3 Birk's MLDRAM 35

2.4.3.1 Circuitry 35

2.4.3.2 Operations 37

2.4.3.3 Increasing the Number of Signal Levels 41

2.4.3.4 Advantages and Disadvantages 42

2.5 MLDRAM Challenges 43

Chapter 3 Test Chip Design 44

3.1 Test Chip Design Overview 44

3.2 The Design Flow 47

3.3 Detailed Design 50

3.3.1 The Core 52

3.3.1.1 The Memory Cells..52

3.3.1.2 Friendly Cells and Dummy Cells...53

3.3.1.3 The Sense Amplifiers and Sub-bitline Connections 53

3.3.1.4 Boost Circuits 55

3.3.1.5 Row and Column Address Scrambling 56

3.3.2 The Periphery..57

3.3.2.1 Row Access 58

3.3.2.2 Column Access 60

3.3.2.3 Data IO 60

3.3.2.4 Chip Pads..61

3.4 Chip Packaging and Fixturing 63

Chapter 4 Design Verification and Simulation 65

4.1 Design Verification 65

4.2 Schematic Simulation Results...66

4.2.1 Circuit Simulation Background 66

4.2.2 Simulation Results 67

4.2.2.1 Simulations for 6-levels-per-cell Operation 69

4.2.2.2 Simulations for 4-levels-per-cell and 3-levels-per-cell Operation...75

4.2.2.3 Simulations for 2-levels-per-cell Operation 84

4.2.3 Discussion..87

Chapter 5 Test Chip Evaluation 90

5.1 The Tester and the DUT Board Interface...................................90

5.2 Test Results....94

5.2.1 Test Vector Generation....95

5.2.2 A Basic Functional Test....97

5.2.3 A Cell Retention Time Test....99

5.3 Conclusions....101

Chapter 6 Conclusions....102

6.1 Chip Design and Simulation....102

6.2 Chip Evaluation....103

6.3 Future Work....103

References....105

Appendix A Test Results for Chip 2....107

Appendix B Test Chip Bitmaps....113

List of Tables

Table 2-1. Data conversion table 30

Table 2-2. Restore voltage conversion 35

Table 2-3. Reference generation for Birk's MLDRAM 40

Table 2-4. Reference generation for a 5-levels-per-cell MLDRAM 42

Table 3-1. Cell capacities using cell pairs 44

Table 3-2. Configuration of the VDC reference generation voltage sources 45

Table 3-3. Assignment of waveforms to the reference and generate wordlines 59

Table 3-4. Column and data bus decoding in one section 61

Table 3-5. Chip pads 63

Table 4-1. Comparison of voltage signals for 4-levels-per-cell operation 87

Table 4-2. Output data conversion table 89

Table 5-1. ML5 pin list 93

Table 5-2. Percentage of good cells 99

List of Figures

Figure 1-1. A 1T-1C DRAM cell 3

Figure 1-2. DRAM charge sharing 4

Figure 1-3. A DRAM block 6

Figure 1-4. A two-bit-per-cell storage scheme 8

Figure 2-1. A DRAM cell array 12

Figure 2-2. Open-bitline and folded-bitline array architectures 13

Figure 2-3. Bitline twisting schemes 14

Figure 2-4. DRAM core circuitry 16

Figure 2-5. Standard sense amplifier block 18

Figure 2-6. I/O transistor connections 19

Figure 2-7. The basic sense amplifier 19

Figure 2-8. Wordline driver from ATMOS Corporation 21

Figure 2-9. Floorplan of a 256-Mb DDR-SDRAM 23

Figure 2-10. Schematic of a DRAM bitline pair 24

Figure 2-11. Basic read/write timing diagram for the DRAM core 25

Figure 2-12. The schematic and layout of ATMOS' SoC-RAM cell 27

Figure 2-13. Furuyama's basic sub-bitline block design 28

Figure 2-14. The sensing operation in Furuyama's scheme 29

Figure 2-15. Timing diagram for Furuyama's MLDRAM scheme 30

Figure 2-16. Schematic of Gillingham's MLDRAM 32

Figure 2-17. Timing diagram for Gillingham's MLDRAM 33

Figure 2-18. Sub-bitline designs for Birk's MLDRAM 36

Figure 2-19. Schematic of Birk's MLDRAM 37

Figure 2-20. Timing diagram for Birk's MLDRAM 38

Figure 2-21. Sub-bitlines with memory cells connected for restoring 39

Figure 2-22. Sub-bitlines with memory cells connected for reference generation 39

Figure 2-23. Sub-bitlines with memory cells connected for sensing 41

Figure 3-1. The basic 5-by-5 sub-bitline array of the test chip 46

Figure 3-2. Simplified test chip architecture 47

Figure 3-3. Test chip design flow 48

Figure 3-4. Test chip floorplan 50

Figure 3-5. Friendly cells in a memory array 53
Figure 3-6. Sub-bitline connections between memory arrays 55
Figure 3-7. Row and column address mapping in one section 56
Figure 3-8. Cell layout in one section 57
Figure 3-9. Wordline access waveforms 58
Figure 4-1. Stimulus waveforms for 6-levels-per-cell operation 68
Figure 4-2. Bitline and reference signals in 6-levels-per-cell operation 72
Figure 4-3. Simulation waveforms for data = "00000" 72
Figure 4-4. Simulation waveforms for data = "10000" 73
Figure 4-5. Simulation waveforms for data = "11000" 73
Figure 4-6. Simulation waveforms for data = "11100" 74
Figure 4-7. Simulation waveforms for data = "11110" 74
Figure 4-8. Simulation waveforms for data = "11111" 75
Figure 4-9. Stimulus waveforms for 4-levels-per-cell operation 77
Figure 4-10. Bitline and reference signals in 4-levels-per-cell operation 78
Figure 4-11. Simulation waveforms for data = "000" 78
Figure 4-12. Simulation waveforms for data = "100" 79
Figure 4-13. Simulation waveforms for data = "110" 79
Figure 4-14. Simulation waveforms for data = "111" 80
Figure 4-15. Stimulus waveforms for 3-levels-per-cell operation 81
Figure 4-16. Bitline and reference signals in 3-levels-per-cell operation 82
Figure 4-17. Simulation waveforms for data = "00" 82
Figure 4-18. Simulation waveforms for data = "10" 83
Figure 4-19. Simulation waveforms for data = "11" 83
Figure 4-20. Stimulus waveforms for 2-levels-per-cell operation 85
Figure 4-21. Bitline and reference signals in 2-levels-per-cell operation 85
Figure 4-22. Simulation waveforms for data = "0" 86
Figure 4-23. Simulation waveforms for data = "1" 86
Figure 5-1. Agilent 81200 tester configuration 91
Figure 5-2. Data segment structure 95
Figure 5-3. Data sequence, blocks and segments 96
Figure 5-4. Basic functional test for 6-levels-per-cell operation 97
Figure 5-5. A gray-scale bitmap for section A in chip 2 98
Figure 5-6. Diagram of cell leakage currents 100

Figure B-1. Bitmap for chip 2 section A 113
Figure B-2. Bitmap for chip 2 section B 113
Figure B-3. Bitmap for chip 2 section E 114
Figure B-4. Bitmap for chip 3 section A 114
Figure B-5. Bitmap for chip 3 section B 114
Figure B-6. Bitmap for chip 3 secticn E 115
Figure B-7. Bitmap for chip 5 section A 115
Figure B-8. Bitmap for chip 5 section B 115
Figure B-9. Bitmap for chip 5 section E 116

List of Abbreviations

ASIC	Application-Specific Integrated Circuit.
BL	Bitline in a memory array.
C_b	Bitline parasitic capacitance.
C_c	Cell capacitance of a DRAM cell.
CMC	Canadian Microelectronics Corporation.
CMOS	Complementary Metal Oxide Semiconductor. Refers to a MOS process or circuit which consists of both n-type and p-type MOS transistors.
DDR	Double Data Rate, e.g. DDR DRAM.
DRAM	Dynamic Random-Access Memory.
DRC	Design Rules Check.
DUT	Device Under Test.
EEPROM	Electrically-Erasable Programmable Read-Only Memory.
EPROM	Erasable-Programmable Read-Only Memory.
FeRAM	Ferroelectric Random-Access Memory.
GW	Generate Wordline.
HDRAM	An embedded DRAM design from MOSAID Technologies Inc. that used a pure logic process.
IC	Integrated Circuit.
IP	Intellectual Property.
LSB	Least Significant Bit.
LVS	Layout Versus Schematic.
MLDRAM	MultiLevel Dynamic Random-Access Memory.
MSB	Most Significant Bit.
NMOS	N-type Metal Oxide Semiconductor transistor type.
PMOS	P-type Metal Oxide Semiconductor transistor type.
PROM	Programmable Read-Only Memory.
RAM	Random-Access Memory.
ROM	Read-Only Memory.
RW	Reference Wordline.
SA	Sense Amplifier.
SDRAM	Synchronous DRAM.

SRAM	Static Random-Access Memory.
SoC-RAM	System-on-a-Chip Random-Access Memory, an embedded DRAM design from ATMOS Corporation.
TSMC	Taiwan Semiconductor Manufacturing Corporation.
UVEPROM	Erasable-Programmable Read-Only Memory that uses Ultraviolet light to erase the memory.
V_{BB}	Back Bias Voltage for memory array substrate.
V_{BLP}	Bitline Precharge Voltage.
V_{CP}	Cell Plate Voltage.
V_{DD}	Power Supply Voltage.
V_{SS}	A common ground voltage, usually set to 0 V.
V_{signal}	Differential signal voltage seen by the SA during sensing operation.
V_{cell}	Initial cell storage voltage in a DRAM cell.
WL	Wordline in a memory array.

Chapter 1 Introduction

Semiconductor memories are a vital microelectronic component in digital logic systems. From computers to microprocessor-based applications, from stand-alone memory chips to embedded intellectual property (IP) blocks in application-specific integrated circuit (ASIC) designs, memory allows data and other digital information to be stored and later retrieved. By using only one access transistor and one cell capacitor to form each storage cell, DRAM is often the most cost-effective solution for high-density semiconductor storage and is, in fact, the most widely used semiconductor memory type. DRAM revenue was $11.2 billon in 2001. The major applications for DRAMs include main memory in computers, consumer products such as mobile systems, video and games (multimedia), communication systems and industrial applications such as medical systems and embedded controllers. Tremendous advances in DRAM process technology have reduced the minimum feature size dramatically. Indeed, the storage density of DRAM chips has been approximately quadrupling every three years for the past 30 years [1]. Currently available DRAM capacity is up to 512 Mb. 1 Gb DRAM has been introduced by some DRAM manufacturers and should be available soon.

Multilevel DRAM differs from conventional DRAM by storing more than one bit per storage cell. In this chapter, we will give a basic introduction to semiconductor memories, including both conventional DRAM and multilevel DRAM. This thesis explores the architecture and circuit design, as well as chip design, implementation and testing of an MLDRAM with adjustable cell capacity, which was intended to facilitate the characterization of a promising MLDRAM scheme that was developed as part of Gershom Birk's M.Sc. thesis research [2] .

1.1 Semiconductor Memory

Semiconductor memory devices are generally categorized into two types: volatile memories and nonvolatile memories. Memories that retain stored data only when the power is applied are called *volatile memories* and those memories that retain data even when the power is turned off are called *nonvolatile memories* [1].

Nonvolatile memory technologies include mask-programmable *read-only memories* (ROM) in which the data are written permanently during manufacturing; *user-programmable ROMs* (PROMs) in which data can be written only once; and erasable and programmable ROMs in which stored data is erasable and re-programmable (EPROM). UVEPROMs (or EPROMs) are erased by removing the memories from the target system and exposing the upper silicon surface to ultraviolet light. Electrically-erasable PROMs (EEPROMs) are electrically erasable to permit in-circuit programming. *Flash memory* is another kind of electrically erasable and reprogrammable nonvolatile memory technology in which the contents of blocks of cells can be erased simultaneously through the use of an electrical erase signal. Flash memory is also two to three times denser than EEPROM. *Ferroelectric memory* (FeRAM) is a relatively new type of non-volatile memory to watch. It uses ferroelectric material to build cell capacitors. A ferroelectric material has the property that a remnant electric dipole polarization is imprinted in the crystal structure. The polarization can be changed by applying a sufficiently strong electric field. FeRAM potentially has many advantages such lower power consumption and higher writing speed than flash memory, but it is currently handicapped by relatively low cell density and many challenges related to processing [3].

Volatile memories are usually called *random-access memories* (RAMs). Data can be read from and written to any specific memory location in the same amount of time and without affecting any other stored data. *Static random-access memories* (SRAMs) store data into a cell by setting the state of a bi-stable flip-flop. As long as the power is applied and no write signals are received, data are retained. *Dynamic random-access memories* (DRAMs), however, store data by charging a capacitor to one of two possible voltages. The electrical charges stored on the capacitors have to be refreshed periodically to prevent loss of data due to leakage current from the storage capacitors to their surroundings. Because of the active drive capability of SRAM cells, SRAMs are inherently faster than DRAMs. SRAM is normally used to provide cache memory or embedded memory in high-performance applications. DRAMs, with only a capacitor for data storage in each cell, require roughly four to six times less silicon area per cell than SRAMs and are the cheapest and densest semiconductor memory. They have the largest sales volumes among memories and are widely used as computer system main memory as well as in graphics systems, consumer products, communication systems, and industrial applications. Historically, DRAMs have been the major technology driving force for the semiconductor industry in many aspects from process development, to production improvement, to reliability control and improvement. DRAM's role as technology pioneer is now less prominent, however, as DRAM processes have increasingly diverged from mainstream logic processes.

1.2 DRAM Technology

Conventional DRAM stores one bit of data, either logic 0 or logic 1, in each memory cell. The two possible values of one bit of binary data correspond to two analog voltage levels, which are usually called V_{SS} (ground) and V_{DD} (power supply voltage). A modern DRAM memory cell, shown in Figure 1-1, contains only one transistor and one capacitor, and is thus referred to as a *1T-1C cell.* The transistor operates as a switch in between the bitline and cell capacitor, whose state is controlled by the wordline signal. This transistor is usually called the *cell access transistor* and it can be of either NMOS or PMOS type. When the cell access transistor turns on, charge can be transferred between the bitline and the capacitor as the voltages equalize. The side of the capacitor connected to the drain of the cell access transistor is called the *storage node*; the other capacitor node, which is shared by all memory celi capacitors, is called the *cell plate* (CP). The capacitance of the cell capacitor ranges from 20 fF to 60 fF in a typical DRAM process technology [4].

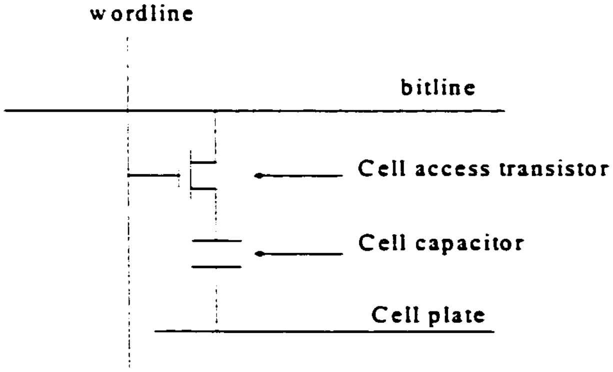


Figure 1-1. A 1T-1C DRAM cell

Digital information is encoded as analog cell voltages into two possible ways. In the *positive logic encoding*, logical 0s and 1s are stored as V_{SS} and V_{DD}, respectively.

In the *negative logic encoding,* logical 0s and 1s are stored as V_{DD} and V_{SS}. Depending on the analog voltage data on the bitline, the charge stored in the cell capacitor is given by either

$$Q = (V_{DD} - V_{CP}) \bullet C \text{, or}$$

$$Q = (V_{SS} - V_{CP}) \bullet C, \qquad (1.1)$$

where C is the cell capacitance in Farads, V_{DD} and V_{SS} are the two possible driven analog voltage levels on the bitline, and V_{CP} is the cell plate voltage. After the cell capacitor has been isolated from the bitline, the charge stored in it depletes slowly due to the various leakage paths from the storage node. In order to restore the strength of the charge in the capacitor and thus maintain the stored data, sufficiently frequent refresh (i.e. re-write) operations are necessary. Because refresh operations are required, DRAM is called a dynamic memory technology, as opposed to a static one.

To write data into the memory cell, the bitline is driven to the corresponding analog voltage by a write driver buffer (not shown in Figure 1-1) and the wordline is asserted to turn on the cell access transistor. The cell storage node is therefore connected to the bitline and is rapidly driven to the bitline voltage. Then the wordline is de-asserted to turn off the cell access transistor, disconnecting the bitline and the cell storage node and trapping the desired voltage-encoded data on the cell capacitor.

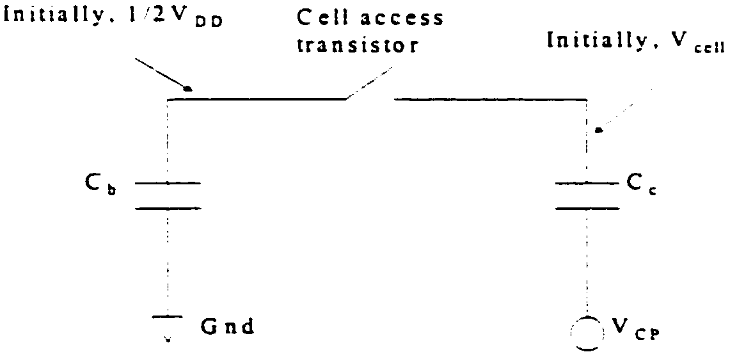


Figure 1-2. DRAM charge sharing

To read the data from the memory cell, the bitline is first precharged to a voltage midway between the two data signal voltages. The midway voltage, which is 1/2 V_{DD} if $V_{SS} = 0$, is called the *bitline precharge voltage* (V_{BLP}). With the bitline floating at the precharge voltage, the wordline is asserted and the cell access transistor turns on to connect the cell capacitor to the bitline, causing charge sharing to occur between the cell capacitor and the bitline parasitic capacitance, as shown in Figure 1-2. In a DRAM cell array, one bitline is connected to a multitude of memory cells. Because of the large number (e.g. 128, 256 or 512) of cell transistors connected to the bitline, as well as its physical length and its proximity to other nearby conductors, the bitline is highly capacitive. Typically, the bitline parasitic capacitance C_b is about eight to ten times the memory cell capacitance C_c [1]. Therefore, the cell signal voltage will be greatly attenuated and thus the bitline voltage changes only slightly about V_{BLP}. Because the cell signal is lost, the read operation in a DRAM is called a *destructive read*. The voltage signal on the bitline (with respect to V_{BLP}) can be calculated according to the following equation, which is implied by conservation of charge:

$$V_{signal} = (V_{cell} - \frac{V_{DD}}{2}) \bullet \frac{C_c}{C_b + C_c} \qquad (1.2)$$

For 0.18-μm technology, the typical power supply voltage is 1.8 V [5]. Assuming the bitline parasitic capacitance to cell capacitance ratio is roughly 9, the bitline signal is only 90 mV with respect to V_{BLP}. The differential voltage between the bitline holding the attenuated cell voltage and a second bitline holding a reference voltage of V_{BLP} is amplified to the power rails by a *sense amplifier* and written back to the memory cell. The wordline can then be de-asserted and the destructive read and restore operation finishes.

Figure 1-3 shows the high-level organization of a DRAM block. The center part is the memory array formed by a two-dimensional matrix of memory cells, as shown in Figure 1-1. Wordlines decoded from a *row decoder* run vertically and each wordline is connected to the gates of the access transistors in a *row* of memory cells. Bitlines run horizontally and each of them is connected to the sources of access transistors in a *column* of memory cells. *True bitlines* are connected to the "+" terminals of their sense-amplifiers, while *complement bitlines* are connected to the "-" terminals. Notice that a *"folded bitline"* architecture is used, as is typically the case in modem DRAM memory arrays. Thus the true bitline (bitline) and the complement bitline (bitline*, where "*" denotes the complement signal) are folded together at the sense amplifier to form one column (a bitline pair) that is selected by the *column decoder* for the same column address.

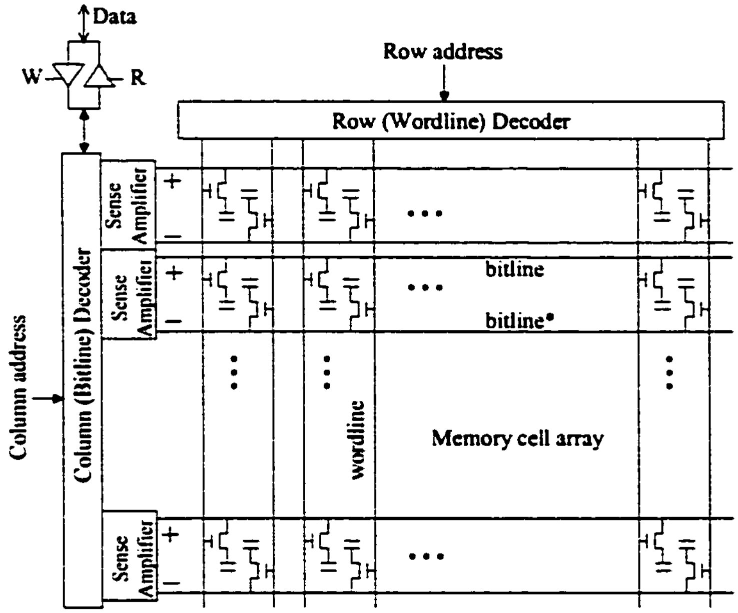


Figure 1-3. A DRAM block [6]

Sensing and restoring techniques in conventional two-level DRAM are described in [1] and [4] in considerable detail. When a wordline is asserted from V_{SS} to V_{DD}, all cell access transistors connected the wordline turn on. For sensing, the charge stored in a memory cell is shared with one bitline of a bitline pair while the other bitline continues to hold the precharge voltage to be used as a reference voltage. The sense amplifiers behave as differential amplifiers that compare the voltages on the two adjacent bitlines to determine the data from the activated memory cell. They amplify the two bitline signals to the point that they reach the array supply voltages, i.e., V_{DD} and ground. For writing new data into a cell, the state of sense amplifier is overwritten by a low-impedance write driver (shown at the top left corner in Figure 1-3). This ensures that the new data is written back into the memory cell.

1.3 MLDRAM Technology

By shrinking the feature size of the transistors and adopting complex three-dimensional cell capacitor structures, DRAM manufacturers have kept increasing the storage density of DRAM and lowering the per-bit cost. However, density innovations in DRAM are approaching economic limits in terms of the cost of manufacturing. It is more and more difficult to further increase the memory density cost-effectively. Multilevel DRAM, by storing more than one bit per memory cell, explores an additional dimension to further increase the storage density without requiring changes in the DRAM process.

In a conventional two-level DRAM cell, the cell voltage corresponding to one bit of data can be defined as one of two possible values as follows [6]:

$$V_{cell} \in \{ a_0 V_{DD},\ a_1 V_{DD} \},\ \text{where } a_0 = 0 \text{ and } a_1 = 1 \qquad (1.3)$$

For a multilevel DRAM that stores N analog voltage levels per cell, the cell voltage can be one of the following voltage levels between V_{DD} and V_{SS}:

$$V_{cell} \in \{ a_0, a_1, a_2, \ldots, a_{N-1} \} V_{DD} \qquad (1.4)$$

In order to maximize the noise margins between possible signals, these voltage levels are equally spaced. So equation (1.4) can be rewritten as (1.5):

$$V_{cell} \in \{0, 1, 2, \ldots, N-1\} \frac{V_{DD}}{N-1} \qquad (1.5)$$

To reliably sense these N possible analog voltages, N-1 reference voltages are needed to compare with the cell voltages. These reference voltages are positioned midway between each pair of adjacent cell voltage levels to maximize the worst-case noise margin seen by the sense amplifier. The reference voltage levels are given in equation (1.6):

$$V_{REF} \in \{1, 3, 5, \ldots, 2N-3\} \frac{V_{DD}}{2(N-1)} \qquad (1.6)$$

Note that n bits can have 2^n possible joint states. Therefore one cell with N possible analog signal levels can be used to encode [$\log_2$N] whole bits. Fractional bits can be stored in a cell if groups of cells are considered together, with fractional bits combined into whole bits during data storage and retrieval. In the limit, when converting N analog voltage levels into binary data, n digital bits can be encoded according to equation (1.7):

$$n = \log_2 N \qquad (1.7)$$

Figure 1-4 [6] illustrates how two bits of data can be encoded into one memory cell using four equally-spaced voltage levels from V_{SS} to V_{DD}.

The storage density of an N-level MLDRAM can be potentially increased over that of DRAM by a factor of n-1 without any changes in the DRAM process technology or the optimized DRAM cell array. The most critical challenge for MLDRAM technology to overcome is probably the reduced noise margins, especially for the already more closely spaced power supply levels in today's deep sub-micron technology. The noise margins for an *N*-level MLDRAM cell are reduced by a factor of *N*-1 over those of a conventional DRAM cell in the same technology. As in the earlier example, in 0.18-μm technology, with a 1.8 V power supply voltage and a bitline-to-cell capacitance ratio of 9, the signal voltage for a conventional 2-levels-per-cell DRAM is 90 mV. However, the signal voltage is only 30 mV for a 4-level MLDRAM, and 18 mV for a 6-level MLDRAM. The reduced noise margins make the MLDRAM much more vulnerable to on-chip noise, cell leakage, sense amplifier offsets, and environmental ionizing radiation. Note also that the maximum retention time of the memory cell will decrease as the number of voltage levels increases and the inter signal level voltage decreases. Consequently, the refresh rate must be increased to maintain the data, and this would increase stand-by power consumption.

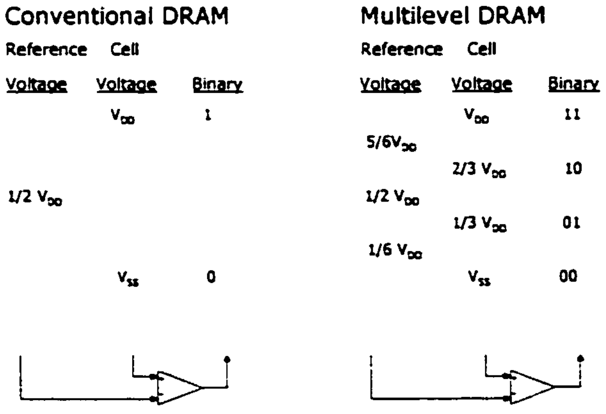


Figure 1-4. A two-bit-per-cell storage scheme [6]

The sensing and restoring operation for MLDRAM is also more complex than for DRAM. In MLDRAM, multiple analog voltage levels other than V_{SS} and V_{DD} are required. The voltages levels should be obtained on chip from the desired digital write data. For sensing, more reference voltage levels need to be created to compare with the cell voltages to allow the analog cell voltage to be converted back to digital form. A more complex control sequence will inevitably increase the access time as well as the silicon area of an MLDRAM. Therefore, the theoretical gain in storage density is reduced by the additional area overhead.

Although multilevel floating-gate nonvolatile flash memory has already appeared in the market [7], there are still no commercial multilevel DRAM products due to all of these disadvantages. Whether or not MLDRAM can be commercially cost-effective is still unclear. Our research aims to help clarify the technical trade-offs.

A suitable circuit design for MLDRAM should minimize all of these disadvantages and also satisfy a number of other criteria. First, the feasibility of commercial production depends on the robustness of the circuits to a large degree. The correct operation of the MLDRAM circuits must be insensitive to the inevitable IC process variations that can cause the circuit parameters to differ from the nominal values. Second, the MLDRAM must be insensitive to the effects of on-chip noise. Noise insensitivity becomes more important for MLDRAM than DRAM due to the reduced noise margins. Great care must be taken when designing balanced circuits in MLDRAM to ensure that noise sources, such as gate to source/drain charge injection through the cell access transistors, will be cancelled as common mode noise by the differential sense amplifiers. Finally, the access time for MLDRAM should not be too slow compared with DRAM. Otherwise its performance will not meet the same operational requirements already met by DRAM and expected by customers. Much slower access time would restrict MLDRAM to file storage applications, where the random-access time is not as important.

1.4 Thesis Outline

Several multilevel DRAM schemes have been described in previous papers and some of them have been implemented in experimental chips [8] [9] [10] [11] [12]. Recently, Birk *et al.* [6] [13] proposed a robust MLDRAM scheme that combines the speed advantages of [9] with the noise cancellation advantages of [10]. This thesis describes the design, implementation and testing of an MLDRAM test chip based on Birk's design, which has adjustable cell capacity and is more flexible than Chan's fixed 2-bits-per-cell implementation [14]. Regardless of the selected cell capacity, data storage and sensing follow the same mechanisms described in [6] and [13]. The test chip, which has an adjustable cell capacity of 1, 1.5, 2, 2.5 bit(s) per cell, should be useful for characterizing the chip implementation and for experimentally determining the practical limits of one particularly promising MLDRAM technology.

The rest of this thesis is organized as follows:

The next chapter reviews basic concepts, useful circuit design techniques, chip architecture, as well as the read and write operations of conventional DRAM. Then, a detailed overview of several MLDRAM techniques is given. The previous MLDRAM projects at the University of Alberta are reviewed. Benefits, drawbacks and technology challenges for MLDRAM are summarized at the end.

Chapters 3 and 4 explore the design, simulation and implementation of the test chip. The test chip was designed in TSMC's 0.18-micron CMOS technology that is supported by the Canadian Microelectronics Corporation (CMC). Simulations for both a PMOS and an NMOS-based memory array demonstrated correct functionality for the four different cell capacities. A fully-customized memory core was combined with periphery logic, which was designed using cells from CMC's standard cell libraries.

Silicon verifications for five prototype chips are presented in Chapter 5. The test environment, some basic functional tests and test results are described.

Chapter 6 makes conclusions and proposes future related work.

Chapter 2 Previous Work

MLDRAM technology borrows many ideas from conventional DRAM technology. Sections 2.1, 2.2 and 2.3 review basic concepts, circuit design techniques, chip architectures and operations for conventional DRAM. Several previous MLDRAM schemes are then reviewed and compared in Section 2.4. The major advantages, disadvantages and technology challenges of MLDRAM are summarized in Section 2.5.

2.1 Basic DRAM Circuits

DRAMs evolved from the earliest 1 Kb generation to the imminent 1 Gb generation through advances in both integrated circuit manufacturing processes and circuit design techniques. A DRAM manufacturing process technology distinguishes itself from a logic process in that it places high density and low per-bit cost, rather than high performance, as its first concern. With continually shrinking transistor feature sizes and innovative cell capacitor structures, which evolved from the early planar capacitors to complex trench and stacked cell capacitors, DRAM has been the technology driver for semiconductor industry for over two decades.

Unlike most logic ICs, DRAMs have a very regular physical structure. Figure 1-3 in Chapter 1 shows a very simplified view of DRAM chip architecture. The memory arrays are composed of rows and columns of memory cells, which take up most of the memory chip area. The densely packed cell array and the immediately surrounding circuits are called the *memory core*. The core normally contains the sense amplifiers, bitline precharge devices, wordline drivers, and some address decoding logic. To achieve an efficiently packed layout, the core circuitry is constrained to be pitch matched to the bitlines and wordlines. The outermost circuitry of the DRAM chip, which contains address decoding logic, the power regulation circuits, the control logic and the interfacing circuitry, is referred to as the *periphery*.

2.1.1 The DRAM Cell Array

A DRAM array is a two-dimensional array of DRAM memory cells. As mentioned before, the modern 1T-1C DRAM cell includes a cell access transistor and a cell storage capacitor. Wordlines run in one direction and are connected to the gates of the cell access transistors. Bitlines run perpendicularly to the wordlines and are connected to the source/drain terminals of the cell access transistors. Figure 2-1 shows a basic DRAM cell array structure with the so-called folded bitline architecture (explained in section 2.1.1.2).

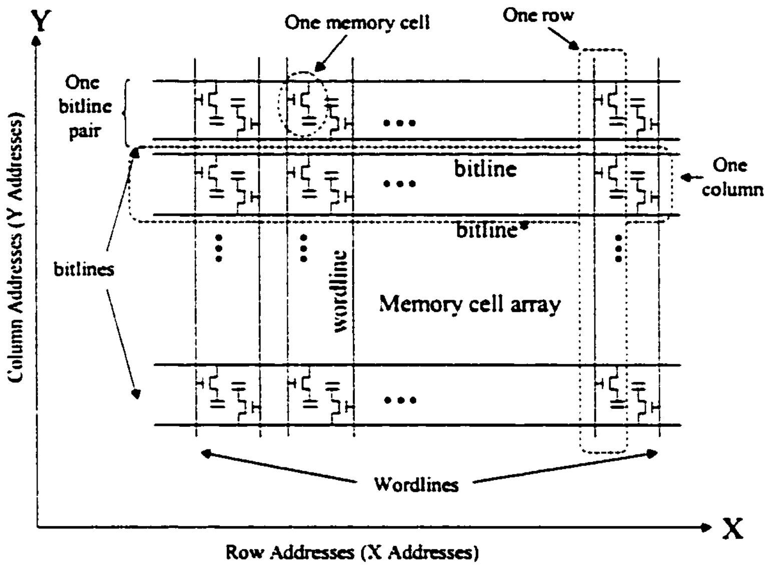


Figure 2-1. A DRAM cell array [6]

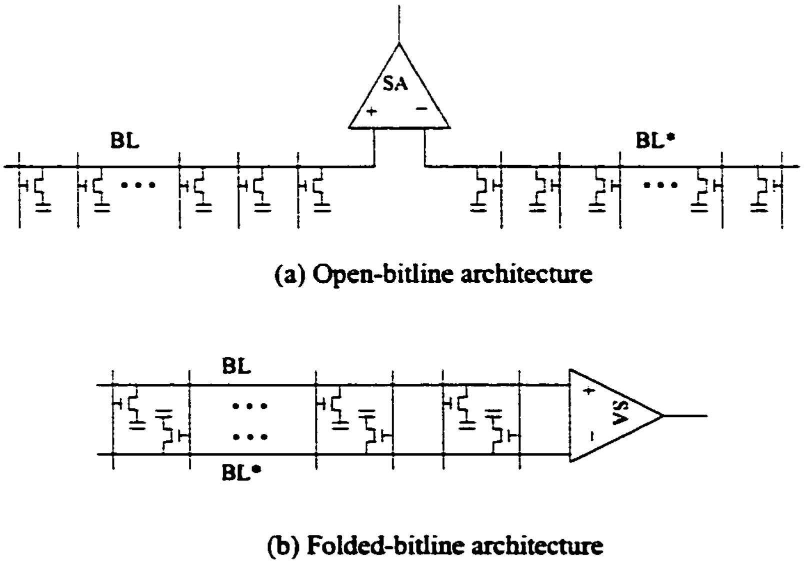
2.1.1.1 Rows and Columns

As indicated in Figure 2-1, one *row* of DRAM refers to one wordline and all the memory cells connected to this wordline. One *column* of DRAM refers to one bitline and all the memory cells connected to this bitline. However, it is common practice to use the terms wordline and row, and bitline pair and column interchangeably. Given a particular wordline and bitline address, one or a few memory cell(s) can be accessed.

Note that when one wordline is addressed, the whole row of memory cells that is connected to this wordline is accessed. Standard DRAM operations, such as page mode access, take advantage of this situation by allowing fast access to cells in the currently accessed row. Depending on the memory access pattern, such a mode can improve the chip average cell access time dramatically.

2.1.1.2 Folded Bitlines and Twisted Bitlines

DRAM arrays prior to the 64 Kb generation (pre 1980) employed the *open bitline* array architecture as illustrated in Figure 2-2 (a). Here two bitlines are connected to the sense amplifier from two separate arrays. In this arrangement, each wordline connects to the memory cell at every bitline-wordline intersection, creating crosspoint-style arrays. Since a memory cell is present at each wordline-bitline intersection, the smallest possible area of the memory cell is between $4F^2$ and $6F^2$, where the *feature size* F is the minimum realizable process dimension (that is, the half bitline pitch and the half wordline pitch). The folded bitline array architecture, shown in Figure 2-2 (b), comes from taking the complement bitline BL* in Figure 2-2 (a) that goes into the same sense amplifier as the corresponding true bitline BL and folding it alongside BL. Since each wordline connects a crosspoint with a memory cell on only every other bitline, the minimum layout area for a memory cell in this arrangement will be $8F^2$ [4].

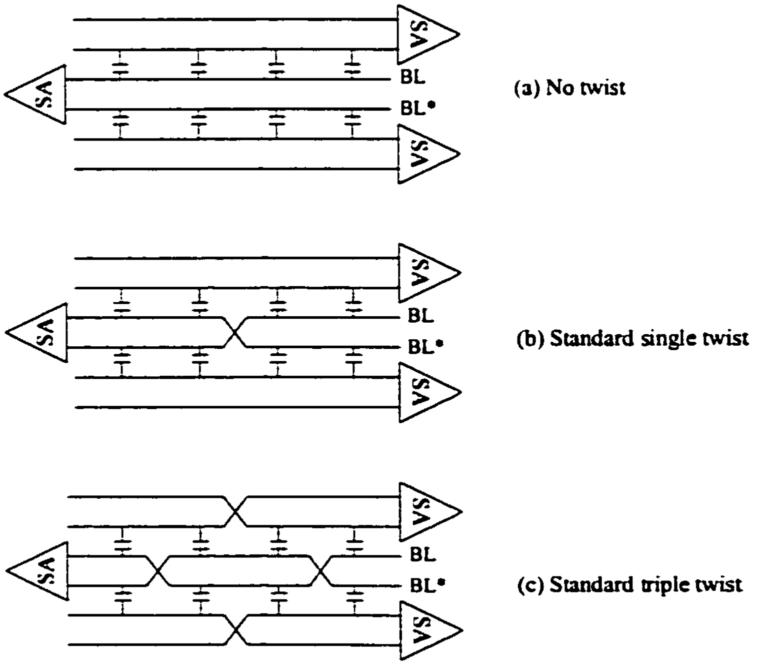


(a) Open-bitline architecture

(b) Folded-bitline architecture

Figure 2-2. Open-bitline and folded-bitline array architectures [6]

Although the open-bitline array architecture permits the greatest possible physical cell density, the folded-bitline array architecture became dominant ever since the 64 Kb generation due its greatly improved noise immunity properties. As we discussed before, bitlines are highly capacitive and normally the bitline parasitic capitance is about 8 to 10 times the memory cell capacitance. The resulting attenuated memory cell signal voltages that must be detected by the sense amplifiers are only a few hundred millivolts. Reduced power supply voltages and increasing cell capacitor leakage currents aggravate the situation by reducing the worst-case signal voltage even more. Therefore, the sensing operation is particularly vulnerable to errors caused by noise and component mismatches. In the folded bitline arrangement, by putting BL and BL* in parallel and close to each other, electrical parameter mismatches in the bitlines due to process variations are minimized. Noise injected onto the bitlines will be tend to be common mode noise, and therefore, effectively rejected by the differential mode sense amplifier. Moreover, the folded bitline arrangement can use bitline twisting (not available in open bitlines) to further increase the noise immunity by reducing effect of the bitline-bitline capacitive coupling [4].



(a) No twist

(b) Standard single twist

(c) Standard triple twist

Figure 2-3. Bitline twisting schemes [4]

Bitline twisting in one or more positions of the bitline physical layout is widely used in DRAM arrays. Twisting balances the coupling capacitance to the adjacent bitline pairs so that coupling between the two columns will not be seen by the sense amplifier. Figure 2-3 (b, c) shows two standard bitline twisting schemes and illustrates how parasitic capacitive coupling will be balanced.

2.1.1.3 Friendly Cells and Dummy Cells

Balanced layout is very important for small-signal analog and mixed-signal IC designs. Memory array design is actually mostly small-signal analog design despite the digital signals at the outside interfaces. The information data stored in the memory cells are held in the form of analog voltages, and the attenuated signals seen during sensing are very vulnerable to any mismatches and noises. Although process variation is inevitable in IC production, designers can still optimize their design to ensure a balanced layout and minimum variations in critical electrical parameters.

In a memory array, all memory cells are designed as much as possible to have the same size and the same physical and electrical immediate environment. So the parasitic capacitance and other electrical parameters of each memory cell should be made the same in the ideal layout design. However, for memory cells at the edges of the array, there are memory cells absent on one or two sides and this topographical discontinuity will affect the quality of the corresponding signals. A common practice to solve this problem is to add "*friendly cells*" all around the edge cells of the memory array [6]. Such friendly cells are usually identical to the normal ones, but they are never used for data storage. The wordline connected to the gates of their access transistors is tied permanently to the power supply voltage (V_{SS} for NMOS and V_{DD} for PMOS). Alternatively, the friendly cells may contain capacitors only to mimic normal cell capacitors. This approach appears to be common in embedded DRAM design.

Dummy cells, which are also called "*reference cells*", provide the reference voltages to be used to compare with the signal voltages from normal memory cells. Unlike friendly cells, dummy cells have wordlines that can be activated. Depending on the design, they may be identical to normal cells. It was popular to use dummy cells in early DRAMs when the bitline precharge voltage V_{BLP} was not equal to 1/2 V_{DD}. This feature is no longer used for modern DRAMs, which have a 1/2 V_{DD}

precharge voltage that can be directly used as the reference voltage. Nevertheless, dummy cells can still be helpful for charge injection cancellation. When a wordline is activated, there will be a small amount of electrical charge injected onto the bitlines of the all accessed memory cells through the parasitic gate-to-drain capacitance of the access transistors. This small amount of charge injected onto the bitlines is very critical for the small-signal sensing, especially for multilevel DRAM designs where noise margins are even smaller. Dummy cells and dummy wordlines may be used in multilevel DRAM to store reference voltages and to cancel out charge injection.

2.1.2 The DRAM Core

As stated earlier, the DRAM core contains the entire DRAM cell array and all the pitch-matched circuits around the array. Figure 2-4 shows a typical floorplan for the DRAM core circuitry. Row decoders and wordline drivers are laid out within wordline pitch. Column decoders and sense amplifiers are laid out within the bitline pitch. As shown in the figure, the column (but not the bitline) pitch is doubled for the folded bitline array architecture compared to that for the open bitline array architecture.

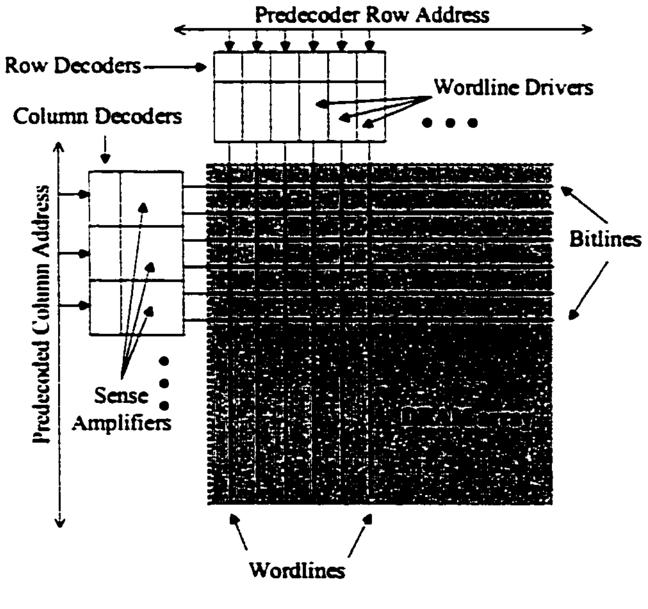


Figure 2-4. DRAM core circuitry [6]

Address decoding is normally split into hierarchical stages for layout and performance reasons. In the first stage, pre-decoders in the periphery take a group of two or three (n) address bits and produce a combination of four or eight (2^n) minterms. Those minterms are fed to the next level decoding circuits (decoders) to produce more minterms decoded from yet more address bits. Final-stage decoding circuits for the complete address of a wordline or a bitline, which logically combine minterms from the previous decoder stages, can then be more easily laid out in the very narrow pitch of a wordline or a bitline pair. Note how Figure 2-4 shows the pitch-matched final stage row and column decoders. Whatever logic style (static or dynamic) decoders or decoder trees are used, the primary objectives in decoder design are to maximize speed, minimize die area, and simplify the constraints for any pitch-matched circuits.

In commercial DRAMs, there are also redundant, fully functional columns and rows of spare memory cells at the edge of each DRAM array (not shown in Figure 2-4), which are used to improve the effective yield of sellable memories. If a defect occurs affecting a particular wordline or a bitline during manufacturing, memory cells associated with that line may not function correctly. Without the redundant rows and columns, the whole chip would have to be discarded due to less than nominal available functional capacity. With these extra rows and columns, the memory cells that do not work can be located and removed from service, and the addresses for redundant rows and/or columns can be permanently modified to replace the defective ones. The required wire reworking is usually accomplished using laser-cut fuses. Therefore many of the memory chips with physical defects and faulty cells can still be saved and sold. This technique is called *static redundancy*. Static redundancy is required to achieve economically high yields in all commercial memory chips of at least 1 Mb capacity. Extra chip area, design complexity and manufacturing time are the main additional costs of this technique.

2.1.2.1 The Sense Amplifier

The term *sense amplifier block* refers to a collection of circuit elements (excluding memory cells) that are physically laid out within the bitline pitch. This block normally includes isolation switches, bitline precharge devices, one or more NMOS sensing transistors, one or more PMOS sensing transistors, and the input/output transistors that connect the selected bitlines to the data buses. Figure 2-5 shows a typical DRAM sense amplifier block circuit.

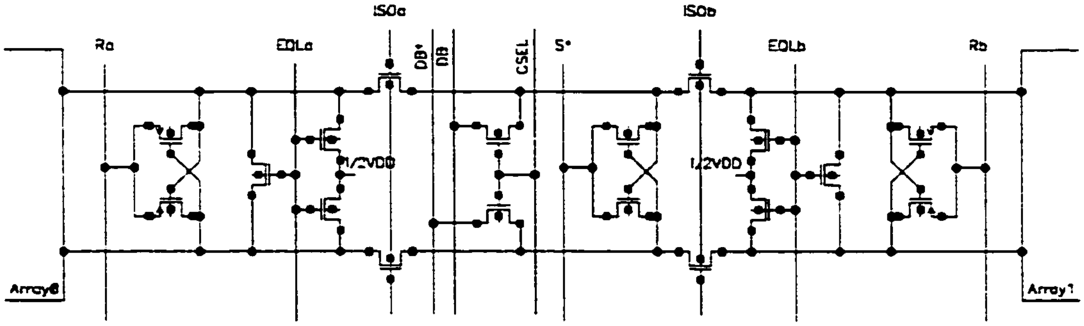


Figure 2-5. Standard sense amplifier block [4]

The *precharge circuit*, controlled by EQLa and EQLb in Figure 2-5, is used to precharge and equalize the bitlines to 1/2 V_{DD} prior to the sensing operation. NMOS transistors with relatively large channels are used to ensure greater drivability and faster precharging. The 1/2 V_{DD} precharge voltage is used in modern DRAMs because it reduces power consumption and read-write time. By applying the 1/2 V_{DD} voltage to the cell plate node, rather than V_{DD} say, the electrical stress is reduced on the very thin dielectric of cell capacitors.

Isolation switches, controlled by ISOa and ISOb in Figure 2-5, allow two adjacent arrays to share the same sense amplifier block. Another benefit of isolation devices is that they provide some resistance between the low capacitance sense node of the sense amplifier and the highly capacitive bitlines. This resistance partially isolates the sensing operation from the bitlines and permits much faster sense amplifier latching.

Input/output transistors are also called column access transistors. Controlled by the decoded signals from the column decoders, which is called CSEL in Figure 2-5, they allow data to be read from and written to specific bitline pairs. These transistors must be carefully sized so that instability is not introduced in the sense amplifiers from the data bus precharge voltage or possible remnant voltages. The ratio of the size of the NMOS sense amplifier transistors over the size of the I/O transistors is referred as the beta ratio. A beta ratio of between five to eight is typical [4]. Figure 2-6 shows how two bitline pairs are selected to connect to two different data buses through four I/O transistors.

NMOS and PMOS sense amplifiers are the fundamental elements of the whole sense amplifier block. An NMOS sense amplifier consists of cross-coupled NMOS transistors which drive the lower potential bitline to ground. Conversely, a PMOS sense amplifier consists of cross-coupled PMOS transistors which drive the higher potential bitline to the positive power supply voltage.

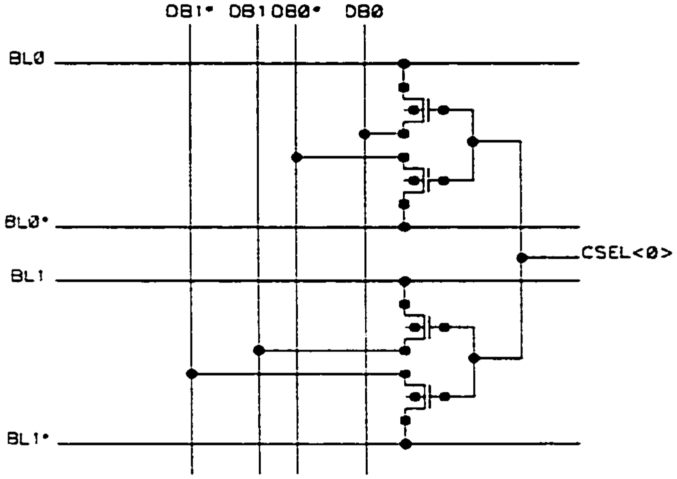


Figure 2-6. I/O transistor connections

Figure 2-7 shows the basic *CMOS sense amplifier* where NMOS and PMOS sense amplifiers are put together. The R and S* nodes of the sense amplifier are connected to the precharge voltage when the bitline pairs are precharged. During the sensing operation, the S and R nodes are connected to ground and VDD, respectively. When the sense amplifier is thus activated, it behaves like a feed-forward latch composed of two inverters. When powered up, it quickly drives the bitline signals to the complementary power rails in one of two possible states.

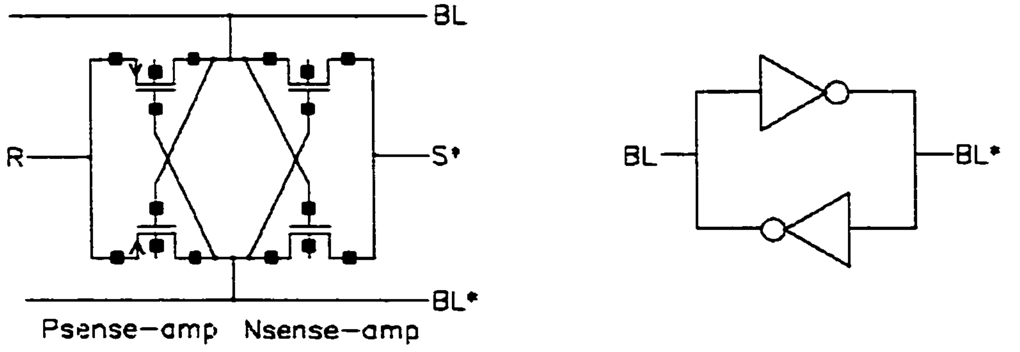


Figure 2-7. The basic sense amplifier

Because the sense amplifiers are used to detect and amplify very small differential cell signal voltages (less than 100 mV) during the sensing operation, they must be designed very carefully to ensure reliable operation. Balanced, symmetrical and duplicated circuit layout techniques are critical to a successful design [15].

Note that in the configuration of the sense amplifier block shown in Figure 2-5, the PMOS sense amplifiers are placed outside the isolation switches because the isolation switches are NMOS enhancement type and their gate control signals are not boosted. If this were not to be done, one of the bitline pairs can never be driven to a full V_{DD} because NMOS transistors can only pass an attenuated high voltage signal.

2.1.2.2 Wordline Drivers

Wordlines in DRAM are highly capacitive due to the connection of a relatively large number (e.g. 128 or 256) of cell access transistors; wordlines are also highly resistive due to the relatively high resistivity of the polysilicon layer. In order to decrease the RC delay of the wordline and to improve the read-write access time, large, low-impedance wordline drivers and parallel metal straps are employed. In addition, for NMOS (PMOS) access transistors, wordline drivers are required to drive the wordline signal to at least one threshold voltage above V_{DD} (below V_{SS}) so that a full V_{DD} (V_{SS}) can be passed from the bitline to the cell storage node. This technique is called *wordline boosting*. There are many different kinds of wordline driver circuit designs. In our test chip, we used the proven wordline driver circuits from ATMOS Corporation, as shown in Figure 2-8. Wordline boosting is realized for these CMOS drivers by simply changing the power supply voltage to the boosted power supply voltage.

2.1.2.2 Data Buses

As illustrated in Figure 2-6, information data are read from and written to the memory core via the data buses that run over or alongside the sense amplifiers. The number of the parallel data buses determines the available data bandwidth to and from the memory core. Since data buses are long and highly capacitive, the sense amplifiers are too small to drive them to the V_{DD} and V_{SS} in a reasonable time. Therefore, just like the bitlines, data buses are precharged to 1/2 V_{DD} and then left floating at that mid voltage. For a sensing operation, when the column select signal activates the I/O transistors that connect the data buses to specific bitlines, the large data bus differential sense amplifiers will quickly detect the weak signals coming from the bitline sense amplifier and help drive the data buses to the power rails. Noise-rejection techniques, such as folded and twisted data buses, could be employed. For a write operation, the data bus voltages are driven to the intended levels by a powerful write driver and the resulting signal will overwrite the state of the comparatively small bitline sense amplifiers.

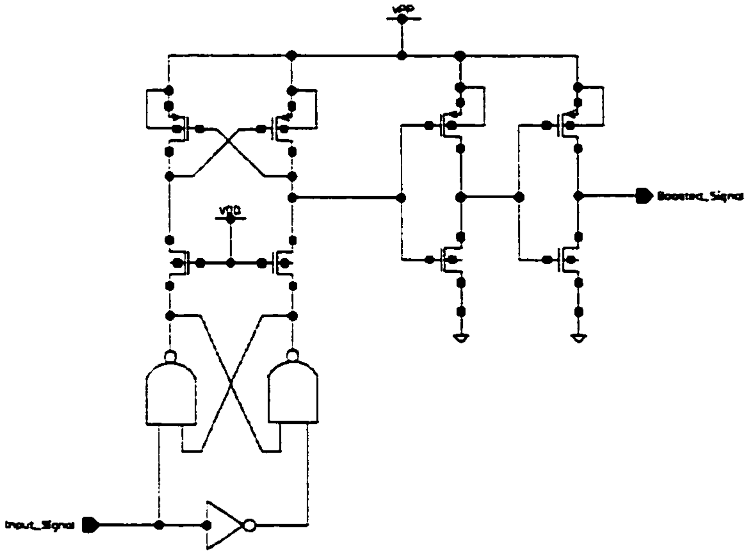


Figure 2-8. Wordline driver from ATMOS Corporation

2.1.3 The DRAM Periphery

The DRAM periphery contains all of the rest of the circuitry around the memory core in the memory chip. It consists of the pre-decoder logic, address buffers, data bus sense amplifiers and write drivers, circuits that control row and column redundancy, voltage regulators used to create different on-chip voltages, as well as some control logic and data output/input interface logic. For synchronous DRAM (SDRAM), the on-chip control circuitry for fast operation modes is quite complicated. Pipelining techniques and dynamic logic gates (or bipolar transistors) must typically be used in the periphery to improve speed.

2.1.4 DRAM Floorplanning

The maximum length of wordlines and bitlines is limited in practical DRAM chips. Long wordlines have large parasitic capacitance and resistance, so the RC delay of the wordlines increases and the access time becomes too long. Similarly, long bitlines have large parasitic capacitance and appreciable resistance, which means high dynamic current and power consumption, and weaker cell signals for sensing. Therefore, large capacity memory chips are always segmented into multiple smaller banks with shorter bitlines and wordlines. In each bank, there is an array of memory array blocks. For example, a 256 Mb synchronous DRAM might have 8 memory banks composed of 32 Mb memory bits each. Each bank may in turn have 8K rows and 4K columns arranged into a 16 × 16 array of basic memory array blocks. Each basic memory array block may have 512 nominal rows plus 4 redundant rows and 256 nominal columns plus 8 redundant columns (512 + 16 bitlines), for a total of 128 Kb. The bank as a whole has one row decoder, major wordline drivers along the edges of the bank, and one column decoder. The major wordline drivers drive sub-wordline drivers in each basic array to activate a specific wordline. Figure 2-9 shows the floorplan of such a 256 Mb DDR-SDRAM.

Cell efficiency and die efficiency are two commonly used parameters to judge the design of a memory chip. *Cell efficiency* is the percentage of the die area occupied by memory cells (not including the redundant cells) over the memory core. Normally it equals the cell area divided by the core area. This parameter highly depends on the process technology and the cell layout design. The *die efficiency* is defined as the die area required by all of the memory cells on average of the whole chip, which can be calculated by dividing the total die area by the total silicon area occupied by the useful memory cells.

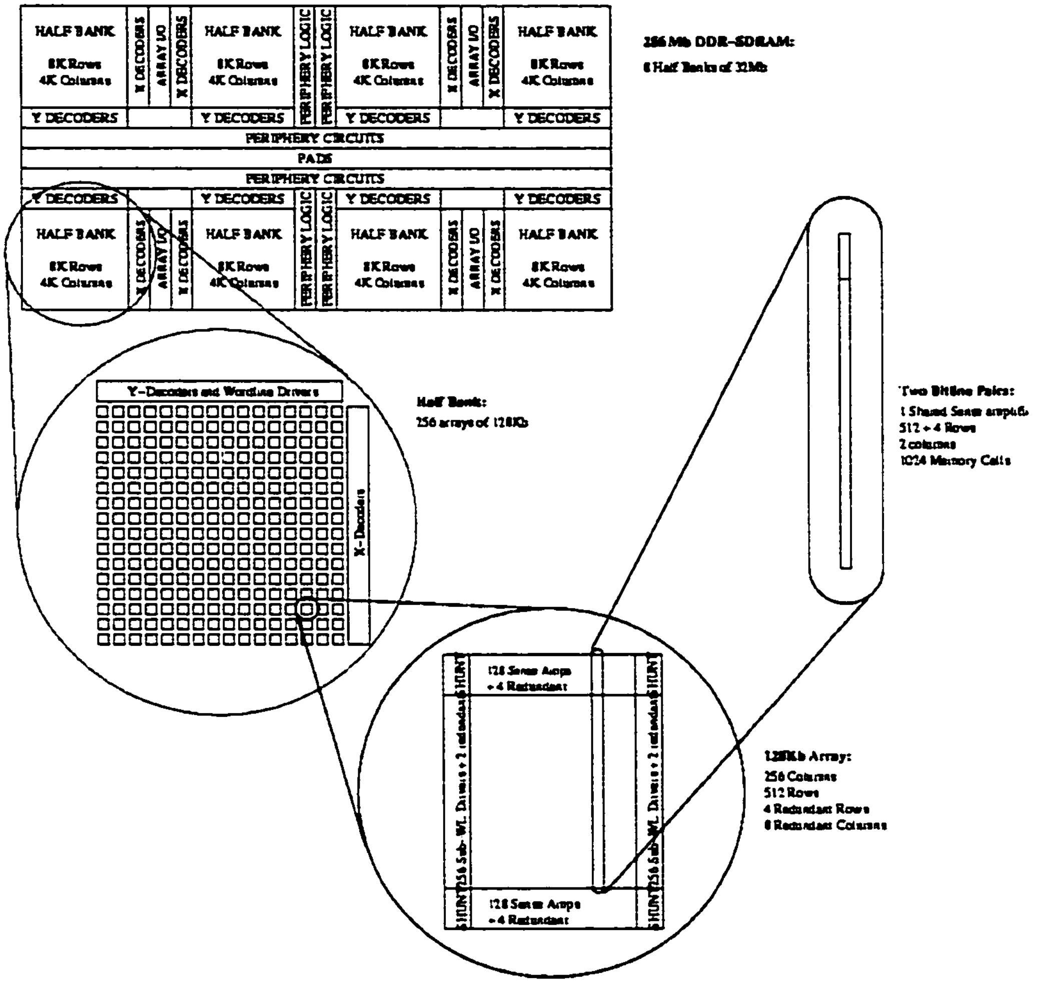


Figure 2-9. Floorplan of a 256-Mb DDR-SDRAM [6]

2.2 DRAM Operations

The operation modes of conventional DRAM have been continually evolved to improve chip performance. In this section, first we will discuss the basic operation of the DRAM core as background knowledge for understanding multilevel DRAM operation. Then different speed improvement operation modes will be introduced.

If we consider the bitline pair of the standard sense amplifier block and its associated arrays in Figure 2-5 and assume that wordline WLi in array0 is addressed, we will now obtain the column schematic shown in Figure 2-10:

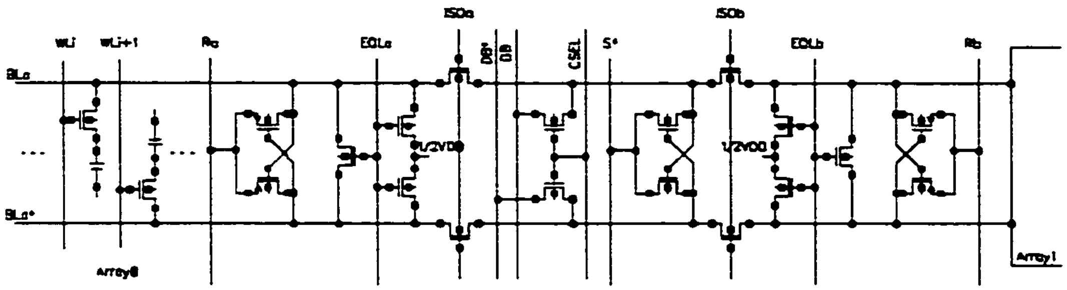


Figure 2-10. Schematic of a DRAM bitline pair

As shown in Figure 2-11, a basic read/write operation cycle starts with the idle state, where all bitlines and sense amplifier nodes are precharged to 1/2 V_{DD}, the isolation switch signals ISOa and ISOb are switched on, and no wordlines in array0 and array1 are activated. When a wordline in array0 is to be accessed, the isolation switch controlled by signal ISOb will disconnect array1 from the shared sense amplifier block and leave only array0 with the complete sense amplifier block. Then the EQLa signal stops precharging the bitline pair BLa and BLa*, leaving this bitline pair floating at the bitline precharge voltage V_{BLP}. One specific wordline in array0, WLi, is activated by the wordline decoder and wordline driver for the given row address. All the memory cells connected to WLi are now accessed. The charge stored in each memory cell will be distributed along one bitline of the bitline pair. In our example, the voltage of the true bitline BLa will be slightly shifted by about a few hundred millivolts or less. Next, asserting signal R and S* to the power supplies activates the PMOS and NMOS sense amplifiers. The weak differential signals on BLa and BLa* are then amplified and driven to the power rails. The data stored in the memory cell is thus amplified and ready to be driven out over the data bus. A column select signal CSEL, decoded from the given column address, will be pulsed to connect the sense amplifier with the data bus. Depending on the operation type, the data is either read out to an external data pin or overwritten by the write driver. Then either the original data (for a read operation) or the newly written data (for a write operation) will be restored by de-asserting WLi. Finally, the sense amplifiers and bitlines are precharged back to 1/2 V_{DD} and array1 is connected to the sense amplifier block once again by asserting the ISOb signal. The circuits are now idle again and ready for the next read/write operation.

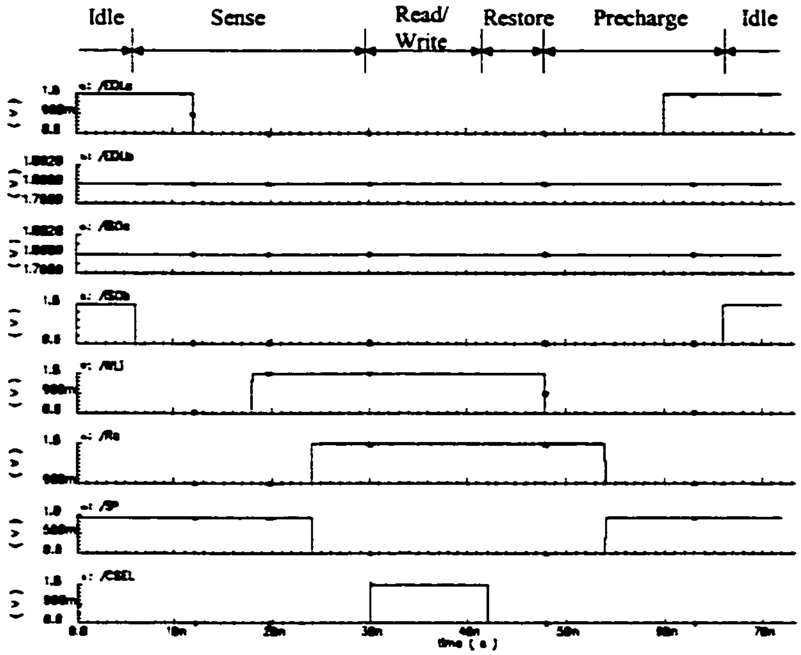


Figure 2-11. Basic read/write timing diagram for the DRAM core

The performance of a DRAM chip can be improved by speeding up the absolute access time and/or delivering data at a faster average rate. Recall that during the normal read operation, after a specific wordline is addressed in a memory bank, a large number of data bits from all of the memory cells connected to this same wordline (collectively called a *page* of bits) are held on the sense amplifiers. The sense amplifiers behave like latches waiting to be selected by the column selecting signals. Many of the fast access DRAM modes that have been developed rely on the ability to rapidly access the data stored in a page without having to initiate a completely new memory cycle. *Nibble mode*, in which four sequential bits of data can be accessed quickly, was used in early DRAMs. *Static mode*, *page mode*, *fast page mode* are used to get faster access to random data in one page by eliminating the row address setup and hold time. *Fast Page with Extended data out* (EDO) mode (also called *Hyperpage* mode) and *Burst Mode* with EDO are operation modes that are widely available in recent DRAMs. Meanwhile, multiple parallel I/O buses and more data I/O pins are used to increase the available data bandwidth and therefore to directly increase the data transfer rate [16].

2.3 Embedded DRAMs

Embedded memory refers to a memory block that is integrated with other logic circuits in one chip. Embedded SRAMs have been widely used to implement cache memory in microprocessor-based system designs (SoC, system-on-a-chip) due to its fast access time and its compatibility with digital logic processes. Embedded DRAMs are also becoming more and more common in all kinds of applications. Elimination of I/O pins, lower power consumption, higher speed, wider internal data bandwidth, and more flexible data granularity make embedded DRAMs suitable for high density on chip memory in high performance systems such as graphics chips and hard disc controllers. Design and cost trade offs have to be made carefully between the DRAM process and the logic process because they have different priorities [17], [18], [19].

ATMOS' SoC-RAM^{TM} is an embedded memory architecture developed by ATMOS Corporation for system-on-a-chip (SoC) applications [19]. SoC-RAM is delivered in the form of silicon-proven hard memory cores. It is intended to be used in the wireless, graphics, imaging and networking industries that require very large on-chip memory. It is available in both a merged DRAM-logic process and a pure logic process. The layout design of the multilevel DRAM test chip used elements of the planar logic (PL) SoC-RAM core. A brief overview of ATMOS' SoC-RAM is presented below and a detailed discussion about the MLDRAM test chip layout design will be given in Chapter 3.

The structure of ATMOS' 0.18-μm embedded SoC-RAM can be determined using their memory layout compiler. For example, the 0.18-μm, 8-Mb embedded synchronous DRAM microcell has an area of 27.3 mm^2 (with a height of 6000 μm and a width of 4545 μm) and a cell efficiency of 63.4%. It has 16 memory banks. Each bank has 128 pages. There are 16 words with a word width of 256 bits per page. The basic component in the memory bank is a memory block. There are 2 blocks in each bank. Each memory block has 64 rows and 4096 columns [19].

The memory cells of the SoC-RAM are NMOS-based. The cell capacitor is implemented using the parasitic channel capacitance of an NMOS transistor. The schematic and layout of a memory cell are shown in Figure 2-12. The common cell plate for the memory cells is implemented in the poly1 layer. Deep n-well (from which a p-well is formed) is used to isolate the memory array from the rest of the chip.

The cell plate is biased at more than one NMOS threshold voltage (V_{tn}) above V_{DD} in order to get linear and maximal cell capacitance. A detailed discussion about the memory cell capacitance will be given in Chapter 4.

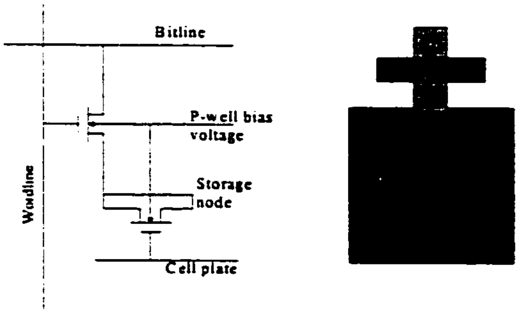


Figure 2-12. The schematic and layout of ATMOS' SoC-RAM cell

2.4 MLDRAM Circuits and Techniques

Storing more than one bit per memory location to achieve higher memory density is not a new concept. D.A. Rich authored a survey of several multi-valued ROMs and RAMs that were known in 1986 [12]. Multiple states can be encoded into a ROM memory cell by varying the threshold voltage of the memory transistor or by varying the transistor's width-to-length ratio in order to change the gain current. M. Aoki et al. proposed the first 4-bit, 16-levels-per-cell DRAM in 1985 [8]. The proposed MLDRAM read and wrote the 16 states from and into the memory cell by applying a staircase waveform voltage to the addressed wordline. Since the sequential access for this scheme is too slow, it has not been pursued further although it is scalable in theory to any number of voltage levels. In 1989, T. Furuyama *et al.* from Toshiba Corporation designed and fabricated an experimental 1-Mb 2-bits-per-cell DRAM chip using a new technique with global reference voltages and fast parallel access [9]. In 1996, P. Gillingham from MOSAID Technologies proposed another 2-bits-per-cell multilevel DRAM sensing and restoring scheme that is robust to small parameter variations and that uses locally-generated reference voltages [10]. In 1997, Okuda *et al.* from NEC developed a 4-Gb 2-bits-per-cell DRAM that used a ratioed coupling capacitor to generate the reference voltages [11]. G. Birk *et al.* from the University of Alberta did a comparative simulation study of the three later MLDRAM techniques and proposed in [6] [13] a new scheme that combines the speed advantages of [9] and the noise cancellation advantages of [10]. A. Chan designed and implemented a test chip (ML3) based on the 2-bits-per-cell mode of Birk's new scheme in [14]. The test chip in this thesis project is also based on Birk's MLDRAM scheme, with more flexible operating modes to facilitate characterization experiments.

Although Intel has announced plans to introduce a 4-bit/cell flash device by 2003 or 2004 [7], there is still no multilevel DRAM product that has been considered promising enough to enter production.

In the following sections, we will review Furuyama's and Gillingham's MLDRAM schemes that preceded to Birk's work. Then, Birk's MLDRAM scheme will be reviewed in a full detail.

2.4.1 Furuyama's MLDRAM

Figure 2-13 shows the basic block of Furuyama's MLDRAM scheme [9]. It consists of a pair of sub-bitlines, normal and dummy memory cells, a bitline precharge circuit, sense amplifier, and the I/O devices. Each sub-bitline has one dummy memory cell, which is used to hold the reference voltage. The reference voltage, V_{DC}, is from a global voltage source created on chip. It is connected to the dummy cells by activating the control signal DCP. Like normal memory cells, dummy cells are connected or disconnected to the sub-bitlines by controlling the dummy wordline signals DWL and DWL*. The sense amplifer is connected to the sub-bitline pair through two pass transistor switches controlled by signal CNCT.

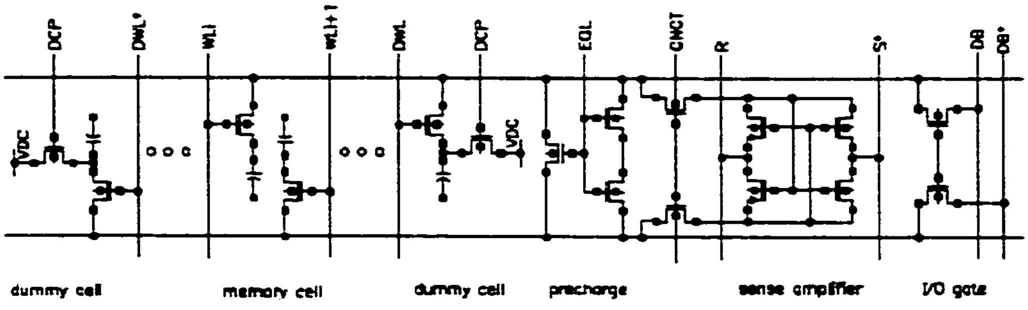


Figure 2-13. Furuyama's basic sub-bitline block design

Recall from Section 1.3 that for a 2-bits-per-cell MLDRAM, one of four analog voltages stored in a memory cell must be compared with three reference voltages during a sensing operation. In Furuyama's scheme, there are three identical basic sub-bitline blocks for each column, which can be connected by horizontal switches, as illustrated in the simplified block diagram in Figure 2-14. The dummy cells in each block hold one of the three reference voltages, which are V_{DCA} for block A, V_{DCB} for block B, and V_{DCC} for block C.

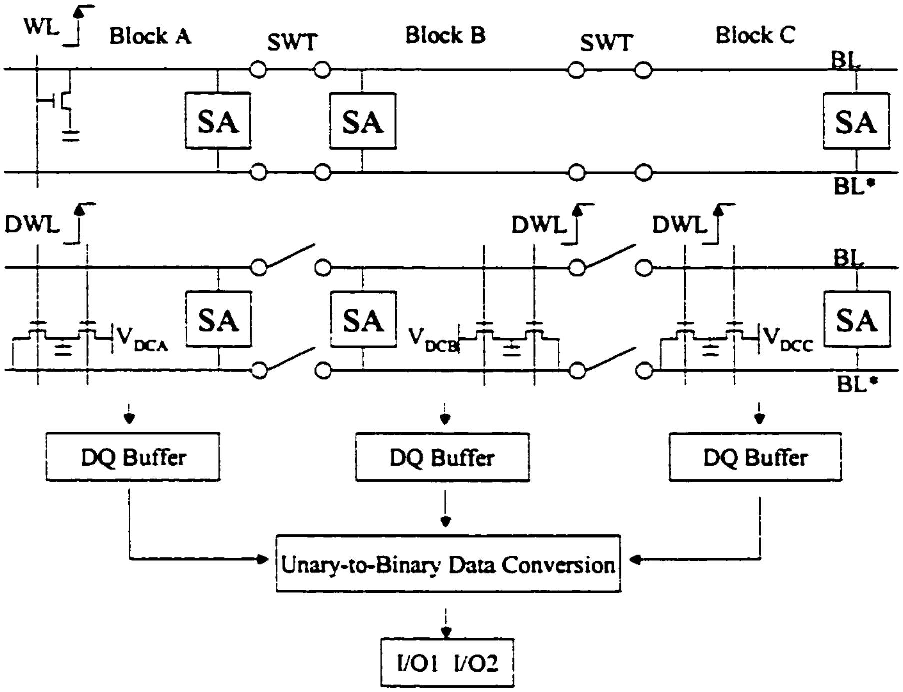


Figure 2-14. The sensing operation in Furuyama's scheme

Figure 2-14 also illustrates the sensing operation. The timing diagram for the control signals in a sense/restore cycle is shown in Figure 2-15. When the circuits are idle, all three sections of the sub-bitlines are connected together via SWT to form one bitline and are precharged to V_{BLP}; at the same time, the dummy cells are connected to the reference voltages via DCP (shown in Figure 2-13). A sensing operation starts with the deactivation of EQL and DCP. Following the timing diagram, a specific wordline is addressed and the charge stored in the memory cell is distributed along one full bitline, as shown in the first row of Figure 2-14. The SWT signal is then deactivated, leaving each sub-bitline in three blocks holding a copy of the attenuated signal voltage. The dummy cells on the opposite sub-bitlines are now activated and the reference voltages stored in them are redistributed along the opposite sub-bitlines. With two sub-bitlines holding the signal and reference voltages, respectively, a sense amplifier in each section is activated and the resulting digital data are latched into the

DQ buffer. Through the data conversion circuits, the 3-bit unary (also called thermometer) codeword is converted into the normal 2-bit digital data according to Table 2-1.

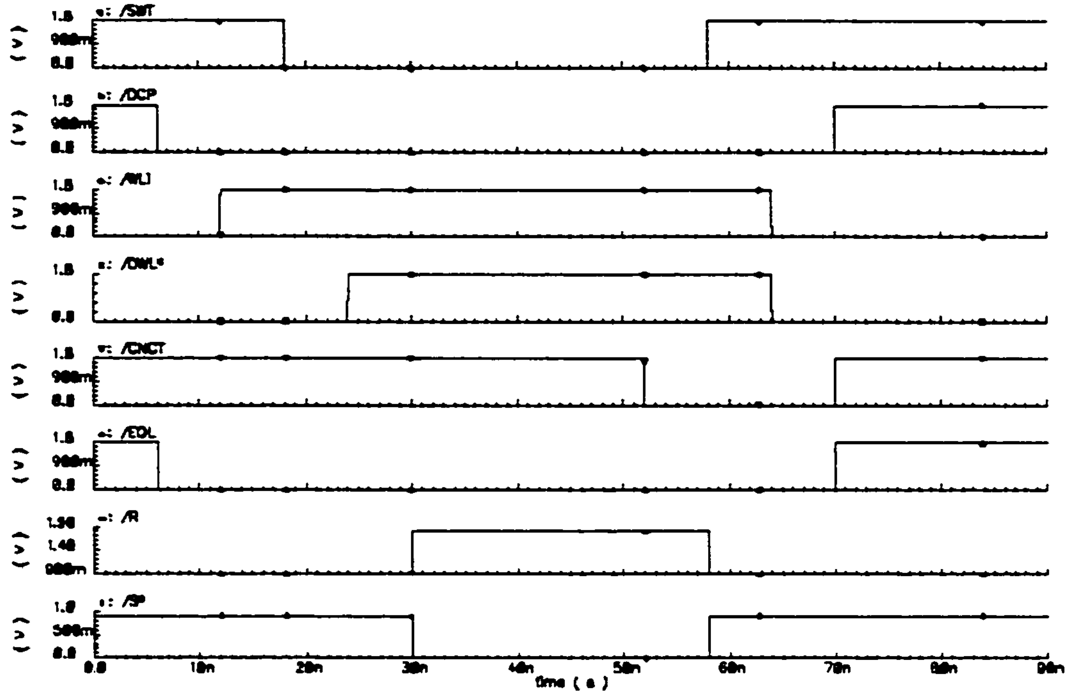


Figure 2-15. Timing diagram for Furuyama's MLDRAM scheme

Table 2-1. Data conversion table

Cell voltage level	Thermometer data from SAs	Two-bit binary data
V_{DD}	111	11
2/3 V_{DD}	110	10
1/3 V_{DD}	100	01
V_{SS}	000	00

Multilevel voltage restoration is completed by charge sharing the sub-bitlines in all three sections. First, the sense amplifiers are disconnected from the sub-bitlines. Then the floating sub-bitlines in the three blocks are connected together by activating SWT. After redistribution of charge along the full sub-bitline, the restored analog voltage is trapped back in the memory cell by deactivating the wordline. Although the three sections are identical, there is still a minor offset error due to the bitline parasitic capacitance during the restore operation. The sub-bitline with the asserted wordline has one memory cell connected to it while the other two do not, which causes the parasitic bitline capacitance of this sub-bitline to be slightly bigger than the others.

One should note that the three global references, V_{DCA}, V_{DCB} and V_{DCC}, are not exactly equal to the reference voltages V_{REF} because the charge in the memory cell is dumped onto three sub-bitlines while the charge in the dummy cell is only dumped onto one sub-bitline. In order to make the voltage changes caused by V_{DC} over one sub-bitline equivalent to those caused by the V_{REF} over three sub-bitlines, V_{DC} should be generated according to the following equation:

$$(V_{DC} - \frac{1}{2}V_{DD}) \bullet \frac{C_c}{C_b + C_c} = (V_{REF} - \frac{1}{2}V_{DD}) \bullet \frac{C_c}{3C_b + C_c} \qquad (2.1)$$

$$V_{DC} = \frac{V_{REF} \bullet (C_b + C_c) + V_{DD} \bullet C_b}{3C_b + C_c} \qquad (2.2)$$

where,

V_{DC} = global reference stored in the dummy cells, V_{DCA}, V_{DCB}, or V_{DCC}

C_c = cell capacitance

C_b = parasitic capacitance of one sub-bitline

V_{REF} = ideal reference voltages, which are 1/6 V_{DD}, 1/2 V_{DD} and 5/6 V_{DD}

The main advantages of Furuyama's MLDRAM scheme are the fast, parallel sensing and relatively simple operation. However, requiring sense amplifiers in all three sections results in greater area overhead, which reduces the potential density gain offered by the multilevel DRAM technology. The biggest disadvantage of this MLDRAM scheme is probably the use of global reference voltages, which are created on-chip, and have to be distributed across the memory array. Any slight inaccuracy in the reference voltages will cause incorrect sensing. Thus this method is vulnerable to process variations, voltage drops and offsets, and induced noise.

2.4.2 Gillingham's MLDRAM

The multilevel DRAM scheme proposed by Gillingham [10] employs similar charge-sharing restore techniques as [9] and a different sequential access sensing method. The circuit schematic of a single column of Gillingham's MLDRAM is shown in Figure 2-16. Bitlines are divided into two equal sub-bitline pairs, left and right, which can be connected in six different ways by the central transistor switch matrix controlled by signals C, C*, X, X*, EL and ER. Each sub-bitline pair has a sense amplifier at one end, which can also be connected or disconnected by the isolation transistors. Note that there is precharge circuitry for both the sub-bitline memory array and the sense amplifiers, which means that these two sections can be precharged independently. Unlike Furuyama's scheme, the dummy cells in this scheme are not used to store the reference voltages but are used instead to cancel out the charge injection caused by the transition of the addressed wordline.

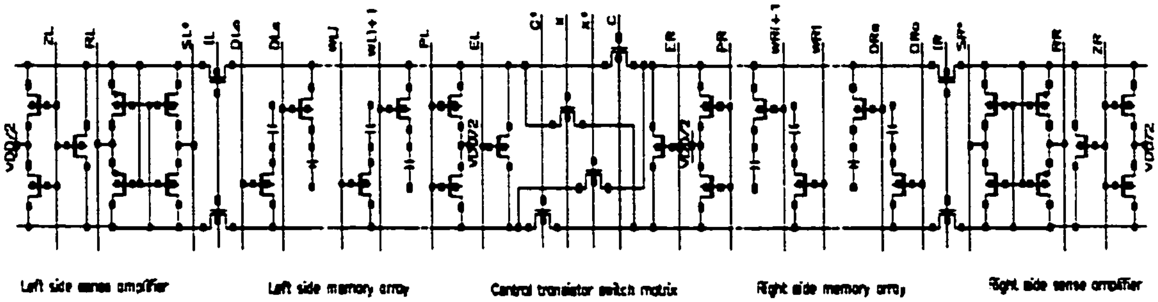


Figure 2-16. Schematic of Gillingham's MLDRAM

The associated control signal timing is illustrated in Figure 2-17. All signals ending with the letter "L" denote signals in the left side and all signals ending with the letter "R" denote signals in the right side.

Sensing of the two-bit data (00, 01, 10 and 11) is performed in two sequential steps: first the most significant bit (MSB) is determined by directly using the V_{BLP} as the reference voltage, then this MSB is used to create the reference voltage for least significant bit (LSB) sensing. Referring to the timing diagram, all sub-bitlines and sense amplifiers are initially isolated and precharged to 1/2 V_{DD}; odd and even dummy wordlines DLo, DLe, DRo and DRe are normally enabled. Before activating an even wordline WLi on the left side, the even dummy wordline DLo on the same side is deactivated so that the charge injection gets cancelled out. Meanwhile, the sub-bitline parasitic capacitance remains the same for four sub-bitlines. After the wordline is activated, signal C* is pulsed to allow the charge in the memory cell to distribute onto the two bottom sub-bitlines. Thus each bottom sub-bitline holds a copy of attenuated cell signal. At this time, by asserting signal IL, the left sense amplifier is connected to the left sub-bitline pair and it senses the MSB by directly using the 1/2

V_{DD} as reference voltage. With the left sense amplifier still holding the MSB, the left sub-bitline pair is isolated from the sense amplifier and the MSB is trapped in the memory cell by deactivating WLi. DLo is asserted immediately after the deactivation of WLi. The next step is to create the reference voltage for the LSB and to sense the LSB. With the left sub-bitlines holding the precharge voltage by giving a pulse to signal PL, signal EL is asserted to short the top and bottom left sub-bitlines together. Then signal C is pulsed to connect the left and right top sub-bitlines together, too. Meanwhile, WLi is activated and the MSB is dumped onto the three sub-bitlines holding 1/2 V_{DD}. After isolating the top right sub-bitline from the left one, the right sense amplifier is connected to the right sub-bitlines and senses the LSB.

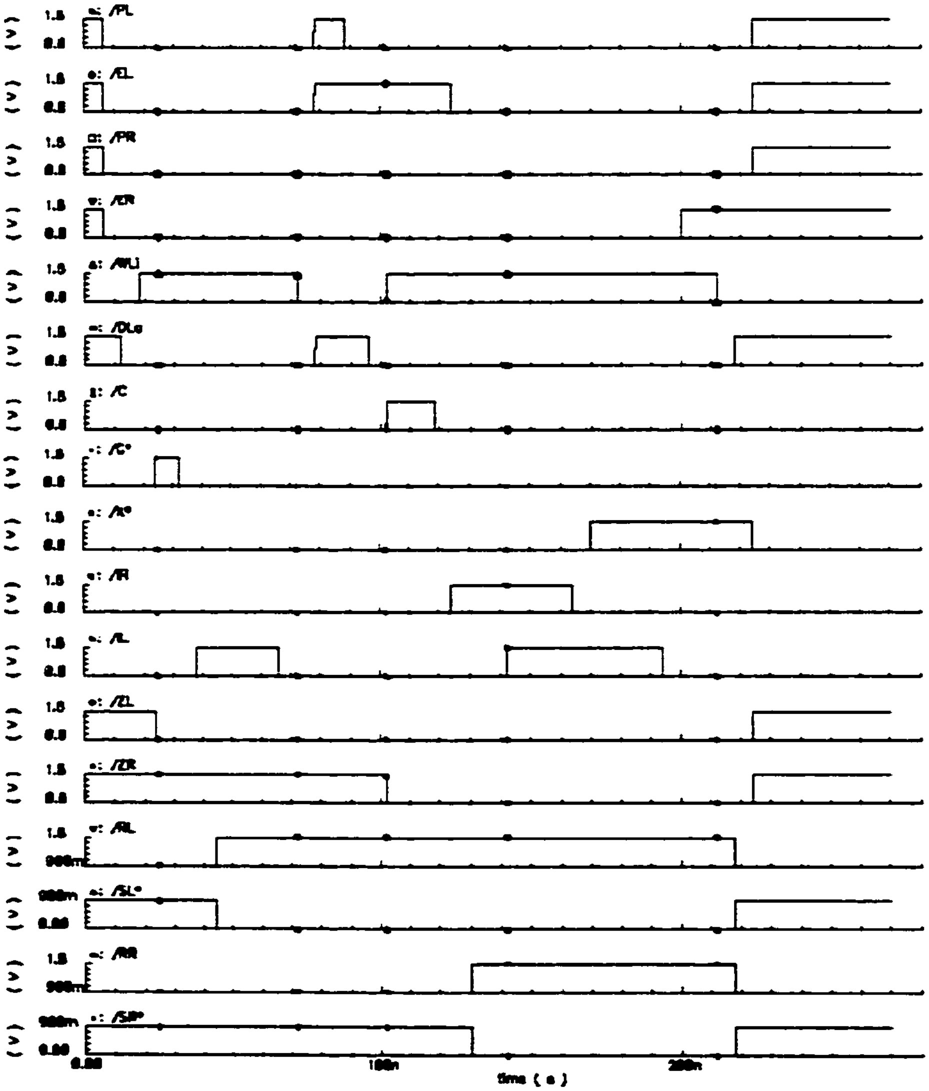


Figure 2-17. Timing diagram for Gillingham's MLDRAM

Notice that in the above sensing operation, the copy of the LSB signal on the right bottom sub-bitline comes from dumping charge of the signal voltage over two sub-bitlines. So the reference levels to be created on the top right bitlines should be defined as the result of dumping the charge of either a 1/6 V_{DD} or 5/6 V_{DD} reference level stored in a cell onto two sub-bitlines. The following equations explain why the charge sharing from MSB (encoded as V_{SS} or V_{DD}) over three sub-bitlines is equivalent to charge sharing of 1/6 V_{DD} or 5/6 V_{DD} over two sub-bitlines. Assuming

r_i = resultant reference voltage

R_i = 1/6 V_{DD} or 5/6 V_{DD}

C_c = cell capacitance

C_b = sub-bitline capacitance

S = MSB (a.k.a. sign bit) voltage V_{SS} or V_{DD},

$$r_i = (R_i - \frac{V_{DD}}{2}) \bullet \frac{C_c}{2C_b + C_c} + \frac{1}{2}V_{DD} = (S - \frac{V_{DD}}{2}) \bullet \frac{C_c}{3C_b + C_c} + \frac{1}{2}V_{DD} \qquad (2.3)$$

$2C_b >> C_c$, $2C_b + C_c \approx 2C_b$, and $3C_b + C_c \approx 3C_b$. Solving equation (2.3) we obtain

$$R_i = \frac{2}{3}S + \frac{1}{6}V_{DD} \qquad (2.4)$$

i.e. 1/6 V_{DD} for $S = V_{SS}$ and 5/6 V_{DD} for $S = V_{DD}$.

Still following the timing diagram, when the LSB is being sensed on the right side sub-bitlines, the left side sense amplifier holding the MSB is connected to the left side sub-bitlines and thus charges the bottom left sub-bitline to the MSB voltage. After the right sense amplifier is isolated and signal X* is asserted to connect the top right sub-bitline and the bottom left sub-bitlines, the left sense amplifier is prevented from charging the two sub-bitlines by deactivating signal IL. Then with two sub-bitlines floating at the MSB voltage (bottom left, top right) and one sub-bitline floating at the LSB voltage (bottom right), signal ER is asserted and the three sub-bitlines are shorted together, creating the correct analog cell restore voltage. Wordline WLi is finally deactivated to trap the cell voltage back to the memory cell and the restore operation finishes. The method of restoring the original cell voltage by charge sharing two bitline capacitances at the MSB voltage and one bitline capacitance at the LSB voltage is described by the following equation and Table 2-2.

$$V_{cell} = \frac{2}{3}MSB + \frac{1}{3}LSB \qquad (2.5)$$

where MSB = V_{DD} or 0 (V_{SS}), and LSB = V_{DD} or 0 (V_{SS}).

Table 2-2. Restore voltage conversion

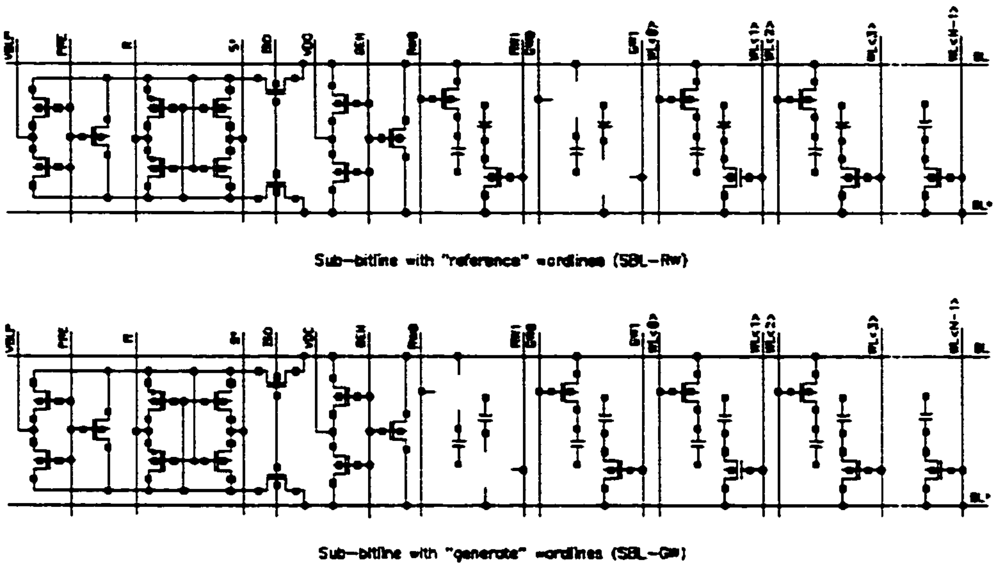
MSB	LSB	Cell voltage
0 (V_{SS})	0 (V_{SS})	V_{SS}
0 (V_{SS})	1 (V_{DD})	1/3 V_{DD}
1 (V_{DD})	0 (V_{SS})	2/3 V_{DD}
1 (V_{DD})	1 (V_{DD})	V_{DD}

The major advantage of Gillingham's scheme is that the reference voltages are generated locally within each column. Therefore small inaccuracies in the reference voltages due to process parameter variations across the chip and voltage drops over long global wires in [9] are avoided. The use of dummy cells and proven standard DRAM blocks also increases the robustness of the design. The main drawback of this scheme is that the control waveforms of the circuits are relatively complex and two sequential steps (first MSB, then LSB) are needed to sense the two-bit data.

2.4.3 Birk's MLDRAM

2.4.3.1 Circuitry

Trying to combine the parallel sensing advantages of Furuyama's scheme and the robustness advantages of Gillingham's design, Birk *et al.* proposed a new MLDRAM scheme [3, 9]. Unlike the previous two, this MLDRAM scheme can be expanded theoretically to any number of levels. Figure 2-18 gives the schematics of two slightly different sub-bitline designs for the new scheme. In each of the two sub-bitlines pairs, from left to right, we have the sense amplifier block, reference voltage generation circuitry, and memory cells. Four special wordlines, called "reference" wordlines (RW) and "generate" wordlines (GW), are connected to the dummy cells that are used to store the reference voltages, to balance charge sharing, and to cancel out noise injection. For the upper sub-bitline pair configuration, only "reference" wordlines are connected to the cell access transistors; while for the lower pair, only "generate" wordlines are connected. As for all wordline notations, the "0" in RW0 and GW0 denotes that wordlines are connected to the true sub-bitline, and the "1" in RW1 and GW1 denotes that wordlines are connected to the complement sub-bitline.



Sub-bitline with "reference" wordlines (SBL-Rw)

Sub-bitline with "generate" wordlines (SBL-Gw)

Figure 2-18. Sub-bitline designs for Birk's MLDRAM

Still taking the 2-bits-per-cell MLDRAM design as our example, Figure 2-19 shows the schematic of the basic building block for the memory core. It is organized as a 3-by-3 array of the sub-bitline pairs shown in Figure 2-18. According to their physical positions, these sub-bitlines pairs are identified as "T" (Top), "M" (Middle), "B" (Bottom) going vertically, and "L" (Left), "C" (Central), "R" (Right) going horizontally. Corresponding sub-bitlines can be connected horizontally via switches controlled by signals SWT0 and SWT1; similarly, corresponding sub-bitlines can also be connected vertically via switches controlled by signals REF0 and REF1.

Dummy cells in the middle row of sub-bitline pairs are used to store the reference voltages, so row M is composed of sub-bitlines with "reference" wordlines. The top and bottom rows consist of sub-bitlines with "generate" wordlines. In order to create three different reference voltage levels for sensing four-level cell signals, the voltage sources "VDC" for the reference generation circuitry (controlled by signal GEN) also differ from one another. Three basic voltage supplies, V_{DD}, V_{SS} and V_{BLP} = 1/2 V_{DD}, are used as "VDC" to create the 1/6 V_{DD}, 1/2 V_{DD} and 5/6 V_{DD} references for the left, central and right sections of sub-bitlines, respectively. Consequently, the "VDC" nodes in TL and BL of sub-bitlines are connected to V_{SS}; the "VDC" nodes in TR and BR of sub-bitlines are connected to V_{DD}; and the remaining VDC nodes are all connected to V_{BLP} (1/2 V_{DD}).

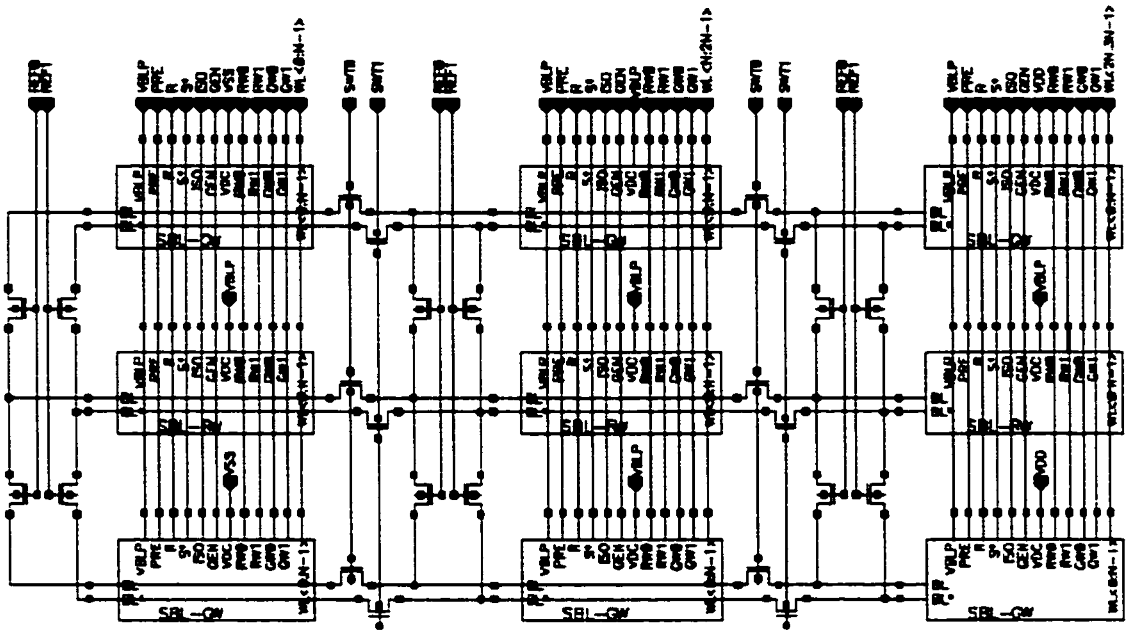


Figure 2-19. Schematic of Birk's MLDRAM [13]

2.4.3.2 Operations

It is easier to explain the operation of Birk's scheme by starting from the time of data restoration or refresh. Referring to the timing diagram in Figure 2-20, we assume that WL0 in the left section has just been accessed and is still active waiting for the restore operation. The sense amplifiers are still activated and connected to the sub-bitlines. As indicated in Table 2-1, the thermometer-coded data is available in the three sub-bitlines for each row. Also, RW1L, RW0C, RW1C, RW0R, RW1R, GW0C and GW0R are all asserted. Note that the position of the addressed wordline is very important here. Which section the wordline is in and whether it is connected to the true sub-bitline or complement sub-bitline will determine the waveforms of the reference and generate wordlines, as well as waveforms of the SWT and REF switch control signals. By deactivating ISO and then pulsing SWT0, the three sub-bitlines floating at the thermometer code voltages are shorted together to create one of the four analog cell voltage levels to be restored. The state of the sub-bitline connections is shown in Figure 2-21. As illustrated in the figure, every true sub-bitline is connected with one memory cell, thus the capacitances for all true sub-bitline are identical to one sub-bitline capacitance plus one cell capacitance. Deactivation of

WL0 now captures the restored analog cell voltages back into the addressed memory cells.

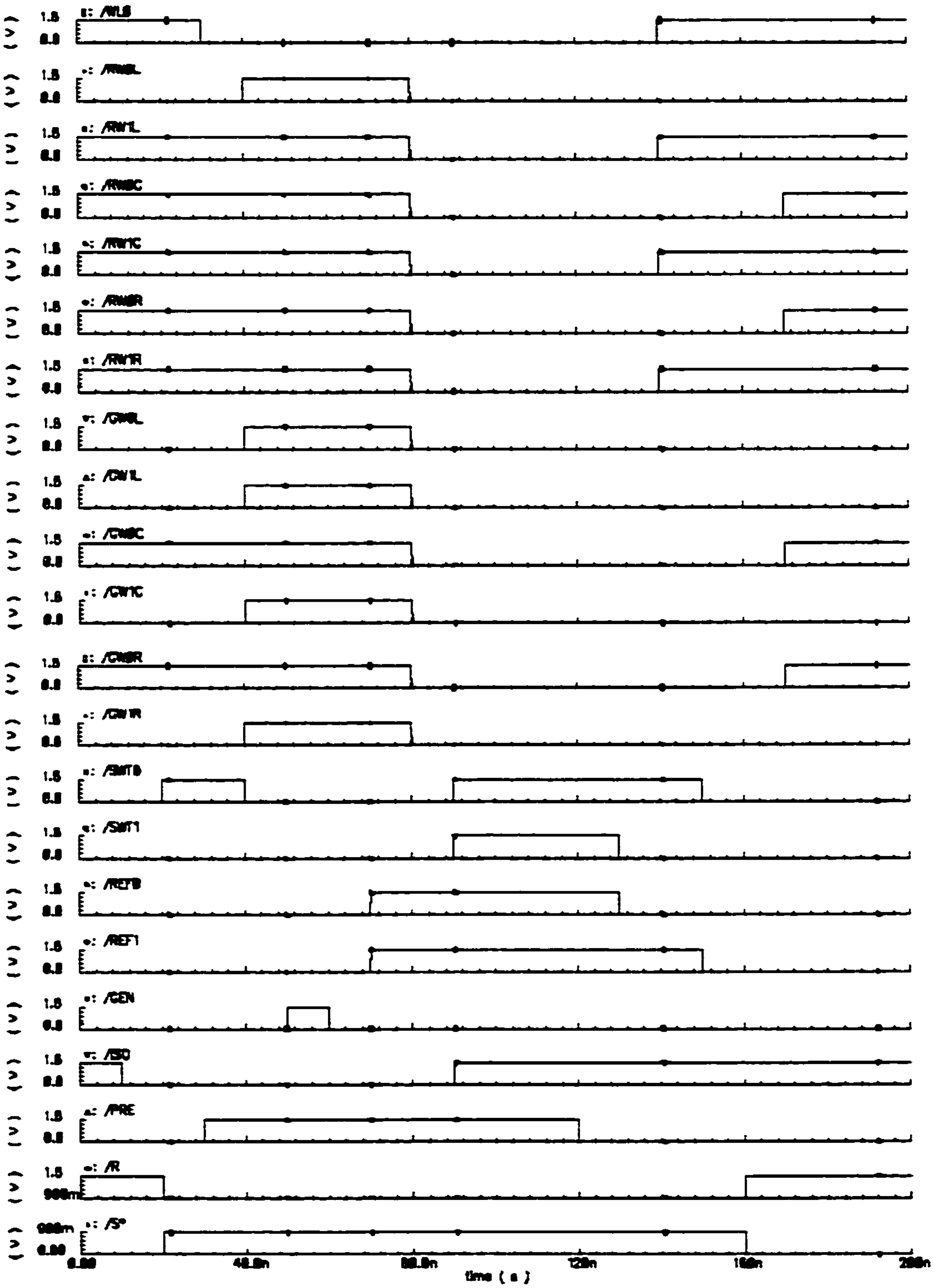


Figure 2-20. Timing diagram for Birk's MLDRAM

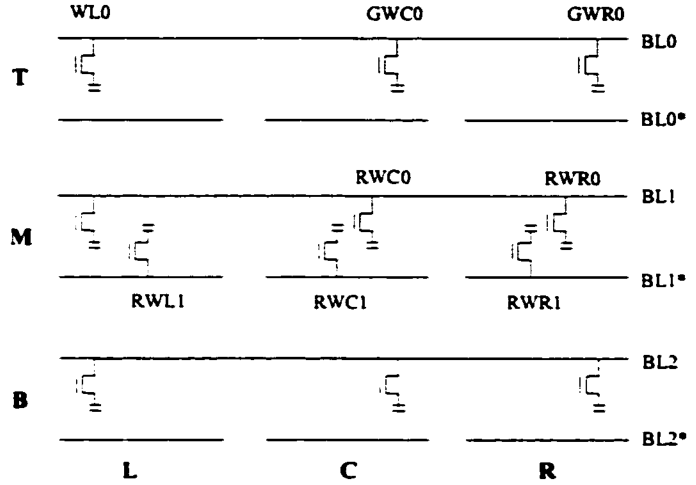


Figure 2-21. Sub-bitlines with memory cells connected for restoring

Prior to the sensing operation, the reference voltages must first be generated. As indicated in Figure 2-20, after a restore operation, all sub-bitlines are isolated from one another. Also, all reference and generate wordlines are asserted, as shown in Figure 2-22. Signal GEN is then pulsed to precharge the sub-bitlines to the corresponding VDC voltages. Then REF0 and REF1 are activated, connecting true sub-bitlines in each section together, and complement sub-bitlines in each section together. After charge sharing along the sub-bitlines, three reference voltages are created in the left, central and right sections, as indicated in Table 2-3. All reference and generate wordlines are then deactivated to trap the desired reference voltages.

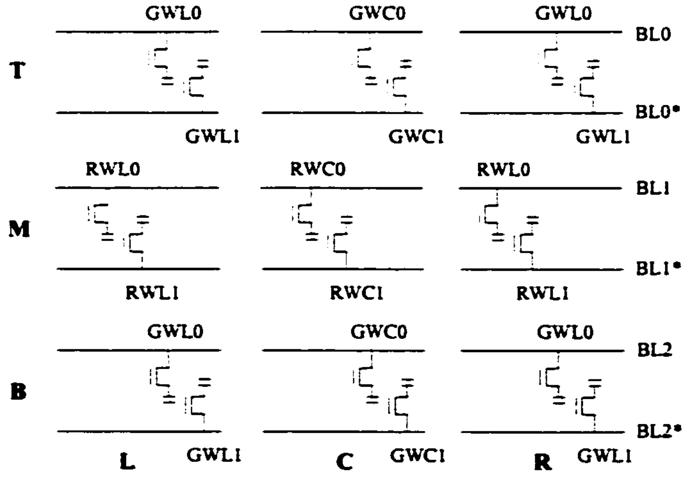


Figure 2-22. Sub-bitlines with memory cells connected for reference generation

With REF0 and REF1 still activated, SWT0 and SWT1 are then asserted so that all sub-bitlines are connected both horizontally and vertically. Meanwhile, sense amplifiers and precharge circuits are connected to the sub-bitlines through asserting ISO. Thus all sub-bitlines are precharged to 1/2 V_{DD}. When PRE stops precharging the sub-bitlines, SWT1 and REF0 are then deactivated, leaving the true sub-bitlines connected horizontally and complement sub-bitlines connected vertically. Activation of WL0 and the three complement reference wordlines RWL1, RWC1 and RWR1 simultaneously distributes the cell voltage along three horizontal sub-bitlines and the reference voltage along three vertical sub-bitlines, as shown in Figure 2-23. The next step is to open switches SWT0 and REF1. Therefore each sub-bitline pair can be sensed independently, and the resulting three-bit thermometer-coded digital data can be read out from the three sections.

Finally, to prepare for the next restore operation, GWC0, GWR0, RWC0 and RWR0 are activated and the circuit returns to the state that was described at the beginning of this subsection.

Table 2-3. Reference generation for Birk's MLDRAM

		Section L Sub-bitline Voltage	Section C Sub-bitline Voltage	Section R Sub-bitline Voltage
Before charge sharing	T	V_{SS}	1/2 V_{DD}	V_{DD}
	M	1/2 V_{DD}	1/2 V_{DD}	1/2 V_{DD}
	B	V_{SS}	1/2 V_{DD}	V_{DD}
After charge sharing		1/6 V_{DD}	1/2 V_{DD}	5/6 V_{DD}

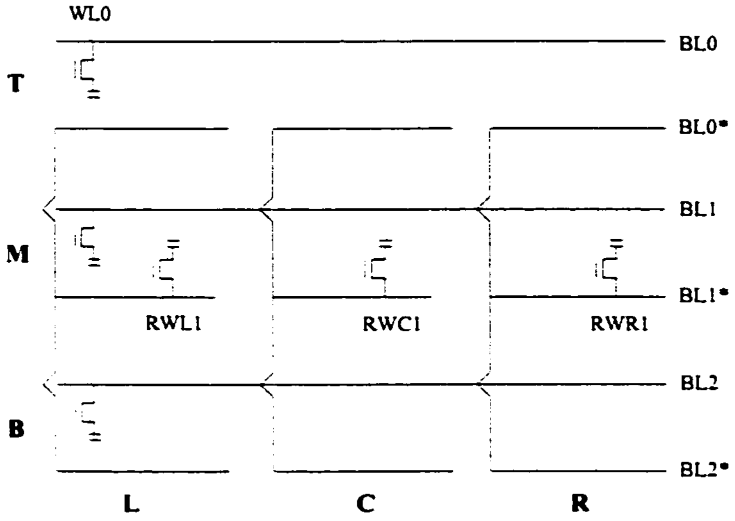


Figure 2-23. Sub-bitlines with memory cells connected for sensing

2.4.3.3 Increasing the Number of Signal Levels

Beyond the 2-bits-per-cell circuit configuration and operation just described, Birk's MLDRAM scheme can be expanded theoretically to a greater number of levels. Recall that *N* different cell voltage levels require *N*-1 reference levels as in equations (1.5) and (1.6), which are repeated here for convenience as equations (2.6) and (2.7).

$$V_{cell} \in \{0, 1, 2, \ldots, N-1\} \frac{V_{DD}}{N-1} \qquad (2.6)$$

$$V_{REF} \in \{1, 3, 5, \ldots, 2N-3\} \frac{V_{DD}}{2(N-1)} \qquad (2.7)$$

By using an (*N*-1) by (*N*-1) array of sub-bitline pairs, *N*-1 evenly-spaced reference voltage levels as in equation (2.7) can be generated and stored for sensing in *N*-1 sections. Since VDC can be connected to either V_{SS} or 1/2 V_{DD} or V_{DD}, charge sharing vertically along *N*-1 sub-bitlines can create any one of the *N*-1 required references. For instance, a 5-level-per-cell configuration requires 4 reference levels, which can be generated in 4 sections according to Table 2-4. Operation of the *N*-levels-per-cell MLDRAM is similar to that of the 4-levels-per-cell MLDRAM.

Table 2-4. Reference generation for a 5-levels-per-cell MLDRAM

	Column 1	**Column 2**	**Column 3**	**Column 4**
Row 1	V_{SS}	V_{SS}	V_{SS}	V_{DD}
Row 2	V_{SS}	V_{SS}	V_{DD}	V_{DD}
Row 3	1/2 V_{DD}	1/2 V_{DD}	1/2 V_{DD}	1/2 V_{DD}
Row 4	V_{SS}	V_{DD}	V_{DD}	V_{DD}
V_{REF}	1/8 V_{DD}	3/8 V_{DD}	5/8 V_{DD}	7/8 V_{DD}

2.4.3.4 Advantages and Disadvantages

The main advantage of Birk's MLDRAM scheme is that the design is very robust against small component variations and not susceptible to most charge injection errors. When a wordline is asserted, the generate and reference wordlines are also asserted to balance out the charge injection. The parasitic capacitances of the corresponding sub-bitline or sub-bitlines are always matched with equal numbers of memory cell(s) connected during the sense and restore operations. Transitions of switch signals cancel out one another most of the time. However, note that some sub-bitlines are connected to two switches; others (i.e., the sub-bitlines at two ends of an array block) are connected to only one switch. Therefore, the charge injection introduced from the transitions of switches is slightly different for these two types of sub-bitlines. Dummy switches or CMOS transmission gate switches must be used to solve the problem. Locally-generated reference voltage levels ensure the reliability of sensing. Another advantage is that, like Furuyama's scheme, data is accessed rapidly at the same time using parallel sensing. Finally, this scheme can be expanded to any number of cell voltage levels.

The main disadvantage of this scheme is the area overhead. The cost paid for fast parallel sensing includes sense amplifier block and reference voltage generation circuitry in each section. Reference and generate wordlines, and horizontal and vertical switches are also sources of additional area overhead. Furthermore, the extra number of sense amplifiers increases the power consumption of the circuit, which is undesirable for many current low-power IC designs. Notice that there is a trade off between the number of levels per cell and the area overhead in this scheme. The greater the number of levels per cell, the greater the area overhead per cell. So finding a practical bit-per-cell limit that allows this MLDRAM scheme to be economical is a very important open problem.

2.5 MLDRAM Challenges

After reviewing three different MLDRAM techniques, it is time to summarize the challenges to overcome for practical MLDRAM designs. The biggest problem is probably the drastically reduced noise margins, regardless of the MLDRAM scheme. As mentioned in Chapter 1, dividing the voltage range from V_{SS} to V_{DD} into N levels decreases the noise margins by a factor of N-1 compared with traditional 2-levels-per-cell DRAM. Consequently, the differential signal voltages to be sensed by the sense amplifier are reduced drastically, making the sensing operation more vulnerable to any imbalance and on-chip noise. Today's lower power supplies aggravate the situation. Meanwhile, reduced noise margins result in shorter data retention times and faster required refresh rates. Vulnerability to soft errors due alpha particle hits will increase as well.

Other problems like increased circuit complexity, longer access time and increased area overhead all need to be considered carefully when designing an MLDRAM. If the density gain of the MLDRAM is not significantly greater than conventional DRAM, the MLDRAM scheme may not be worth the effort.

Therefore, first of all, an MLDRAM design should be robust enough to be able to reject different kinds of noise. It should have a balanced circuit structure and be robust against small process variations. Also, the area overhead should be minimized to increase the die efficiency so that the density gain of MLDRAM over DRAM can be preserved.

The above-mentioned MLDRAM schemes all try to solve one or more of the above problems. Possible solutions to increase the differential signals at the sense amplifiers for MLDRAM also include using a higher memory core supply voltage, employing high dielectric constant materials for the cell capacitors, and especially careful attention to layout design.

Chapter 3 Test Chip Design

3.1 Test Chip Design Overview

The test chip that was designed in this thesis research project is a new implementation of Birk's MLDRAM scheme that takes advantage of the scheme's ability to increase the number of cell signal levels. The test chip, which was simulated and laid out in the 0.18-µm CMOS technology of Taiwan Semiconductor Manufacturing Corporation (TSMC), can be operated either as a conventional 1-bit-per-cell DRAM, or as a 1.5, 2 or 2.5-bits-per-cell MLDRAM. Fractional bits are combined into full bits when groups of two or more cells are considered together. For example, by using three and six cell voltage levels and considering cells in pairs, $3 \times 3 = 9 > 2^3$ and $6 \times 6 = 36 > 2^5$ distinct signal level combinations can be encoded. Therefore it is possible to store 1.5 and 2.5 bits per cell, respectively. Table 3-1 shows how various cell capacities could be obtained by using pairs of cells. The shaded rows correspond to the test chip's four possible operating modes.

Table 3-1. Cell capacities using cell pairs

Number of Levels, N	N^2	$\log_2 N^2$	$[\log_2 N^2]$	Bit(s) per cell
2	4	2	2	1
3	9	3.170	3	1.5
4	16	4	4	2
5	25	4.644	4	2
6	36	5.170	5	2.5
7	49	5.615	5	2.5
8	64	6	6	3

As mentioned in Chapter 1, regardless of the selected cell capacity, data storage and sensing of the test chip follow the same mechanisms described in [3,9]. The ability to adjust the cell capacity will be useful for characterizing the test chip and for experimentally determining the practical limits of Birk's MLDRAM scheme.

Notice that in Table 3-1, using 5 or 7 cell voltage levels will give the same cell capacity as using 4 or 6 levels, respectively. To achieve 3-bits-per-cell density, a 7-by-7 array of sub-bitline pairs would have to be employed to create eight signal levels, which implies an area overhead that is likely to be too much to be acceptable. Therefore, the design of the test chip allows it to operate using 2, 3, 4 and 6 cell voltage levels per cell. Since the maximum number of cell voltage levels is 6, a 5-by-5 array of sub-bitline pairs is employed as the basic construction component for the whole memory array. Inside this 5-by-5 array, one 3-by-3 array can be used to realize 4-levels-per-cell operation, and one of two possible 2-by-2 arrays can be used to realize 3-levels-per-cell operation. Each sub-bitline pair in the memory array can be used directly for 2-levels-per-cell conventional DRAM operation.

In order to create the different reference voltages for the four different operation modes, the reference generation voltage source, VDC, for the sub-bitline pairs must be connected to the proper reference generation voltages. Table 3-2 shows the configuration of the VDC voltages in a basic 5-by-5 array of sub-bitline pairs.

Table 3-2. Configuration of the VDC reference generation voltage sources

VDC		Sections				
		A	B	C	D	E
Sub-bitline pairs	0	V_{SS}	V_{SS}	1/2 V_{DD}	V_{SS}	V_{DD}
	1	V_{SS}	V_{SS}	V_{DD}	V_{DD}	V_{DD}
	2	1/2 V_{DD}	1/2 V_{DD}	1/2 V_{DD}	1/2 V_{DD}	1/2 V_{DD}
	3	V_{SS}	V_{SS}	V_{SS}	V_{DD}	V_{DD}
	4	V_{SS}	V_{DD}	1/2 V_{DD}	V_{DD}	V_{DD}

Referring to Table 3-2, if the entire 5-by-5 array is considered, charge sharing along each section will generate 1/10 V_{DD}, 3/10 V_{DD}, 1/2 V_{DD}, 7/10 V_{DD} and 9/10 V_{DD} in sections A, B, C, D and E, respectively, which are the correct reference voltages for

6-levels-per-cell MLDRAM operation. If only sections B, C, D and sub-bitlines pairs 1, 2,3 are used together, charge sharing in each section will generate 1/6 V_{DD}, 1/2 V_{DD}, and 5/6 V_{DD}, which are the reference voltages for 4-levels-per-cell operation. Similarly, considering sections B and C and sub-bitline pairs 1 and 2 (or sections C and D and sub-bitline pairs 2 and 3) together will create 1/4 V_{DD} and 3/4 V_{DD}, which are the reference voltages for 3-levels-per-cell operation. Conventional 2-levels-per-cell operation does not require any special reference generation circuitry because it can directly take advantage of the 1/2 V_{DD} bitline precharge voltage.

Recall that horizontal and vertical switches are used to connect the sub-bitlines for charge sharing in Birk's scheme. To group only parts of the 5-by-5 sub-bitline array for 4 and 3-levels-per-cell operation, different sets of horizontal and vertical switches have to be employed. Figure 3-1 is a simplified sketch illustrating the horizontal and vertical connection switch settings used in the 5-by-5 sub-bitline array. Note that in fact each switch represents two independently controllable switches for the true and complement sub-bitlines.

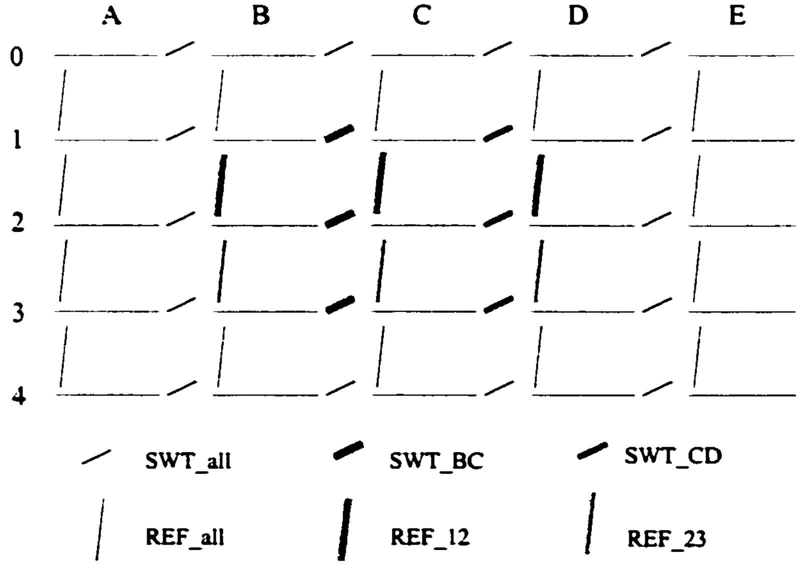


Figure 3-1. The basic 5-by-5 sub-bitline array of the test chip

As we can see from the above figure, there are three sets of switches to group the sub-bitline pairs into different array configurations. SWT_BC and REF_12 together can be used to form a 2-by-2 array. Similarly, SWT_CD and REF_23 together can be used to form another 2-by-2 array. Using SWT_BC, SWT_CD, REF_12 and REF_23 together will form a 3-by-3 array. Lastly, adding the rest of the switches SWT_all and REF_all will form the entire 5-by-5 array. Independent control of these three sets of switches for both the true and complement sub-bitlines makes the test chip operate in the different array configurations and thus different bit(s)-per-cell modes.

Figure 3-2 gives the simplified block diagram of the test chip architecture. In the core, there are five sections of cell arrays that contain equal numbers of cells. Each section of the cell array has a total of 16 wordlines. Four of them are reference and generate wordlines, and 12 are real wordlines that are connected to data storage cells. With 250 bitline pairs, the total number of addressable memory cells in each section is 3000 (12×250=3000). Thus the total number of cells in five sections is 15000 (5×3000=15000), which results in a maximum storage capacity of 37500 (2.5×15000=37500) bits when the test chip is operated in 2.5-bits-per-cell mode.

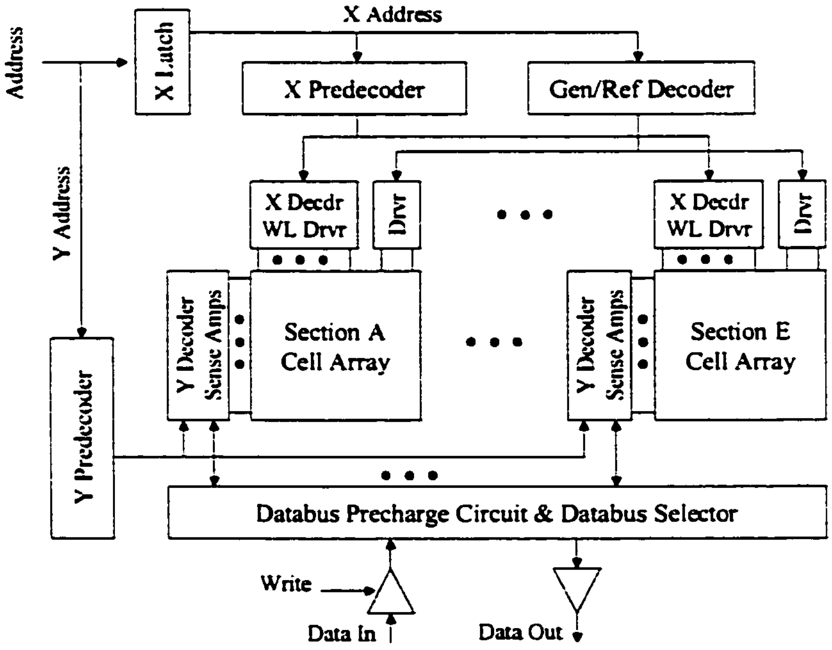


Figure 3-2. Simplified test chip architecture

3.2 The Design Flow

Like a traditional DRAM, the MLDRAM test chip can be divided into two major parts: the periphery and the core. The periphery contains the pads and the logic control circuitry such as decoders, data drivers, etc; the memory core contains mainly the cell array(s), pitch-matched sense amplifiers and wordline decoders/drivers.

The large number of memory cells and pitch-matched components in the core and digital logic gates in the periphery required the MLDRAM test chip design to follow a hybrid design flow, as shown in Figure 3-3. Using TSMC's CMOSP18 technology, the memory core layout design was fully customized so that the layout of memory cells and sense amplifiers would be optimized. All customized layout was implemented in pcells (parameterized cells) using Cadence Skill script codes, which made the layout design flexible and much easier to change. The periphery was implemented using standard cells from the *vst_n18_sc_tsm_c4* and the *tpz973g* black box libraries for 0.18-μm technology provided by CMC. Library *vst_n18_sc_tsm_c4* provides some basic logic gates and library *tpz973g* provides a collection of I/O cells and pads. CMOSP18 has a total of 6 metal layers, but the standard cells use only layer metal1 for internal wiring, leaving the other 5 metal layers for very high-density upper-level on-chip interconnections.

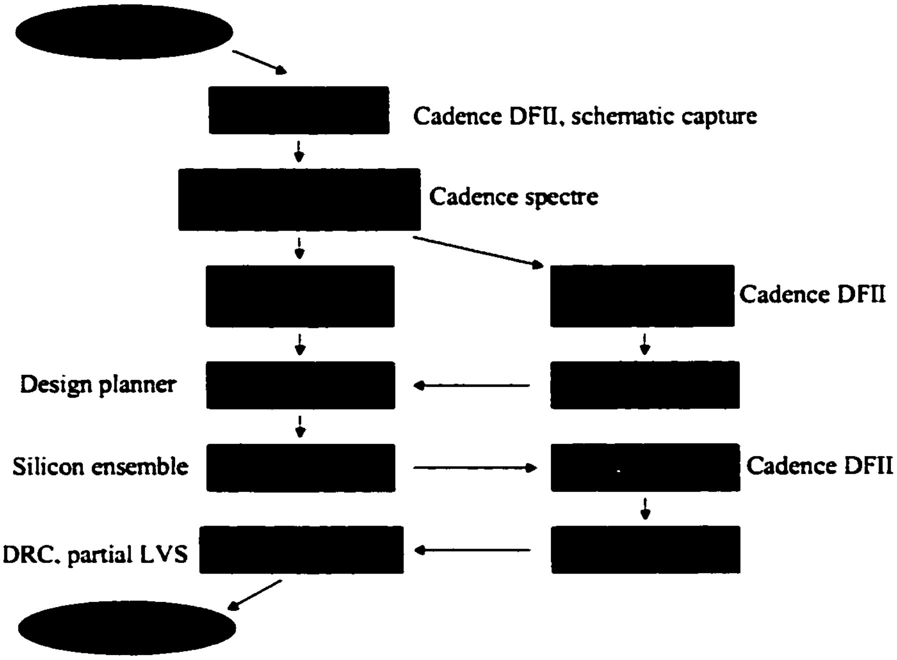


Figure 3-3. Test chip design flow

The test chip was designed according the following steps: First, the test chip schematic was captured in Cadence DFII (Design Framework II). Because the full chip was too big for the analog circuit simulator (Cadence Spectre), the test chip was cut into a small chip having only 10 columns of memory bitline pairs and then simulated in the reduced form. Schematics of the design were modified iteratively according to the simulation results.

After achieving successful schematic simulations, we were ready to build the chip layout. The physical layouts of most of the components in the memory core built in pcells were verified separately by using DRC (Design Rule Check) and LVS (Layout versus Schematic) CAD tools to make sure that the connectivity, the geometry and the spacing were correct and that the layout matched the schematic. Abstract cell views were generated later for the manually-created layouts so that they could be treated as macro-cells during automatic floorplanning and placement. A Verilog netlist of the whole circuit was then created from the schematic and imported into the PDP (Physical Design Planner) for physical floorplanning and placement. At last, Silicon Ensemble (a router tool from Cadence) imported the placed design and routed the standard cells and manually-created macro-cells together. Then the finished layout in DEF (Design Exchange Format) was imported back to Cadence DFII. Post layout verification (like DRC) was repeated to make sure interconnections created by the placer and router did not violate the design rules. For the test chip design, many violations were indeed found during the DRC after placement and routing in the standard logic cell area, so these errors had to be fixed manually. Normally, LVS should be repeated as well to ensure that the netlist matched between the schematics and the layout. Due to the non-standard structure of the memory cells, the cell capacitors could not be extracted automatically from the layout. Therefore, LVS was run separately for most blocks of the design, but not for the whole chip together. A final pad-to-pad simulation was done to the small chip with only 10 columns. However, since the memory cell capacitors could not be extracted from the layout, the memory core could not be extracted. Also, because the periphery was implemented by placing and routing the standard cells, the Verilog netlist name of the pins did not match the pins in the schematic cellview. Thus the final pad-to-pad simulation actually simulated the schematic memory core with only 10 bitline pairs and the extracted standard cells in the periphery without considering the global interconnection parasitics. The last step of the design before submission of the design data was laying out a dummy fill layer on top of the chip layout so that the average density requirement of the poly layer and metal layers would be satisfied.

The final floorplan of the resulting chip is shown in Figure 3-4. The periphery was laid out in standard cells in the upper part of the chip (not shown in accurate proportion in the figure), with the exception of a precharging device that was laid out manually. A customized memory core was imported as a single micro-cell in the lower part of the chip. The chip is surrounded by 56 bonding pads for input/output signals and power supply pins. Four corner cells occupy the chip corners. The silicon area for the test chip is 2796.6 × 2236.6 μm^2 with an absolute cell area of 9.27 μm^2. Silicon area taken per bit in a column for the 6-levels-per-cell operation is 14.54 μm^2/bit, and area per bit is 13.47 μm^2/bit for the 2-levels-per-cell conventional DRAM using the same array architecture. The 68PGA package was selected for the chip packaging.

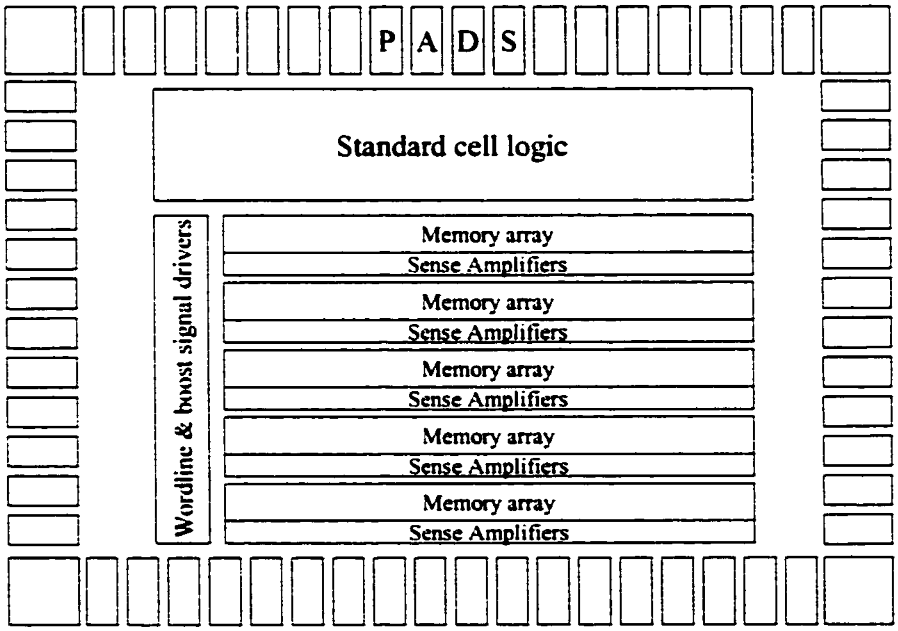


Figure 3-4. Test chip floorplan

3.3 Detailed Design

The periphery and the memory core are the two major parts of the test chip. The periphery contains the pads and logic control circuits such as decoders and drivers, etc. The names and functions of the logic control circuitry are listed below:

- Row Address Register (X_ADDR_reg): A register to latch the row addresses so that the column addresses can be entered later on the same time-multiplexed address inputs.
- Row Section Select Decoder (x_enable_dec): Selects a wordline from one of the five sections of memory arrays.
- Row Pre-decoder (x_pre_dec): Pre-decodes the row addresses into groups of min-terms to avoid large and slow high-fan-in gates.
- Row decoder (x_dec): Selects one wordline from the many wordlines in one memory array by taking the row address minterms from the row pre-decoder as inputs.
- REF and GEN wordline decoder (REF_GEN_dec): Circuitry used to decode the Reference and Generate wordlines from the input row addresses.
- Column Section Select Decoder (y_enable_dec): Selects one bitline pair from the five sections of the memory array.
- Column Decoder (Column_y_dec): Selects a bitline pair from the 250 bitline pairs in one memory array.
- Databus Decoder (block_io): A circuit to decode the column address so that only one out of the 32 databuses will be connected to the data input or output pin; it also contains the data write driver and databus precharge devices.
- Output Buffer (bufferDataOut): A cascade of five inverters used to buffer the output data.
- Signal latches (latched_signal): Latches that are used to create the differential signal pairs for the signal boosting circuitry.

The core contains five sections of memory arrays, pitch-matched sense amplifiers and pitch-matched wordline drivers. The sub-components are described below:

- Memory array: Array of memory cells, including cells for data storage, friendly cells, reference cells, and generate cells.
- Sense amplifiers and data buses: Sense amplifiers and data buses are laid out along the top of each memory array. The sense amplifiers are pitch-matched to each column of the array and each data bus is shared by a group of eight columns.

- Boost signals and wordline driver: Boost circuitry used to boost the signals and wordlines to at least one threshold voltage over V_{DD}.

3.3.1 The Core

3.3.1.1 The Memory Cells

The memory cells were modeled after the memory cell layout of ATMOS' embedded DRAM design in 0.18-μm technology. The cell access transistors are NMOS and the cell capacitors are implemented as a gate-substrate capacitance, as shown in Figure 2-12. The poly layer is used to form the NMOS transistor gate and the common cell plate for the cell capacitors.

The capacitance of a cell capacitor depends on the area of the cell capacitor layout, i.e., the size of the poly and active layer of the cell capacitor. In TSMC's CMOSP18 technology, the calculated gate-substrate capacitance rate for NMOS is 8.54 $fF/\mu m^2$ [20]. The real capacitance is even slightly bigger than this value due to the fringe and coupling parasitic capacitances. Referring to ATMOS' configuration of the memory array and to ensure a safe bitline to cell capacitance ratio, a generous 50 fF of cell capacitance was chosen for the test chip, which requires 5.85 μm^2 of area.

Since the channel capacitance is a non-linear curve depending on V_{GS}, the cell plate voltage was selected to be 2.5 V (i.e. safely greater than $V_{DD} + V_{thn} = 1.8 + 0.4$ V) in order to keep the NMOS transistor channel based cell capacitor in the strong inversion region so that the cell capacitance is stabilized and maximized.

A benefit of the mixed signal process offered by CMC was the deep nwell employed in the memory cell array. Therefore, the separated p-type substrate (pwell) for the memory array could be connected to a back bias voltage of -1.0 V, like in a conventional DRAM. This substrate back bias increases the threshold voltage of the cell access transistors and thus decreases the sub-threshold leakage current of the cell access transistors. It also decreases the junction leakage current from the storage node to the substrate. Thus the retention time of the cells can be maximized. Meanwhile, there will be less noise in the memory arrays because the substrate for memory arrays is better isolated from the other the chip circuits, such as the noisy pad drivers.

3.3.1.2 Friendly Cells and Dummy Cells

Friendly cells are included along the four edges of each memory array in the test chip to make the electrical environment seen by all memory cells the same. The arrangement of the friendly cells is shown in Figure 3-5. On the top and bottom edges of the memory array, there are two friendly wordlines tied to V_{SS}. All memory cells connected to them will never be activated. There is a friendly bitline pair tied to V_{DD} on both the left and right sides of the memory array. Whenever a special wordline is addressed, friendly cells on the left and right sides of the array will be activated as well, but they are never used to store any data.

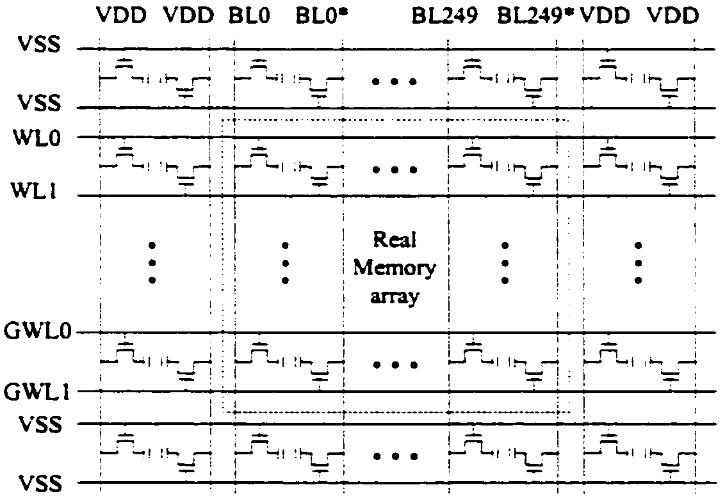


Figure 3-5. Friendly cells in a memory array

The reference and generate dummy cells are connected to the reference and generate wordlines. These dummy cells are exactly the same as the normal memory cells except that the poly layer gates of their access transistors are omitted.

3.3.1.3 The Sense Amplifiers and Sub-bitline Connections

The sense amplifiers used in the test chip are of the basic CMOS latch type shown in **Figure 2-7**. The PMOS and NMOS transistors have exactly the same size. The channel length is 360 nm and the channel width is four times the length. The transistor sizes were chosen based on the experience of the two previous MLDRAM designs, ML2 [6] and ML3 [14].

The PMOS and NMOS transistor sizes are very important for sense amplifier operation. If the sizes are too big, sensing operation might be very fast but writing new data from the data bus to the sense amplifier will be difficult. If the sizes are too small, sensing will be slow and the sense amplifier will be susceptible to noise and soft errors. In order to make the sense amplifiers more robust and also able to reliably sense extremely small differential signals, the sense amplifiers must be designed with extra care to minimize any device parameter imbalances. One way to minimize the effect of small layout imbalances on the sense amplifiers is to increase the channel length of the transistors. If the channel width is kept to be four times the length, the conducting resistance of the transistors will stay the same. At the price of larger layout, sense amplifier transistors with longer channels are more likely to be accurately matched. The impact of process variations against long channel transistors is much smaller than short channel transistors. Traditional sense amplifier designs are effective for resolving signals down to about 100 mV. In the test chip design, new sense amplifiers were not introduced due to time constraints. Instead, large cell capacitors and very short sub-bitlines were used to ensure reasonably strong cell voltages.

Shield lines tied to V_{SS} were employed in the layout design between every real bitline pair as an additional precaution for bitline coupling. Due to the relatively large area taken by the cell capacitors, there was available space between adjacent bitlines, which meant that the shield lines added no area overhead. The relaxed bitline pitch and the shield lines should drastically decrease the coupling noise between adjacent bitline pairs. Bitlines twisting was not used in the test chip due to the lowered expected coupling noise and the short sub-bitlines in each memory array section. The relaxed bitline pitch also made it easier to fit in the sense amplifiers. Recall that the sub-bitlines in the five memory array sections are connected through horizontal switches. Therefore, the bitlines have to run past the sense amplifiers, as shown in Figure 3-6. Note that in this test chip floorplan, bitlines run vertically and wordlines run horizontally. The large bitline pitch made it possible to directly route metal1 for bitline connections in around the side of the sense amplifiers.

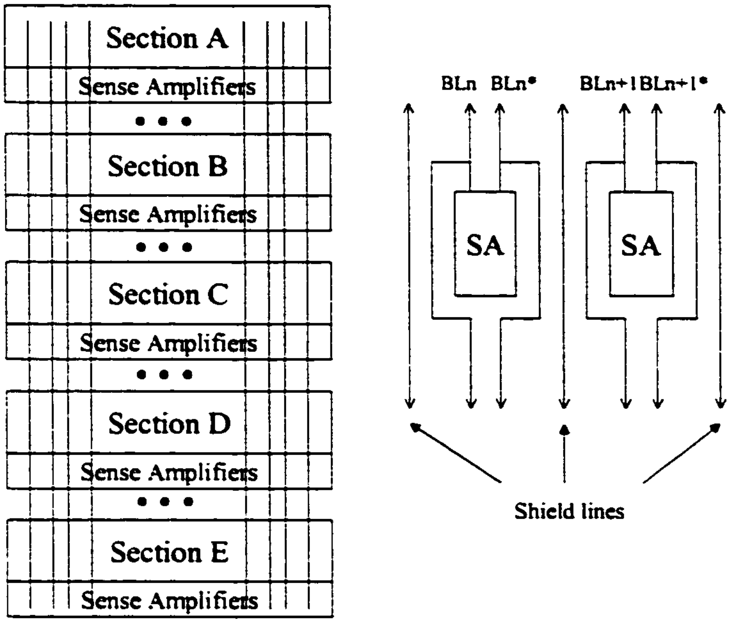


Figure 3-6. Sub-bitline connections between memory arrays

In the layout of the memory core, bitlines run vertically in metal1 and wordlines run horizontally in metal2. Sense amplifiers are at the bottom of the memory array and use only metal1 for interconnections. Data buses run vertically in metal3. Column select signals run in metal4. Both the horizontal and vertical switches are laid out just below the sense amplifiers.

3.3.1.4 Boost Circuits

Voltage boost circuits are used to raise certain logic high signals from V_{DD} to V_{PP}. V_{PP} is at least one threshold voltage higher than V_{DD}. Boost signals are applied to the gates of some NMOS pass transistors in the memory chip to ensure that a full-valued V_{DD} signal can be passed between the transistor switch drain and the source. The voltage boost circuit used in the test chip was designed by ATMOS, as shown in Figure 2-8. The latches used to create the differential signals are actually built in the periphery using standard cells and only the other parts of circuit are laid out in the memory core along the left side of each memory array. The two inverters in the circuit are drivers.

All wordlines, including the reference and generate wordlines, and some switch signals were boosted. All connection switches were implemented as NMOS transistors instead of transmission gates in order to save silicon area and to simplify the chip layout.

3.3.1.5 Row and Column Address Scrambling

Since twisted bitlines were not used in the test chip layout, row address scrambling was straightforward in the test chip. However, each group of eight columns shares one data bus and the data bus addresses are decoded from the 5 LSBs of the column address, thus column addresses do not map directly to their physical positions as row addresses do. Figure 3-7 depicts the row and column address mapping for each section of the memory array.

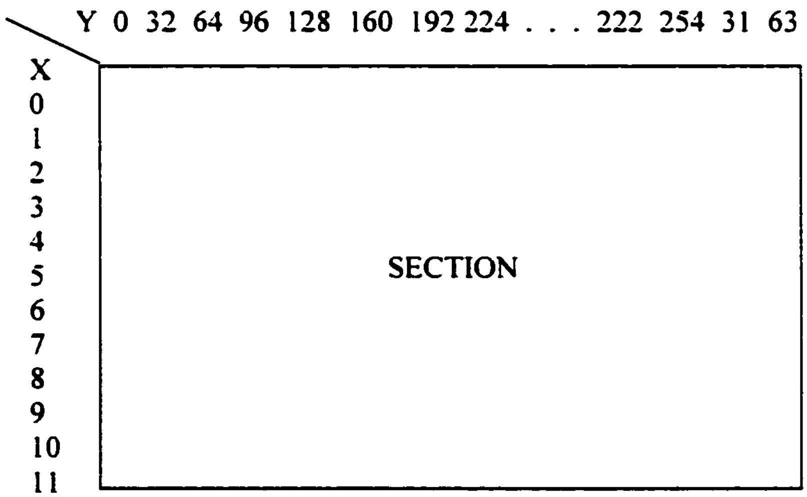


Figure 3-7. Row and column address mapping in one section

Wordline addresses, w: 0 1 2 3 4 5 6 7 8 9 10 11

Physical positions, n: 0 1 2 3 4 5 6 7 8 9 10 11

Column addresses, c: 0 32 64 96 128 160 192 224 1 33 65 97 ... 222 254 31 63

Physical positions, n: 0 1 2 3 4 5 6 7 8 9 10 11... 246 247 248 249

Therefore, for wordline scrambling, $w = n$.

For bitline scrambling, $n = (c >> 5) + (c\&"00011111")*8$,

where c is the binary format of column address, i.e., Y address bits 7 to 0. In other words,

$$c = \left((\text{int})\frac{n}{8} \right) + (n \bmod 8) * 32,$$

where $\left((\text{int})\frac{n}{8} \right)$ is the divident of $\frac{n}{8}$, and "$n \bmod 8$" is the remainder of $\frac{n}{8}$. The ">>5" notation means shift the lefthand argument right by 5 bit positions.

Figure 3-8 shows the detailed cell arrangement in a section of the memory array, in which the bitline number is the physical position number. Notice that in the cell layout, all the even-numbered wordlines are connected to cells on true bitlines, and all the odd-numbered wordlines are connected to the cells on complement bitlines. This special arrangement of the memory cell connections is required by Birk's MLDRAM scheme.

Because the bitlines in each section traverse only 16 wordlines and are very short, and also because there is less coupling noise between adjacent bitlines due to use of shielding bitlines, twisted bitlines were not employed in the test chip layout design.

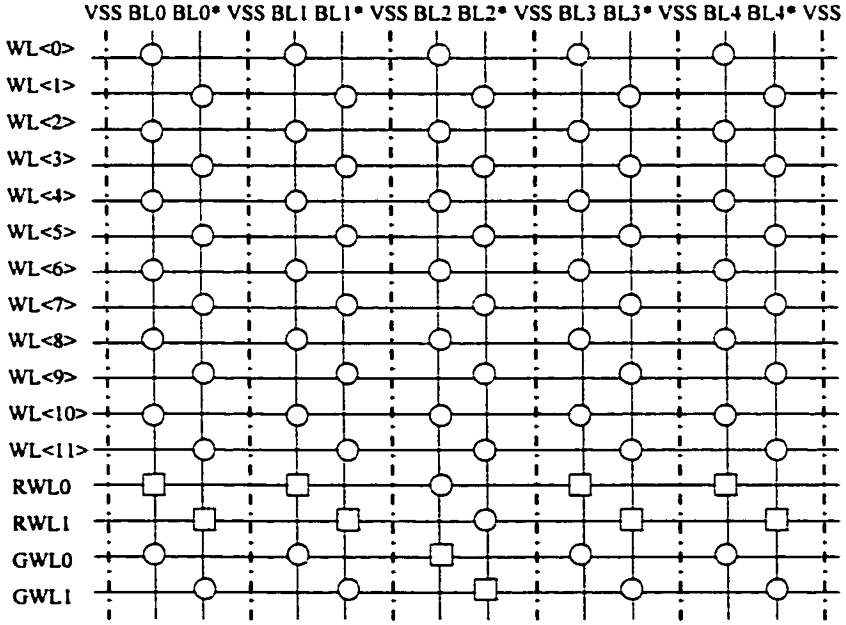


Figure 3-8. Cell layout in one section

□ : A dummy cell without the access transistor;
O: A real cell

3.3.2 The Periphery

3.3.2.1 Row Access

Row access involves accessing one wordline in all five sections of memory arrays. The wordlines, including the normal wordlines and the generate and reference wordlines, are accessed by giving the appropriate row addresses and special waveform control signals.

In the test chip, address input pins multiplex between the row and column addresses. By giving a clock pulse, row addresses are latched at the rising edge of the clock in the row address registers. Column addresses are then passed transparently to the column address decoder.

Each of the five sections of the memory array has 12 normal wordlines as well as four reference and generate wordlines. Seven row address bits are required to select one wordline from all sections of the memory arrays. The four least significant bits are used in the X decoder to select one wordline from the 12 wordlines within each memory array. The timing waveforms of the normal wordlines are controlled by the row address enable signal X_DEC_EN. The three most significant bits of the row address and the X_DEC_EN signal select one section out of the five sections.

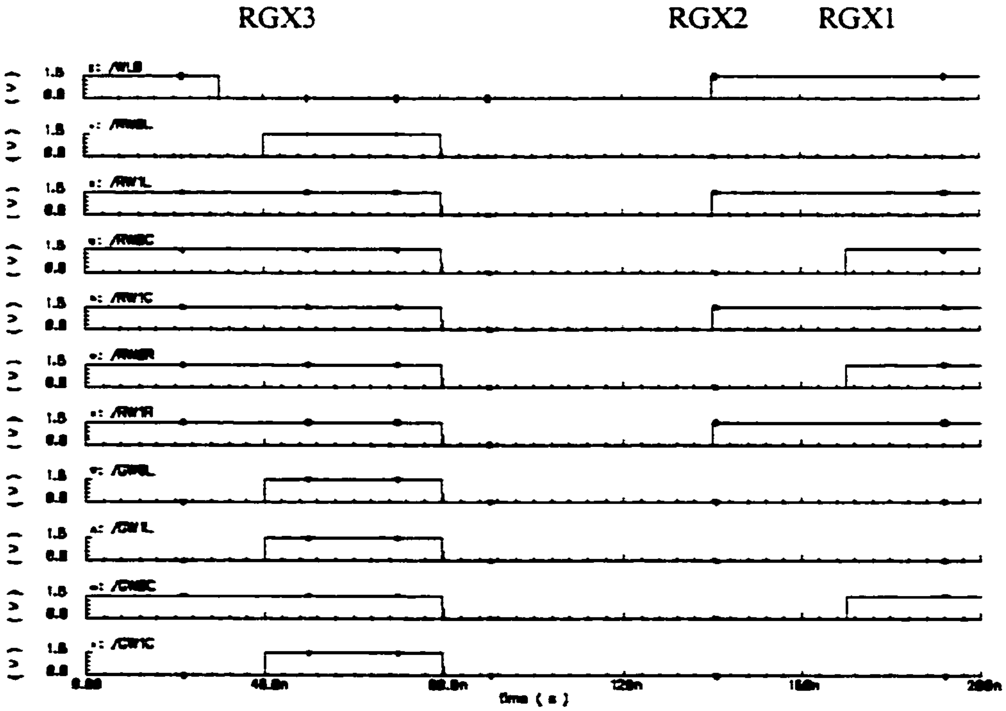


Figure 3-9. Wordline access waveforms

Wordline access use the timing waveforms illustrated in Figure 3-9, which is a portion of Figure 2-20.

Notice that for an addressed wordline, there are three distinct types of timing waveforms assigned to the reference and generate wordlines. As discussed in Chapter 2 concerning Birk's MLDRAM, depending on which section the addressed wordline is in and depending on the whether the addressed wordline is connected to a true or complement sub-bitline, a reference or generate wordline is driven by one of these three waveforms. Address bits 4 to 6 determine the section location of a wordline. Address bit 0 determines whether a wordline address is odd or even, and therefore determines whether the wordline is connected to a true sub-bitline or a complement sub-bitline. In Figure 3-9, from left to right, if we denote the three kinds of reference and generate wordline waveforms as RGX3, RGX2 and RGX1, Table 3-3 specifies the waveform assignment for the reference and generate wordlines given row address bits 6, 5, 4 and 0.

Table 3-3. Assignment of waveforms to the reference and generate wordlines

Signals	Waveform numbers for accessing true sub-bitlines					Waveform numbers for accessing complement sub-bitlines				
	Section A	Section B	Section C	Section D	Section E	Section A	Section B	Section C	Section D	Section E
RW0A	3	1	1	1	1	2	2	2	2	2
RW1A	2	2	2	2	2	3	1	1	1	1
RW0B	1	3	1	1	1	2	2	2	2	2
RW1B	2	2	2	2	2	1	3	1	1	1
RW0C	1	1	3	1	1	2	2	2	2	2
RW1C	2	2	2	2	2	1	1	3	1	1
RW0D	1	1	1	3	1	2	2	2	2	2
RW1D	2	2	2	2	2	1	1	1	3	1
RW0E	1	1	1	1	3	2	2	2	2	2
RW1E	2	2	2	2	2	1	1	1	1	3
GW0A	3	1	1	1	1	3	3	3	3	3
GW1A	3	3	3	3	3	3	1	1	1	1
GW0B	1	3	1	1	1	3	3	3	3	3
GW1B	3	3	3	3	3	1	3	1	1	1
GW0C	1	1	3	1	1	3	3	3	3	3
GW1C	3	3	3	3	3	1	1	3	1	1
GW0D	1	1	1	3	1	3	3	3	3	3
GW1D	3	3	3	3	3	1	1	1	3	1
GW0E	1	1	1	1	3	3	3	3	3	3
GW1E	3	3	3	3	3	1	1	1	1	3

Note that the center shaded regions of the table show the waveform assignments for the reference and generate wordlines that are required for 4 and 3-levels-per-cell operation. Based on Table 3-3, a decoder for the reference and generate wordlines was designed (REF_GEN_dec). The three types of waveforms were then assigned appropriately to the 4 × 5 = 20 internal reference and generate wordlines.

3.3.2.2 Column Access

An 8-bit column address is required to select one column from among the 250 columns inside a section. Three more address bits are required to select one section from among the five sections in the test chip. Column access occurs in two stages: from the bitlines to data buses and from the data buses to the chip input output pins.

Referring to ATMOS' embedded DRAM design, each memory array has 32 differential data bus pairs running vertically in metal3, in parallel with the bitlines which are implemented in metal1. Every eight adjacent sense amplifiers share one data bus pair. Column address bits 5, 6, 7 are used in the column decoder to create the column select signal CSEL, and address bits 8, 9, 10 determine the section of the selected column. To select one data bus pair from the 32 data bus pairs, for connection to the data input output pad of the chip, column address bits 0 to 4 are used. Table 3-4 shows how column select signals are used to connect the sub-bitlines to the data buses in one section. Since there are 250 instead of 256 columns in the test chip, data bus DB31 is used by only two columns.

3.3.2.3 Data IO

Data buses in memory chips are usually bi-directional. To write data onto a data bus, a tri-state buffer is used as a write driver. There are 32 differential data buses in the test chip design, which offers relatively high data bandwidth (32 bits) at the sense amplifier interface. Therefore, 32 differential write drivers are used in the periphery. Data bus select signals DBSEL 31-0, which are decoded from column address bits 0 to 4, select the one write driver to be enabled. For data bus output during read operation, a 32-to-1 multiplixer is used with the same column address bits 0 to 4 as the select signals. Data bus decoding is also shown in Table 3-4.

Table 3-4. Column and data bus decoding in one section

Column Select Signals	Column Address bit A7~A0	Databus pairs	Column Select Signals	Column Address bit A7~A0	Databus pairs
CSEL0	00000000	DB0	CSEL0	00000001	DB1
CSEL1	00100000	DB0	CSEL1	00100001	DB1
CSEL2	01000000	DB0	CSEL2	01000001	DB1
CSEL3	01100000	DB0	CSEL3	01100001	DB1
CSEL4	10000000	DB0	CSEL4	10000001	DB1
CSEL5	10100000	DB0	CSEL5	10100001	DB1
CSEL6	11000000	DB0	CSEL6	11000001	DB1
CSEL7	11100000	DB0	CSEL7	11100001	DB1
......	......	......	......	......	......
CSEL0	00011110	DB30	CSEL0	00011111	DB31
CSEL1	00111110	DB30	CSEL1	00111111	DB31
CSEL2	01011110	DB30	N/A	N/A	N/A
CSEL3	01111110	DB30	N/A	N/A	N/A
CSEL4	10011110	DB30	N/A	N/A	N/A
CSEL5	10111110	DB30	N/A	N/A	N/A
CSEL6	11011110	DB30	N/A	N/A	N/A
CSEL7	11111110	DB30	N/A	N/A	N/A

3.3.2.4 Chip Pads

The test chip has 56 pads in total. It has two sets of V_{DD} power supplies: core power and ring power. The core power supply provides power to the core circuits of the chip, while the ring power supply provides power to the relatively noisy and power-hungry pad I/O cells so that input and output signals will have enough current drive. The number of power supply pads is normally determined by the number of input/output pads. A very general "rule of thumb" is to allow 20% of the pads for power and ground. More power/ground pads should be considered if the majority of the I/O pads are output pads because output pads draw more current than input pads. Although the test chip has only one output pad, in order to distribute the power supplies evenly across the chip, and also because the test chip layout is not pin-limited and had enough space to add more pads, 16 power/ground pads were used. On each side of the chip, there is a V_{DD}_core, a V_{SS}_core, a V_{DD}_ring, and a V_{SS}_ring.

Besides the VDD core power and ring power, there are four other extra power supplies for the test chip. Just as all control signals are provided externally from the input pins, all constant power supplies are provided from off-chip instead of being generated on-chip, as they would be in a production memory part. The power supplies are all listed and explained below:

- V_{DD}_core: core power supply voltage. Set to 1.8V.
- V_{SS}_core: ground return. Set to the 0 V tester reference potential.
- V_{DD}_ring: ring power supply voltage. Set to 3.3V.
- V_{SS}_ring: ground return, same as V_{SS}_core.
- V_{CP}: common cell plate voltage for the memory cell capacitors. Set to 2.5V.
- V_{BB}: memory array substrate bias voltage. Set to –1.0V.
- V_{BLP}: bitline and databus precharge voltage, half V_{DD}_core, Set to 0.9V.
- V_{PP}: Pumped voltage supply for the boosted signals. Set to 2.5V.

Table 3-5 below lists the names of all chip pads, their types (cell name in capital letters and library name in brackets), as well as a brief description of their functions.

Table 3-5. Chip pads

Pad Names	Types	Descriptions
V_{DD}_core, V_{SS}_core (4 pairs)	PVDD1DGZ, PVSS1DGZ (tpz973g)	Core power supplies
V_{DD}_ring, V_{SS}_ring (4 pairs)	PVDD2DGZ, PVSS2DGZ (tpz973g)	Ring power supplies
V_{CP}, V_{BB}, V_{BLP}, V_{PP}	PANALOG (cmosp18)	Extra constant power supplies
Addr<10:0>	PDIDGZ (tpz973g)	Input address bits
Clk	PDIDGZ(tpz973g)	Clock to latch X addresses
Xdec_en	PDIDGZ(tpz973g)	Enable signal for X addresses
Ydec_en	PDIDGZ(tpz973g)	Enable signal for Y addresses
Cnct	PDIDGZ(tpz973g)	To connect the SA to sub-bitlines
Sense	PDIDGZ(tpz973g)	To activate the SAs
D_in	PDIDGZ(tpz973g)	Input data
Write	PDIDGZ(tpz973g)	Write enable signal
Gen	PDIDGZ(tpz973g)	Reference generate signal
Eqln	PDIDGZ(tpz973g)	Activate low precharge signal
D_out	PDO08CDG(tpz973g)	Data output
Swt0_all, Swt1_all, Swt0_bc, Swt1_bc, Swt0_cd, Swt1_cd	PDIDGZ(tpz973g)	Switch control signals to connect sub-bitlines horizontally
Ref0_all, Ref1_all, Ref0_12, Ref1_12, Ref0_23, Ref1_23	PDIDGZ(tpz973g)	Switch control signals to connect sub-bitlines vertically
Rgx1, Rgx2, Rgx3	PDIDGZ(tpz973g)	Waveform control signals for reference and generate wordlines

3.4 Chip Packaging and Fixturing

Packaging and fixturing are integral parts of a microchip design. Packaging provides an interface between the chip and the outside world; fixturing provides an interface (the test fixture) between a chip package and a testing system. Ideally, packaging should not interfere or modify the internal signals in any way. However, electrical characteristics, such parasitic inductance, capacitance and impedance mismatches in the packaging, will result in distortion, signal loss, crosstalk and ringing of signals, especially when frequency and circuit sensitivity increase. Also, the available packages and test fixtures provided by CMC are limited. To fit a specific chip design into an existing generic test fixture, the pin configuration and packaging must be taken into account when designing the layout. There are mechanical constraints in choosing a package as well. In addition to the pin count, the relationship between the package cavity size and the die size has to be considered to avoid tail shorts, shorts or twists between bonding wires, etc.

The PGA68 (ceramic Pin Grid Array package with 68 pins) package was selected for the test chip. It has 17 pins on each side and has a maximum design cavity size of 7.09 mm × 7.09 mm. Typically many metal and ceramic layers are used to accommodate the routing of the signal traces from the die cavity to the pin arrangement underneath the package; these conductors and dielectrics contribute to the undesirable parasitics. Therefore the PGA68 package is not recommended by CMC for designs operating over 50 MHz unless the package parastics are considered during chip design. The first priority was to verify the functionality of the chip, and the through-hole package can be easily fit into lots of sockets for test fixtures, therefore the PGA68 was selected for the test chip. A generic DUT board fixture from Agilent Technologies was used to interface the package to the tester pin drivers and receivers.

Chapter 4 Design Verification and Simulation

The goal of design verification is to verify that the designed circuits meet the specifications before fabrication so that design errors will be detected and corrected. There are many aspects to design verification. Circuit simulation predicts the response of a particular circuit to a particular stimulus. It is an essential step in contemporary circuit design verification. Design verification and circuit simulation of the test chip are discussed in this chapter.

4.1 Design Verification

As mentioned in the test chip design flow, the schematics of the circuits were first captured and verified in Cadence; then the layout was created based on the schematics. Normally, pre-layout circuit simulation is done at first to make sure that the schematic design meets the specifications. Then the circuit is physically laid out in the various layers and a Design Rule Check (DRC) is run to ensure there are no violations of the various physical size constraints that represent the manufacturing requirements. Also, a Layout versus Schematic (LVS) check is run to verify that the circuit layout topology matches that of the circuit schematic. Finally, a post-layout circuit simulation is run using more accurate estimates of the physical parasitic resistances, capacitances and inductances. If the post-layout simulation still meets the circuit specifications, the design is ready for manufacture. Otherwise, the design has to be fine-tuned until all the specifications are satisfied.

The periphery of the test chip is composed of standard cells from the standard cell library provided by CMC, and no schematics were provided for these standard cells. Therefore, the simulation of the periphery could only use the layout-extracted version of the circuit for the basic gates. For the memory core, only schematics could be used in the simulation because the cell capacitor layout could not be extracted by the Cadence Diva extractor. Thus, to simulate the memory array, a cell capacitor of the calculated value 50 fF and an estimated bitline parasitic capacitance of 70 fF per section had to be inserted separately into the schematic. Because the bitline parasitic capacitance is actually determined by the process technology, the bitline parasitic capacitance model of 70 fF per section was inserted based on a conservative bitline to cell capacitance ratio of 7. The LVS check was also only run for some components in the memory core due to the problems encountered during cell capacitor extraction. A final DRC against the whole circuit was run before submission of the design data for fabrication. New rule violations that occurred from standard cell placing and routing were caught and fixed.

4.2 Schematic Simulation Results

4.2.1 Circuit Simulation Background

As feature sizes scale down further and as the number of transistors on a chip increases, integrated circuit designs become larger and more complicated. Prototypes are more expensive to build and more difficult to troubleshoot without many time-consuming simulations. In particular, transistor-level circuit simulation is essential. It enables the designer to verify to various degrees of accuracy that the design meets the specifications. Meanwhile, it also provides additional insight into circuit operation and allows the designer to fine-tune the design prior to its fabrication.

Among the various available circuit simulators, Simulation Program with Integrated Circuit Emphasis (SPICE) is by far the most widely used [21]. SPICE was developed at the University of California, Berkeley, in the early 1970's. In the past several decades, several improved versions of SPICE have been released, such as SPICE2 and SPICE3. Berkeley released a new type of circuit simulator, named Spectre, accompanying with the new release SPICE3 in the late 1980's. Then, Cadence picked it up and made some modifications to improve its speed without losing accuracy and reliability [21]. In a simulation, to numerically compute the response of a circuit to its stimulus, the simulator formulates the circuit equations and then solves them numerically. There are three kinds of simulation analyses in Spectre [21]:

Transient Analysis: Transient analysis solves the nonlinear ordinary differential equations that describe the circuit behavior over a user specified time interval. Based on the user specified stimulus waveforms and initial conditions, transient analysis computes the voltage and current waveforms. It first generates a system of non-linear ordinary differential equations. To solve the equations, the simulator considers a sequence of discrete time steps and converts the problem from the solution of a single system of nonlinear differential equations into solving a sequence of systems of nonlinear algebraic equations. Accuracy and efficiency are heavily dependent on the time step. Also, there is history in the calculations which means the solution at every time point depends on the solution from the previous time point.

DC Analysis: DC analysis solves for an equilibrium point, which is a solution that does not vary with time. An iterative process, called Newton's method, is employed to solve a sequence of linear systems of equations in order to find the solution of a nonlinear system of equations. The process continues until some criteria for stopping are satisfied. The accuracy of the solution depends directly on the stopping or convergence criteria.

AC Analysis: AC analysis computes the steady-state response of the circuit to a small sinusoidal signal. The circuit is linearized about the DC operating point before computing the response due to the small signal inputs.

DC and AC analysis are very important for analog simulations. For digital simulations, transient analysis is commonly used. The simulations applied to the test chip involved only transient analysis to verify the basic functions of the chip. No DC and AC analyses were made of the test chip.

4.2.2 Simulation Results

In this section, the simulations of 6-levels-per-cell operation will be discussed in detail. Simulations of 4 and 3-levels-per-cell operation are very similar to those of 6-levels-per-cell operation. All of these simulations are pad-to-pad simulations using mainly the schematic version of the circuit. Although the core circuits of the chip operate at 1.8 V (V_{DD}_core) internally, the logic high voltage of the input and output signals is 3.3 V (V_{DD}_ring) at the external chip pads, because the I/O cells provided by CMC are driven by the V_{DD}_ring voltage.

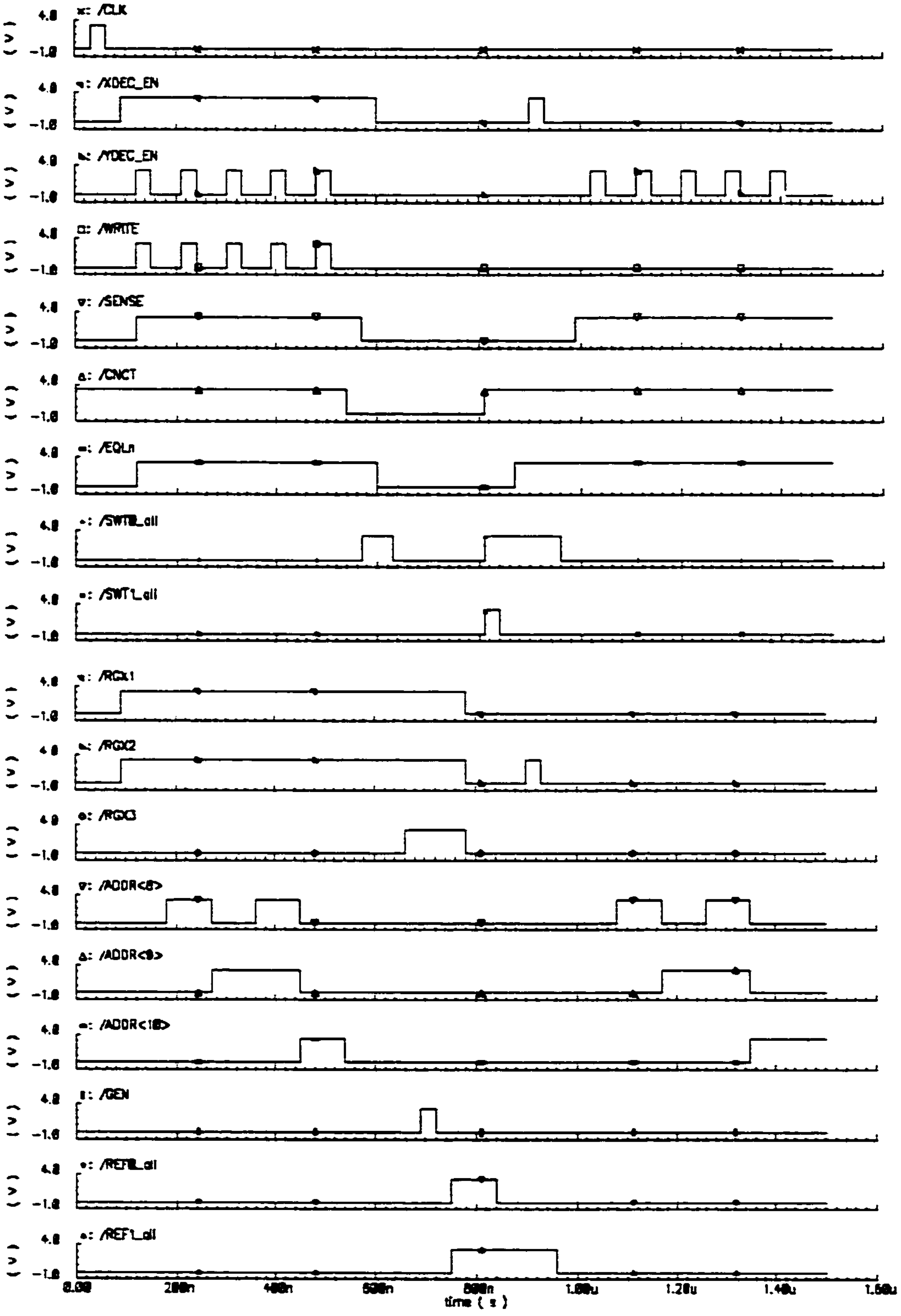


Figure 4-1. Stimulus waveforms for 6-levels-per-cell operation

4.2.2.1 Simulations for 6-levels-per-cell Operation

Figure 4-1 shows the stimulus waveforms for the MLDRAM test chip in the 6-levels-per-cell operating mode. Three address bits (10 to 8) are shown because these three bits are not only the Y address section-select bits, but also are necessary control signals for controlling MLDRAM operation. Address bits 0 to 7 are used purely to select an addressed memory cell (their waveforms are not shown in the figure). Thus one of 15000 memory cells can be selected. The serial input data signal D_IN takes one of six possible values encoded as a 5-bit thermometer codeword, which represent one of 6 cell voltage levels. In order to compare the input (D_IN) and output (D_OUT) data, D_IN is shown in the simulation result waveforms instead of in the input stimulus waveforms. Since in 6-levels-per-cell operation, all five sections of the memory arrays are used together, all horizontal switches and all vertical switches are used together. SWT0_BC and SWT0_CD use exactly the same waveforms as SWT0_all; similarly, SWT1_BC and SWT1_CD use the same waveforms as SWT1_all. Also, REF0_12 and REF0_23 use the same waveforms as REF0_all; while REF1_12 and REF1_23 use the same waveforms as REF1_all. So only four waveforms SWT0_all, SWT1_all, REF0_all and REF1_all are shown in the stimulus waveforms. All signal transitions in the stimulus waveforms are separated by thirty nanoseconds, which corresponds to safe relaxed timing.

In this simulation example, the first memory cell located in Section A, which is addressed by WLA<0> and column <0>, is accessed. When the simulation starts, at time 0 ns, all switches are turned off and the five memory array sections are disconnected. Signal CNCT is "high", connecting the SAs in each section to the corresponding BLs. All BLs are precharged to 1/2 V_{DD} through the precharge circuits in the SA blocks. All real wordlines and reference and generate wordlines are turned off. The row addresses are applied and ready to be latched. Then at time 30 ns, CLK is pulsed to latch the 7 bits of the row address. At time 90 ns, XDEC_EN is asserted and the addressed wordline goes to the boosted voltage V_{PP}. Signal RGX1 goes high, turning on the true reference wordlines and generate wordlines in sections other than the section containing the addressed real wordline. Therefore, all of the true sub-bitlines have one memory cell accessed and therefore have equal capacitance. Meanwhile, the address input pins are changed to provide 8 bits of column address to select one column from the 250 that are present in each section. Then the precharge circuits stop precharging the sub-bitlines and the sense amplifiers are activated at time 120 ns. At the same time, YDEC_EN and WRITE are activated to write the input D_IN into the addressed sub-bitlines. From time 120 ns to 510 ns, input address bits 8 to 10 changed serially to choose the corresponding sub-bitlines in Sections A, B, C, D and E. YDEC_EN and WRITE pulse five times to write the 5-bit thermometer-coded data into the five sections. Sense amplifiers are isolated from the sub-bitlines at time 540 ns by deactivating CNCT, leaving the addressed sub-bitlines in the five sections floating at either V_{DD} or V_{SS}. Then, the horizontal switches controlled by SWT0_all, SWT0_BC and SWT0_CD connect the true sub-bitlines in all five sections from time 570 ns to 630 ns, charge sharing the sub-bitline voltages and creating one of the six possible analog voltage levels. The addressed wordline goes down by deactivating XDEC_EN at 600 ns and traps the analog data in the memory cell. This finishes the write operation. If YDEC_EN and WRITE are not pulsed, the row access alone will cause a restore operation for the data already available in the sub-bitlines.

During the time interval from 660 ns to 780 ns, five reference voltages are generated. RGX3 goes high at 660 ns so that all reference and generate wordlines are activated. Pulsing GEN precharges the sub-bitlines to one of V_{DD}, V_{SS} and 1/2 V_{DD}, depending on the particular voltage connected to the VDC for that column, as indicated in Table 3-2. At time 750 ns, all REF signals are asserted, vertically connecting every group of five true sub-bitlines and five complement sub-bitlines. After charge sharing along the five columns, an appropriate reference voltage is created in each section. For example, in section A, four sub-bitlines are connected to V_{SS} and one is connected to 1/2 V_{DD}. After charge has been shared, 1/10 V_{DD} is created. Similarly, 3/10 V_{DD}, 5/10 V_{DD}, 7/10 V_{DD} and 9/10 V_{DD} are created in sections B, C, D and E, respectively. All RGX signals are deactivated at 780 ns and these reference voltages are trapped in the reference cells and generate cells. However, only the reference cells are used later on to provide the proper reference voltages for the sensing operation.

All horizontal sub-bitlines in the five sections are connected together by asserting SWT0 and SWT1 at time 810 ns. Notice that the vertical switches are still turned on. Meanwhile, CNCT connects the SAs to all the sub-bitlines and precharges them all to 1/2 V_{DD}. SWT1 and REF0 go low at time 840 ns, leaving five true sub-bitlines connected horizontally and five complement sub-bitlines connected vertically. The sub-bitline precharging stops at time 870 ns to minimize the charge injection when deasserting the switch signals. At time 900 ns, pulsing XDEC_EN and RGX2 asserts the addressed wordline and dumps the signal voltage from the addressed cell onto its full-length true bitline; simultaneously, the reference wordlines RW1 in sections A, ..., E are asserted and dump their reference cell signals onto the full-length complement sub-bitlines. Then prior to sensing, SWT0 and REF1 are deactivated, disconnecting all sub-bitline pairs. Parallel sensing with all sub-bitline pairs is then started by asserting the SENSE signal. By changing address bits 10 to 8 and pulsing YDEC_EN signal, each bit of the thermometer encoded data is read out serially from sections A to E.

Figure 4-2 gives the simulated sub-bitline signals for 6-levels-per-cell operation. The 6 cell signal levels and 5 reference signal levels are shown superimposed on top of each other to show more clearly the cell signal and reference signal spacings. As discussed above, all sub-bitlines are precharged before time 120 ns when the SAs are activated. Then from 120 ns to 510 ns, the five-bit data is written into the sub-bitlines. The 6 equally-spaced cell signal voltage levels from 570 ns to 630 ns, and the 5 equally-spaced reference voltage levels from 750 ns to 810 ns are shown in the figure. Sensing starts at 900 ns and the data are read out serially starting at time 1020 ns. Note that the attenuated differential signals shifted down at time 960 ns right before activating the sense amplifiers because of the de-asserting of the NOMS horizontal and vertical switches. Negative charge injection was received by all of the sub-bitlines.

The simulation results for the six possible inputs of "00000", "10000", "11000", "11100", "11110" and "11111" are shown in Figures 4-3 to 4-8. The output signal D_OUT is normally high because of the precharge in the databus. D_IN and D_OUT are only valid when data being written or read by setting YDEC_EN and/or WRITE to high. From the simulation waveforms, it can be observed that the output data (D_OUT) read during time 1020 ns to 1410 ns are always the same as the input data (D_IN) written earlier to the cell during time 120 ns to 510 ns. The other internal signals were "probed" in simulation to verify that the addressed wordline is activated at the right time and that the addressed sub-bitlines are being written and sensed properly.

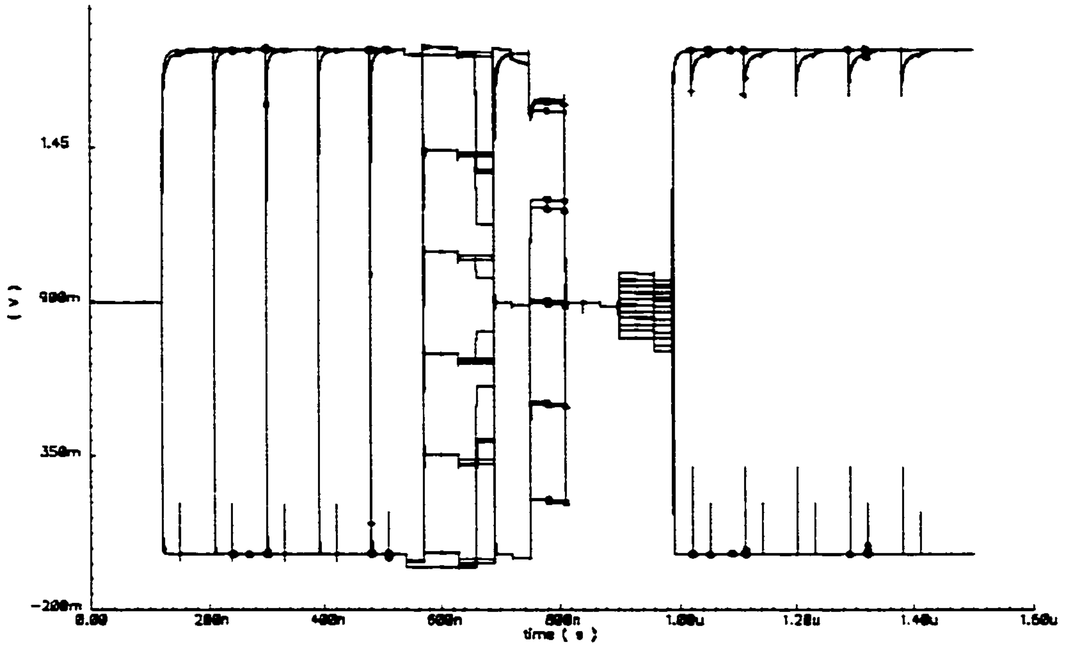


Figure 4-2. Bitline and reference signals in 6-levels-per-cell operation

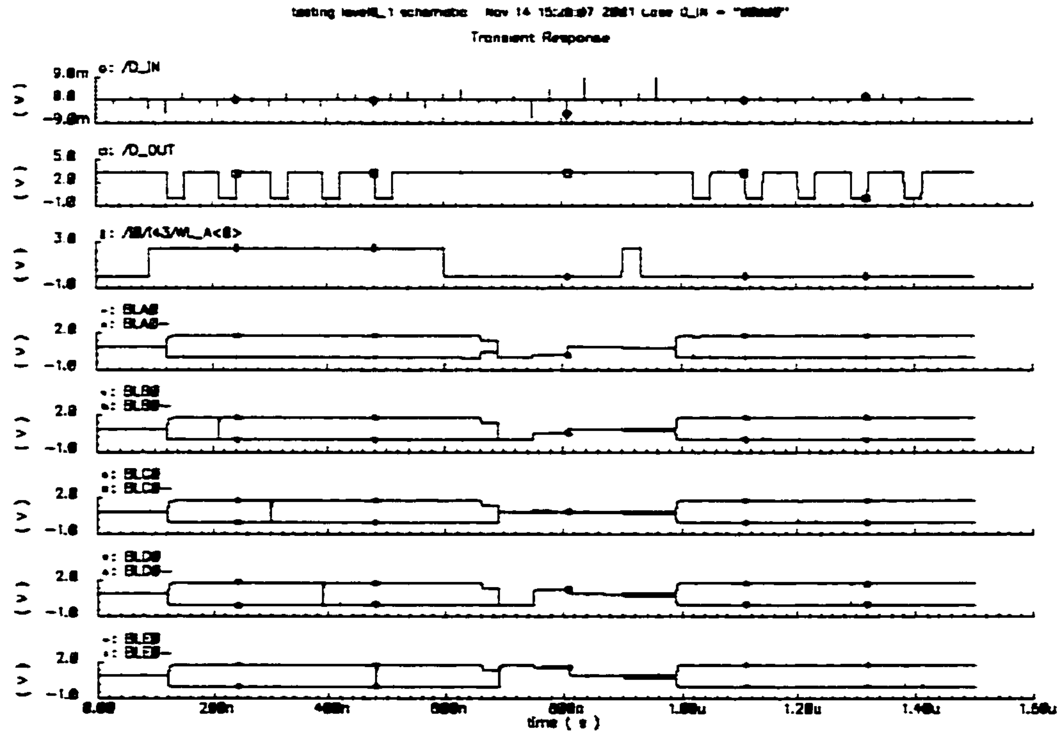


Figure 4-3. Simulation waveforms for data = "00000"

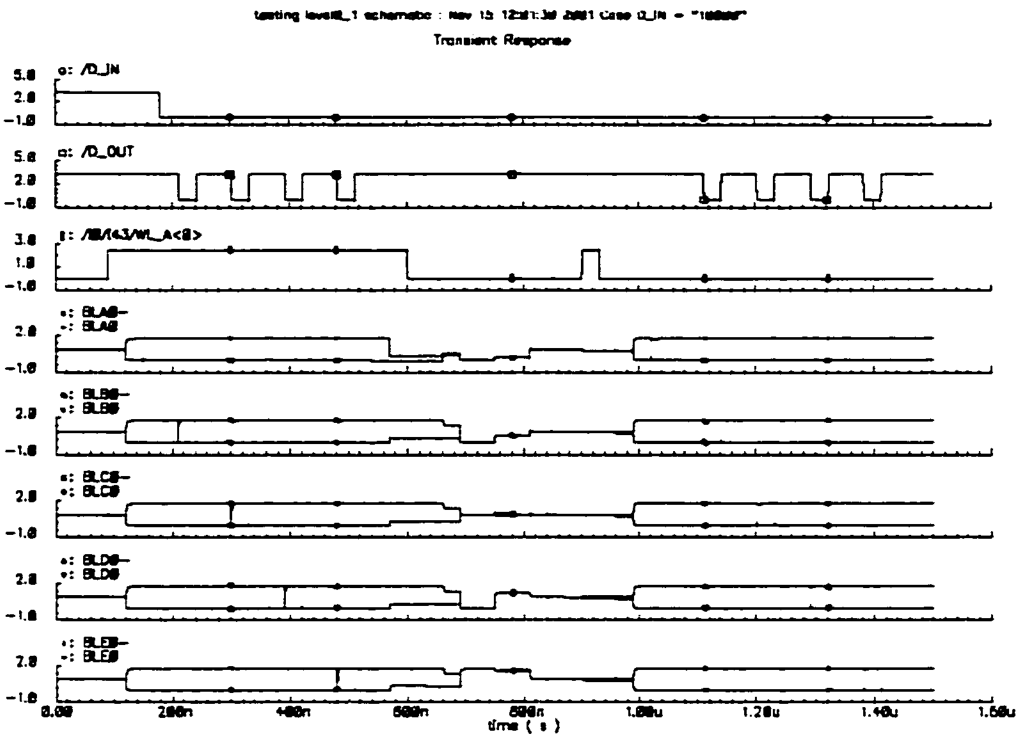


Figure 4-4. Simulation waveforms for data = "10000"

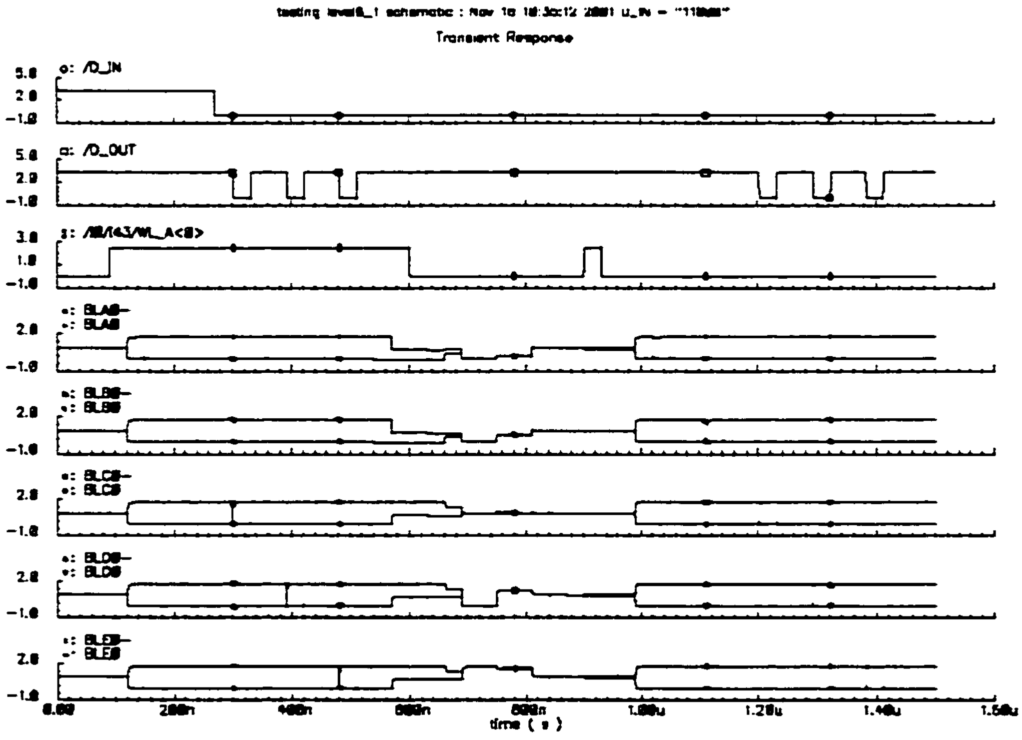


Figure 4-5. Simulation waveforms for data = "11000"

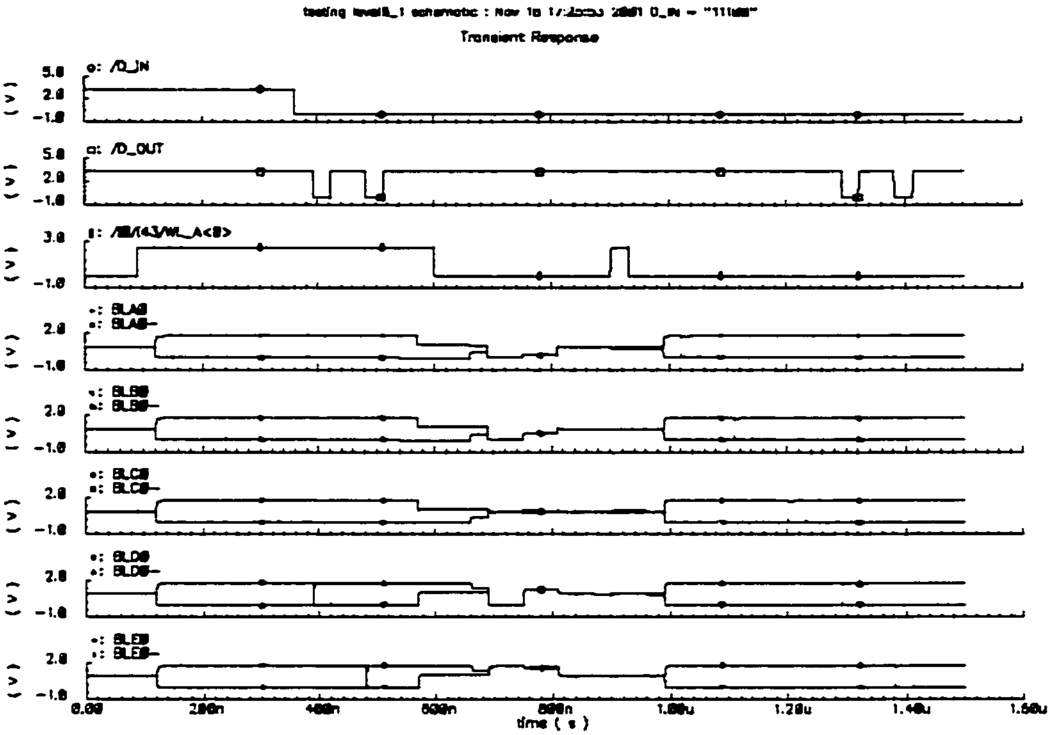


Figure 4-6. Simulation waveforms for data = "11100"

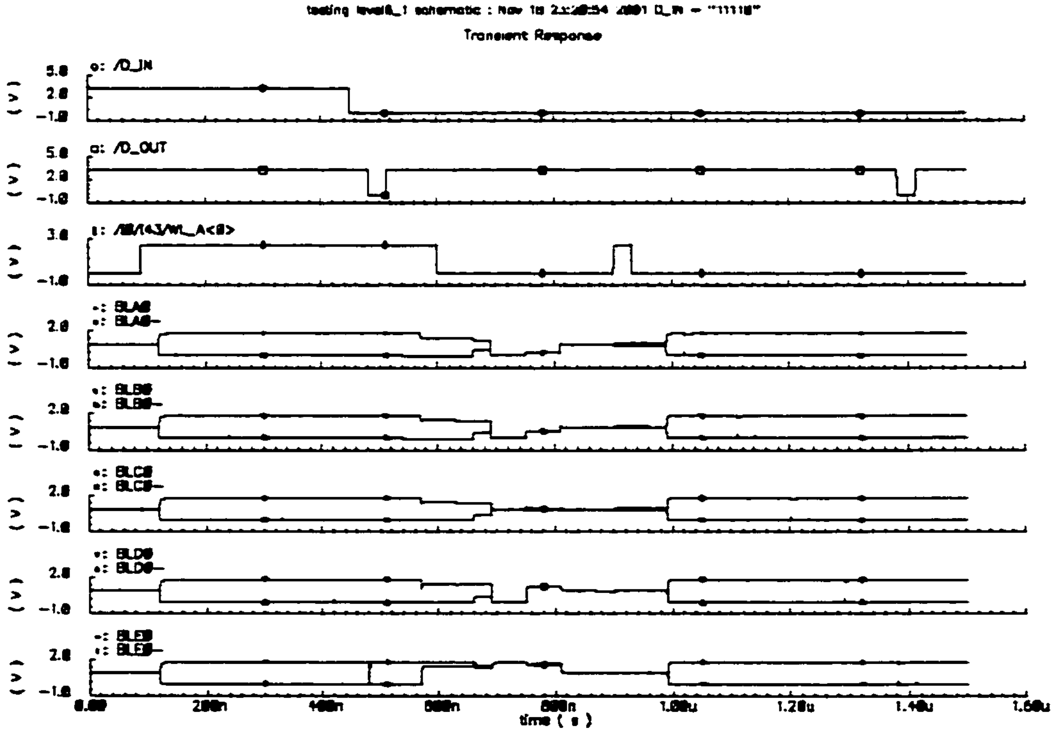


Figure 4-7. Simulation waveforms for data = "11110"

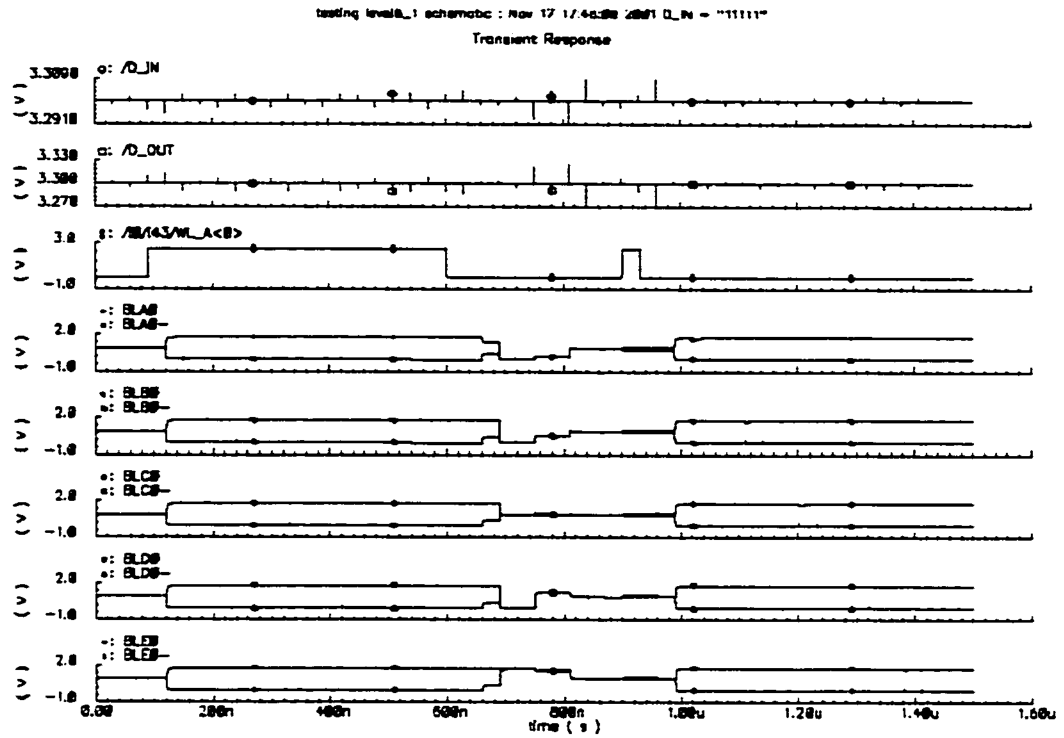


Figure 4-8. Simulation waveforms for data = "11111"

4.2.2.2 Simulations for 4-levels-per-cell and 3-levels-per-cell Operation

The signal control sequences for the 4 and 3-levels-per-cell operating modes are similar to those of the 6-levels-per-cell mode. However, in the 4-levels-per-cell operating mode, three sections, namely sections B, C and D, are used to create the three required reference signals, 1/6 V_{DD}, 1/2 V_{DD}, and 5/6 V_{DD}. YDEC_EN and/or WRITE are pulsed three times when writing/reading the data into/from these three sections. The four possible thermometer codewords are "000", "100", "110" and "111". In the 3-levels-per-cell operating mode, two sections, either B and C or C and D, are used to create the two required reference signals, 1/4V_{DD} and 3/4V_{DD}. Data codewords "00", "10" or "11" are written/read serially into/from the two sections by pulsing YDEC_EN and /or WRITE two times.

In order to more directly compare the noise margins for MLDRAM, the control sequences for 4-level and 3-level operation were slightly modified from the 6-levels-per-cell sequence: just before powering up the sense amplifiers, both the cell and reference signals were charge shared over full-length horizontal and vertical bitlines each, i.e., over all five sub-bitlines connected horizontally or vertically together. This was done to simulate the situation of a 2-bit (4-level) or 1.5-bits-per-cell (3-level) chip that has the same bitline length as in the actual 2.5-bits-per-cell (6-level) test chip.

When operating in 4-levels-per-cell mode, only the wordlines in section B, C or D should be addressed. Also, only the middle three of every group of five columns should be addressed. The addressable column numbers are given by the following equation:

$$N = a + 5b \qquad (4.1)$$

where N is the column number. For example, in column 22, a is either 1 or 2 or 3, and $b \in \{0, 1, 2, \ldots, 49\}$.

Similarly, when operating in 3-levels-per-cell mode, the addressable column numbers can also be expressed by equation (4.1). However, a is either 1 or 2 when sections B and C are used; and a is either 2 or 3 when sections C and D are used. Figure 4-9 gives the input waveforms for 4-levels-per-cell mode. Figure 4-10 shows the superimposed bitlines and reference signals. Figures 4-11 to 4-14 illustrate the simulation output signals. Figure 4-15 gives input waveforms for 3-levels-per-cell mode. Figure 4-16 shows the superimposed bitlines and reference signals. Figure 4-17 to Figure 4-19 illustrate the simulation output signals. As observed from the result waveforms, the read output data (D_OUT) are always identical to the written input data (D_IN) in both the 3 and 4-levels-per-cell operating modes.

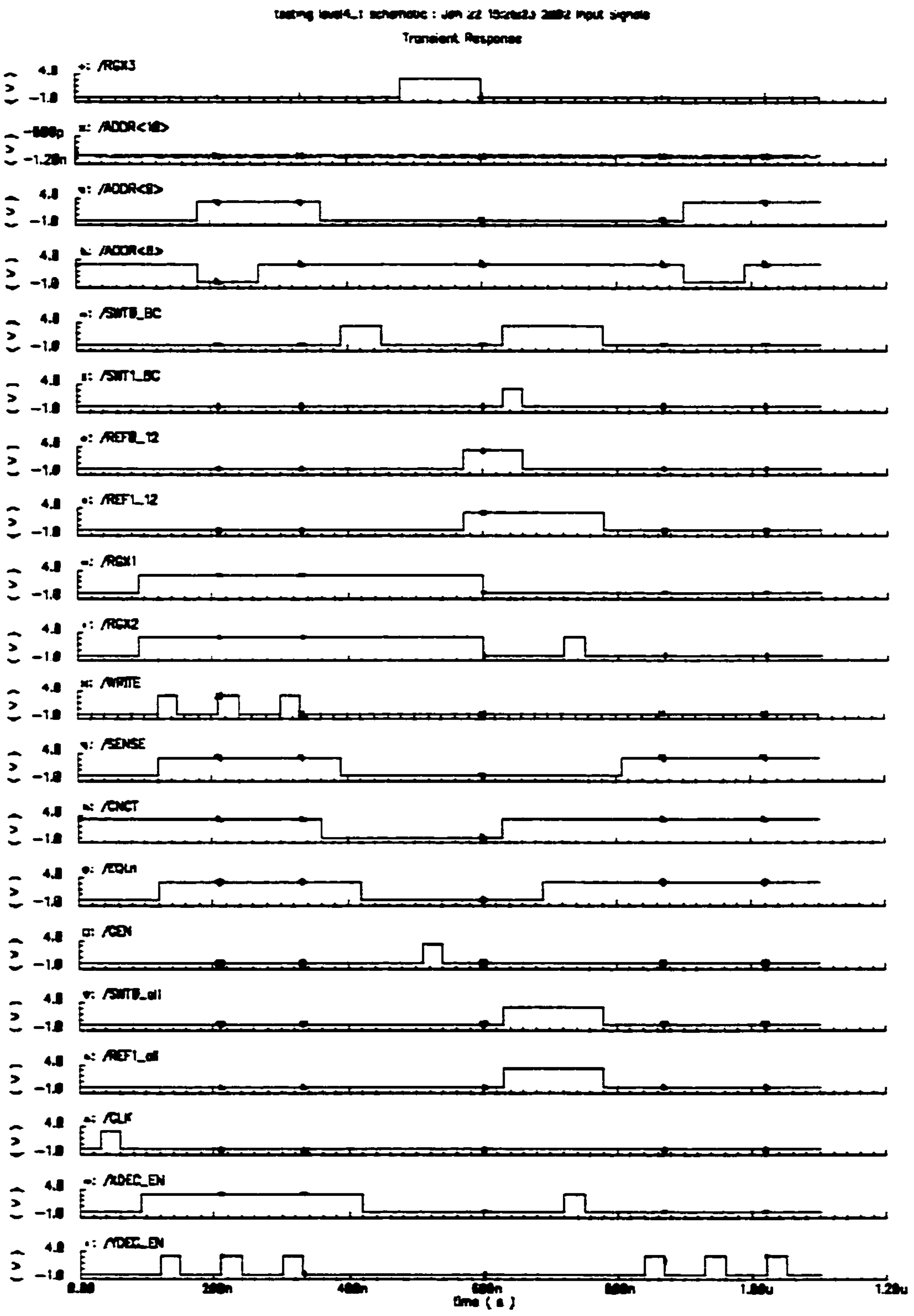


Figure 4-9. Stimulus waveforms for 4-levels-per-cell operation

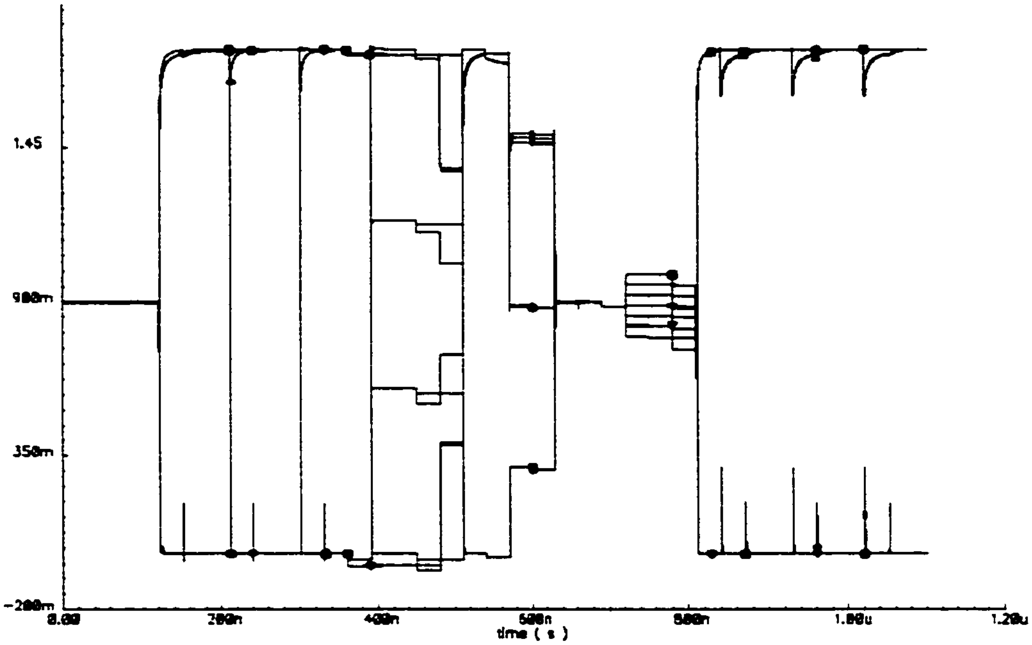


Figure 4-10. Bitline and reference signals in 4-levels-per-cell operation

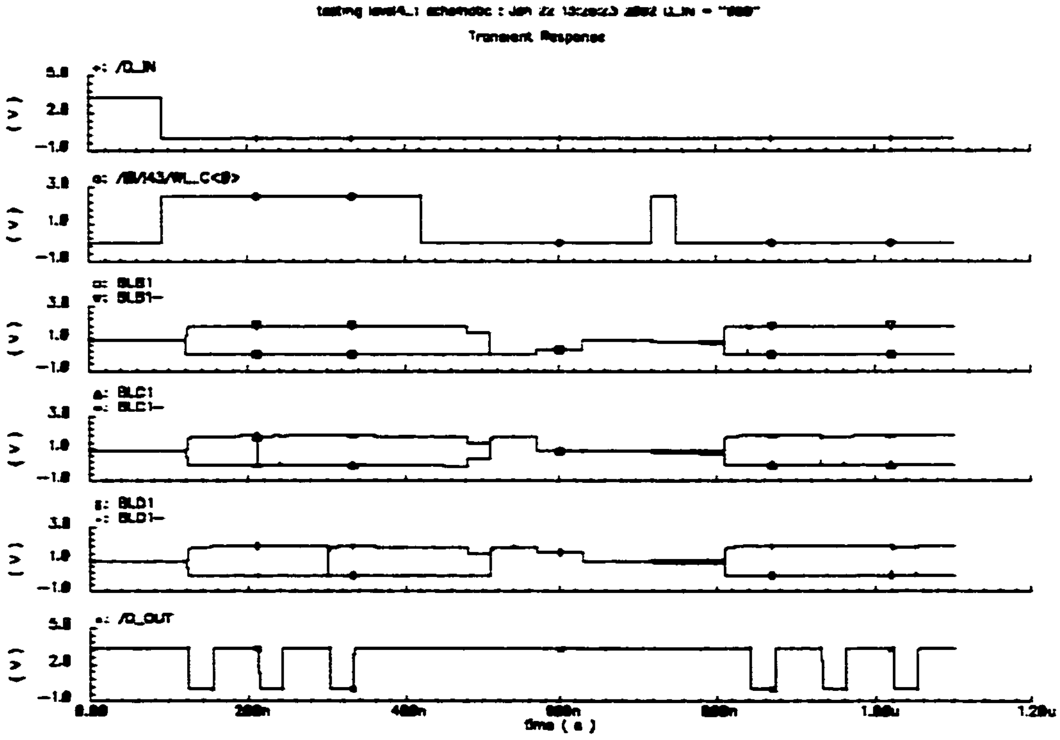


Figure 4-11. Simulation waveforms for data = "000"

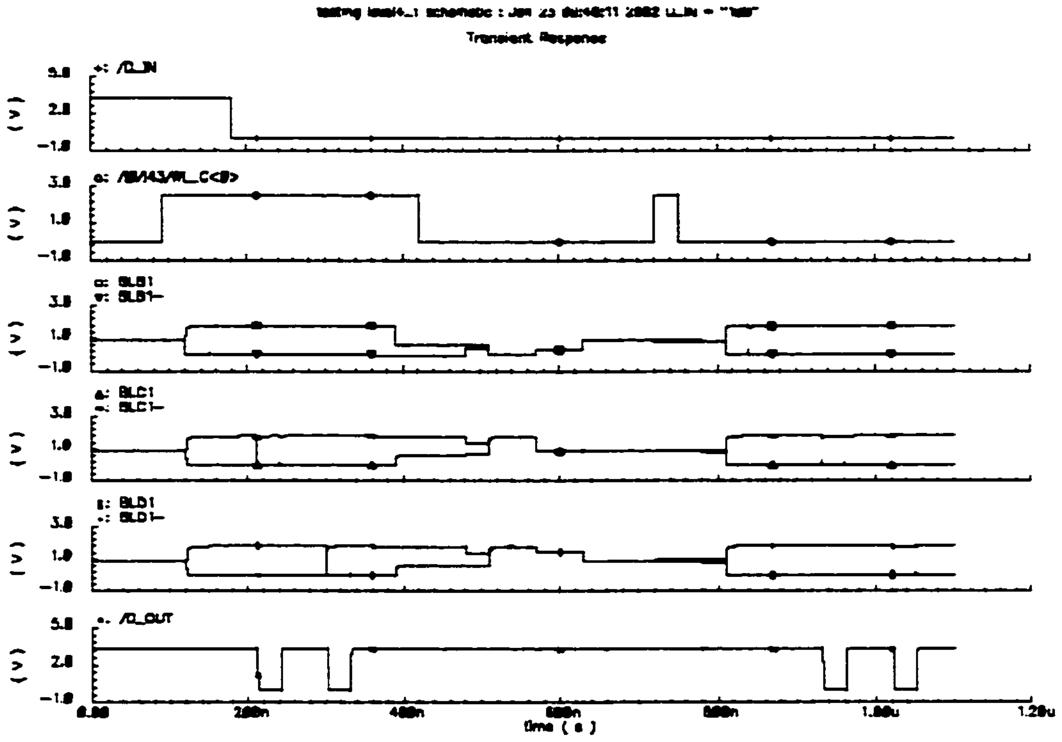


Figure 4-12. Simulation waveforms for data = "100"

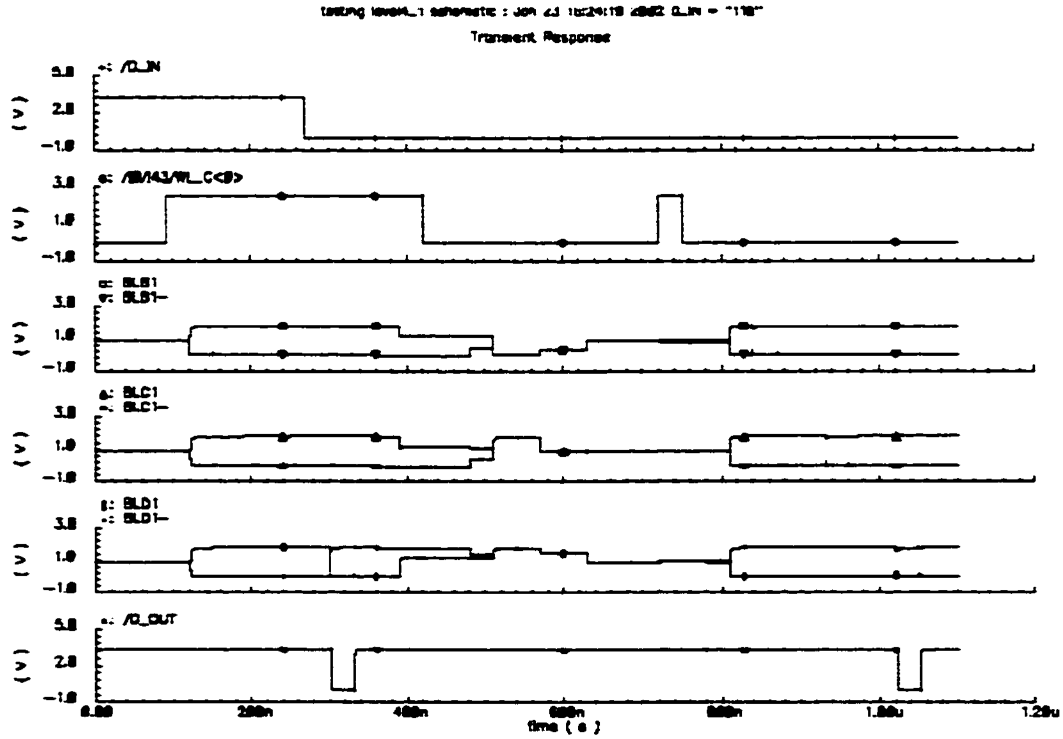


Figure 4-13. Simulation waveforms for data = "110"

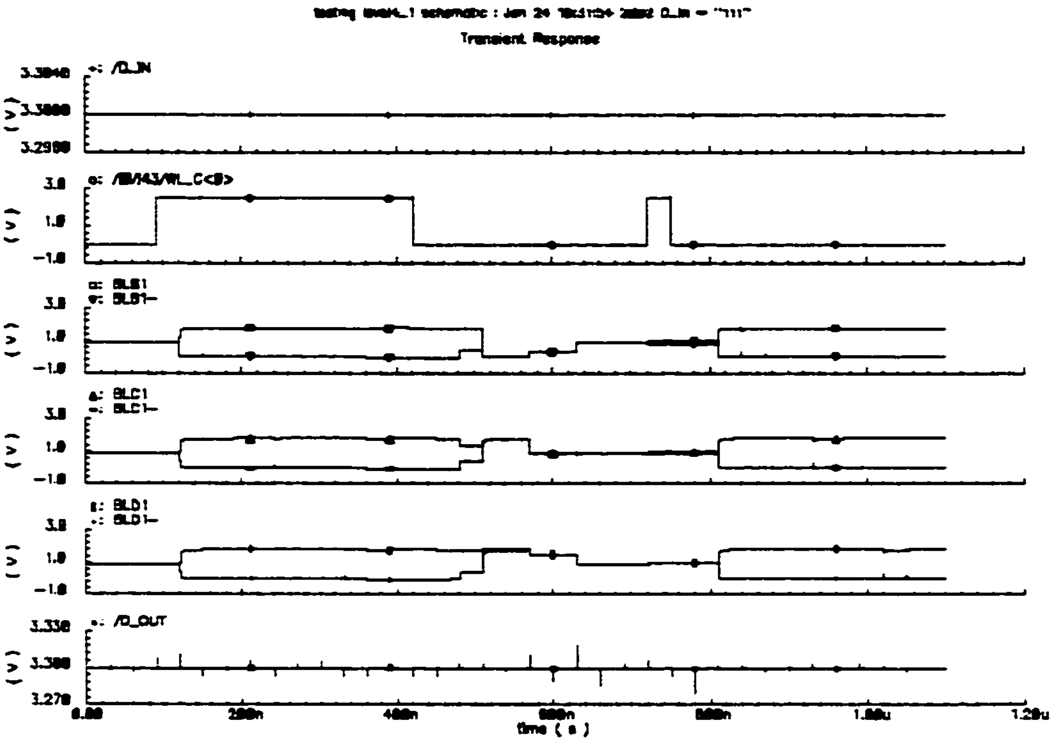


Figure 4-14. Simulation waveforms for data = "111"

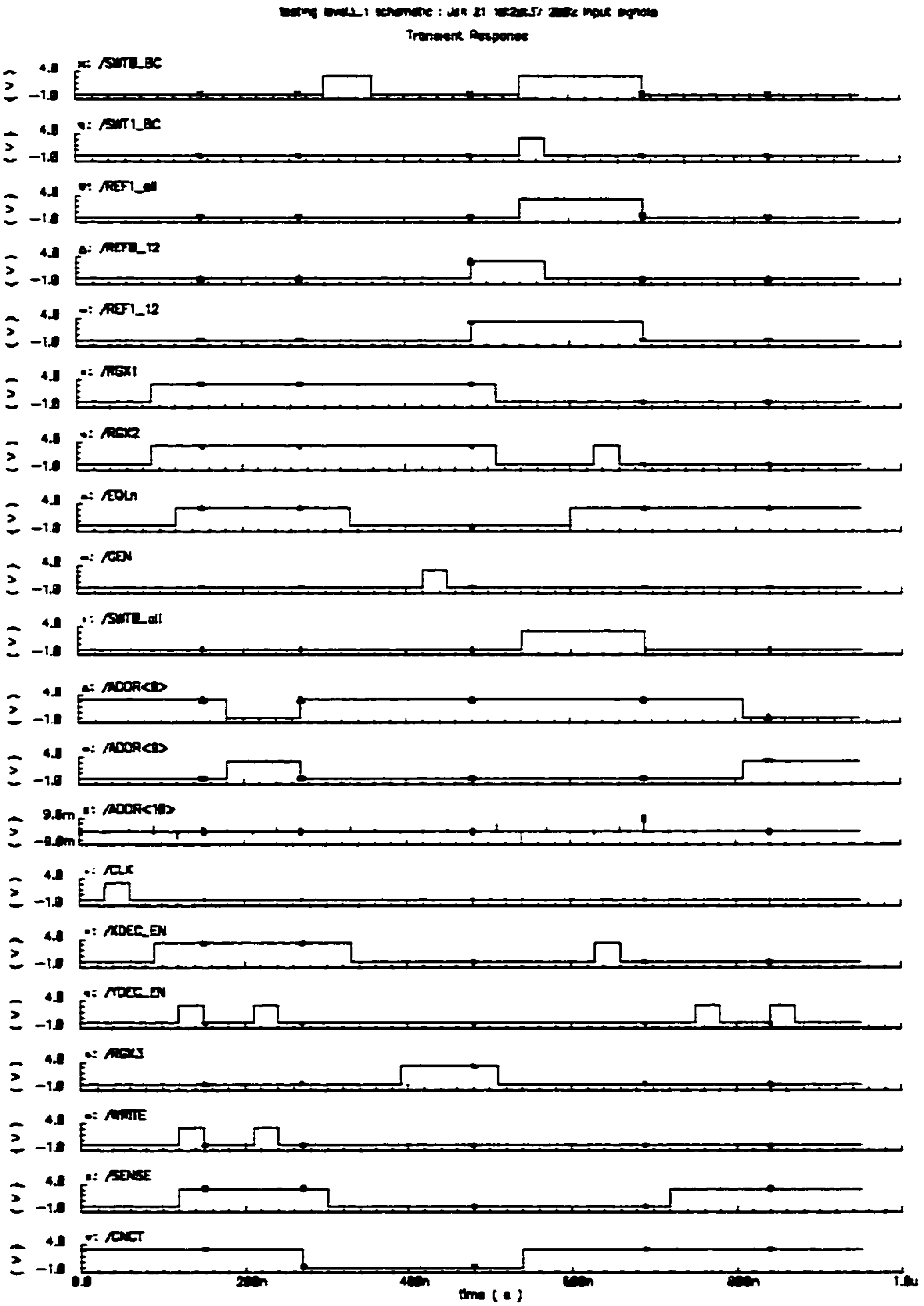


Figure 4-15. Stimulus waveforms for 3-levels-per-cell operation

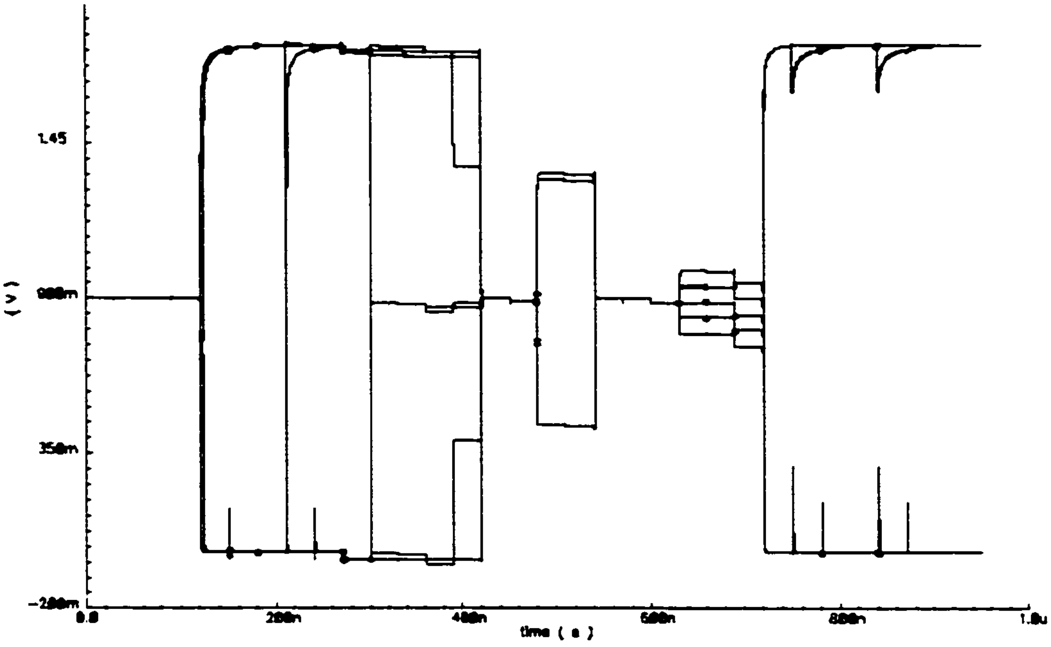


Figure 4-16. Bitline and reference signals in 3-levels-per-cell operation

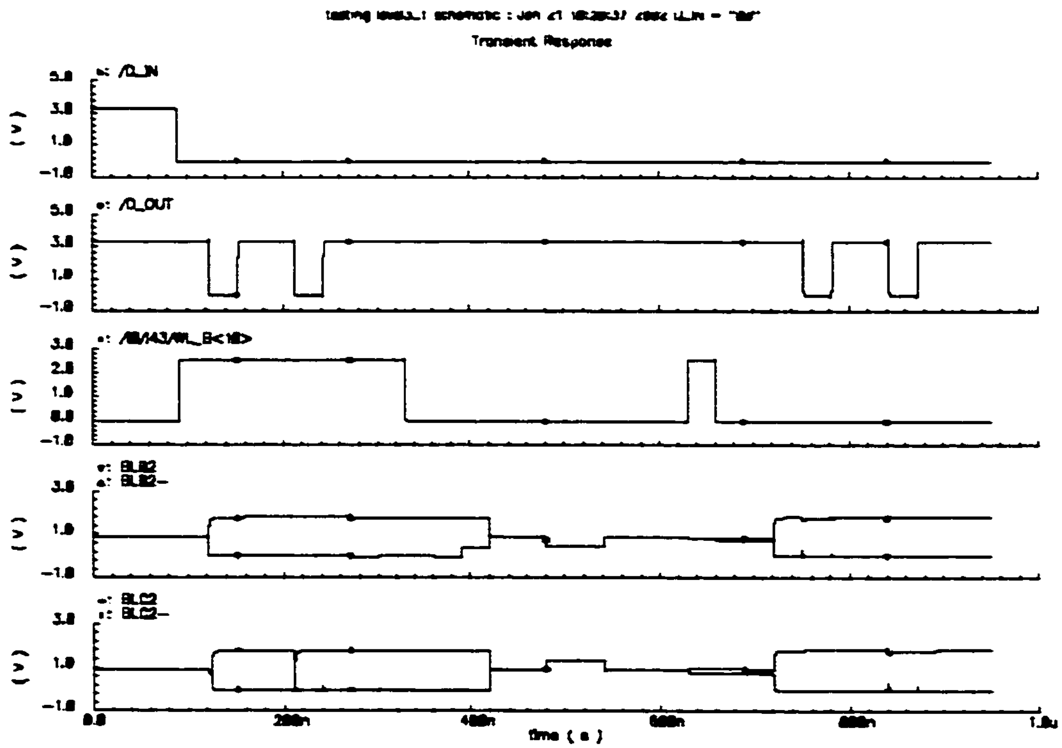


Figure 4-17. Simulation waveforms for data = "00"

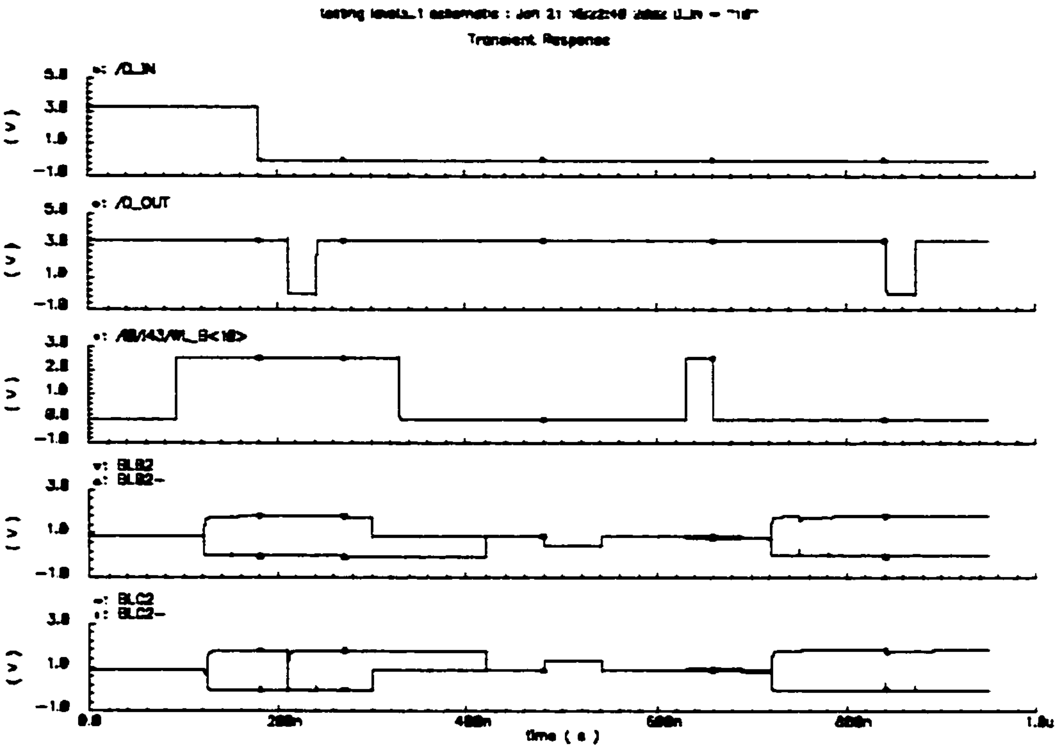


Figure 4-18. Simulation waveforms for data = "10"

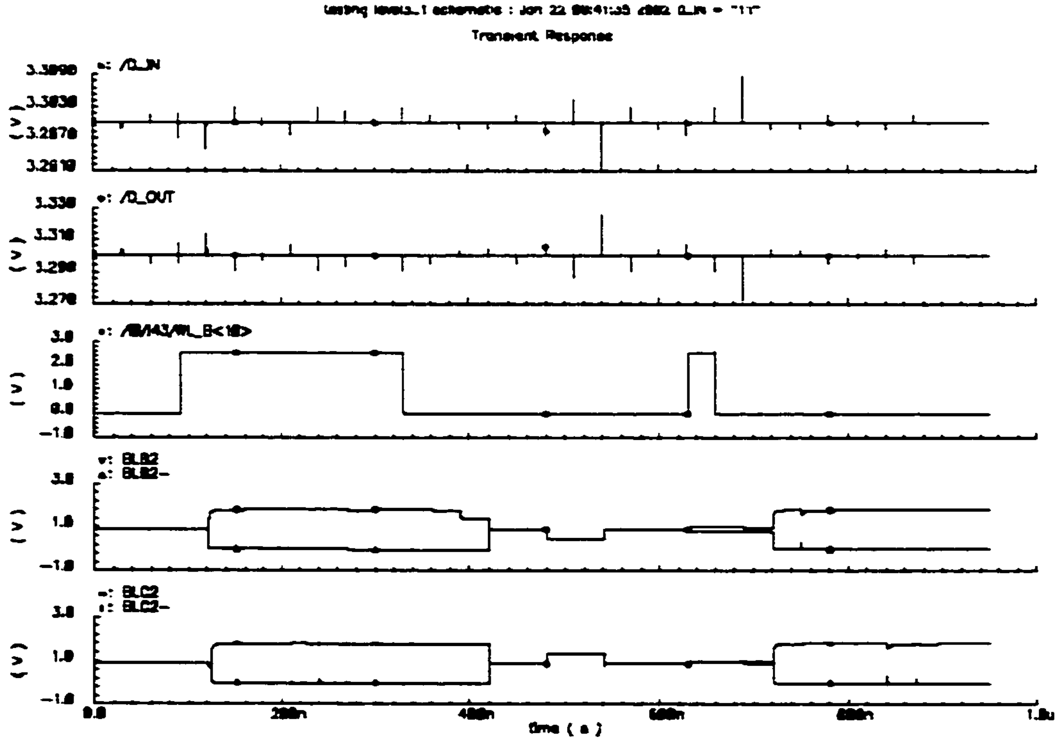


Figure 4-19. Simulation waveforms for data = "11"

4.2.2.3 Simulations for 2-levels-per-cell Operation

There are fewer control signals for the 2-levels-per-cell operating mode of the test chip. The control sequence for 2-levels-per-cell operation is similar to that of a conventional DRAM and is much simpler than those of other operation modes. Since there are only two cell voltage levels, and since the precharge voltage 1/2 V_{DD} is used as the reference voltage, the complicated multilevel reference generation process is avoided. It is not necessary to create any intermediate cell voltages other than V_{DD} and V_{SS}, either. Therefore, the reference and generate cells are never used and all RGX signals can be kept low. The SAs are kept always connected to the sub-bitlines by setting CNCT to high. GEN is set to low because the special reference generation circuits are not used. In order to obtain the same bitline to cell capacitance ratio as in the other modes, if an even number wordline is addressed, all SWT0 and REF1 signals will go high and connect five sub-bitlines to form full-length bitlines horizontally and vertically before the sensing operation; if an odd number wordline is addressed, all SWT1 and REF0 signals will go high.

Figure 4-20 gives the stimulus waveforms for the 2-levels-per-cell operating mode. Since WL<11> in section E is addressed, all SWT0 and REF1 signals are tied to V_{SS}; only SWT1 and REF0 are used as control signals. Figure 4-21 illustrates the bitline and reference signals waveforms and Figure 4-22 and Figure 4-23 show the simulation results. Note that both logic "0" and "1" data values can be written and read back correctly.

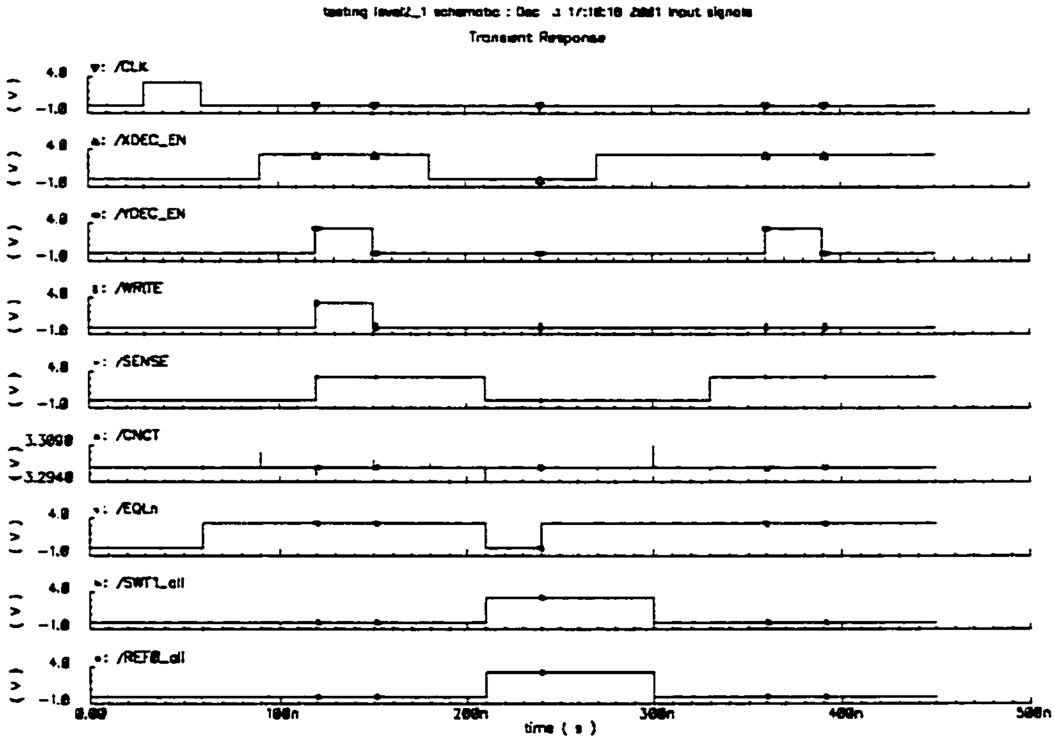


Figure 4-20. Stimulus waveforms for 2-levels-per-cell operation

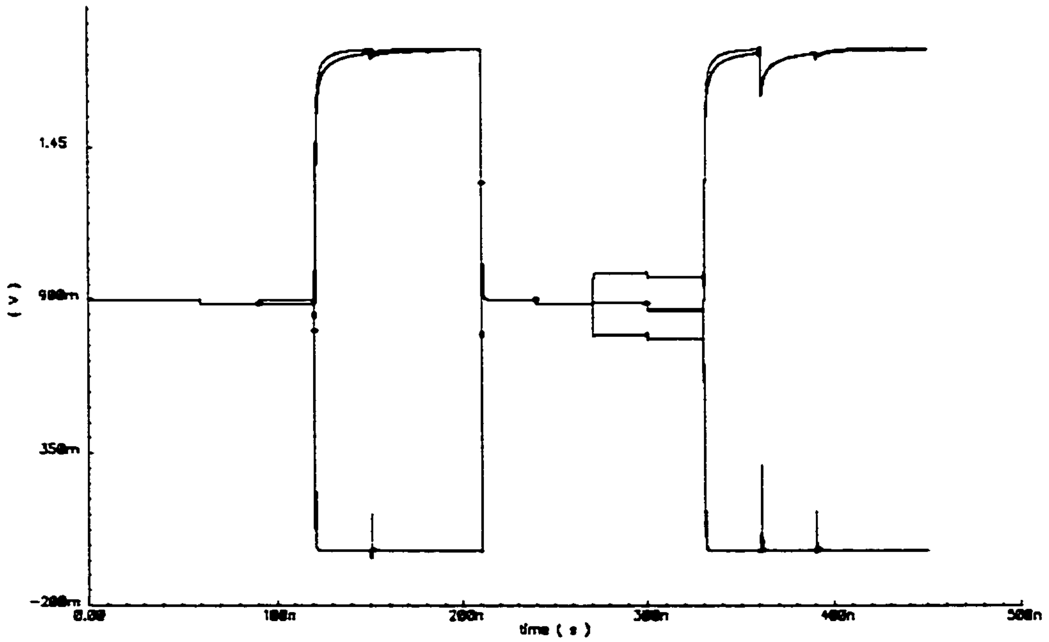


Figure 4-21. Bitline and reference signals in 2-levels-per-cell operation

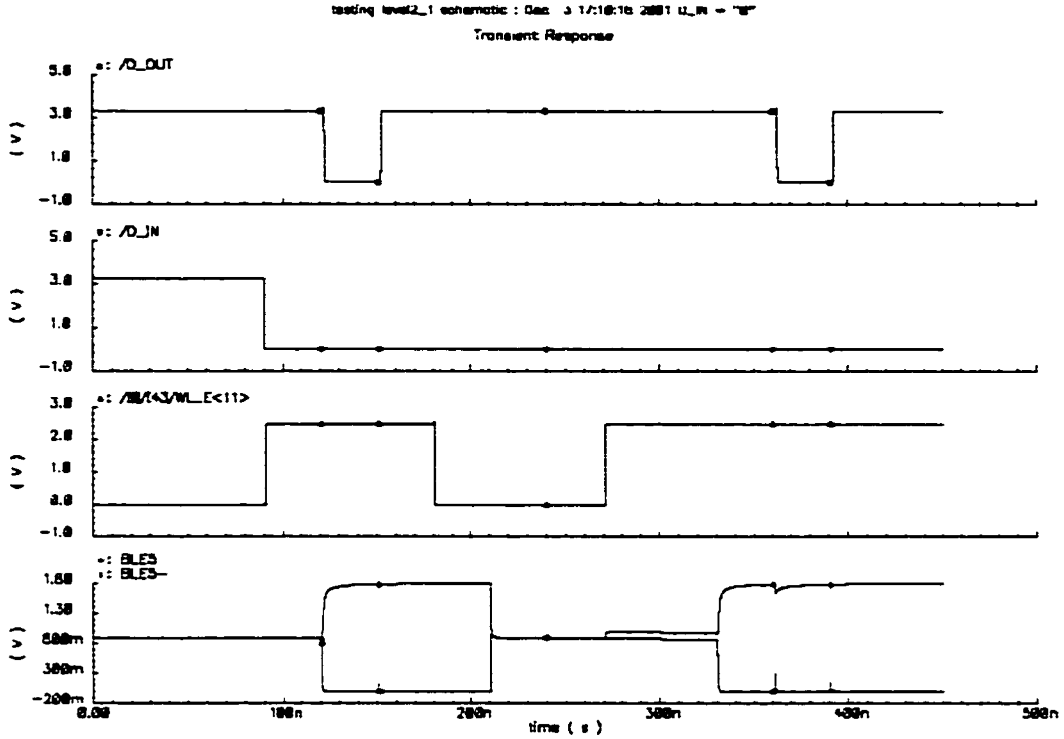


Figure 4-22. Simulation waveforms for data = "0"

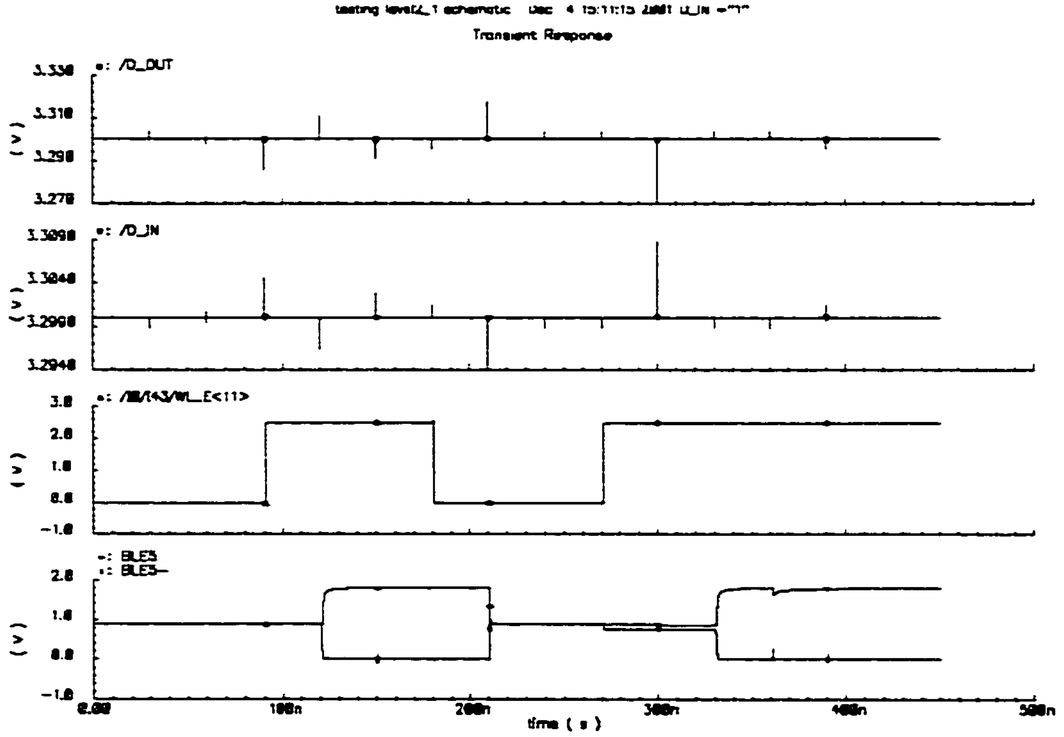


Figure 4-23. Simulation waveforms for data = "1"

4.2.3 Discussion

Recall that in section 4.1, we assumed that the full bitline to cell capacitance ratio is 7. Thus the critical attenuated differential voltage applied to the sense amplifiers before sensing is 180 mV/8 = 22.5 mV for 6-levels-per-cell operation, 300 mV/8 = 37.5 mV for 4-levels-per-cell operation, 450 mV/8 = 56.25 mV for 3-levels-per-cell operation, and 112.5 mV for 2-levels-per-cell operation. From Figure 4-2, Figure 4-10, Figure 4-16 and Figure 4-20, we can clearly observe that during the sensing operation, the attenuated voltage signals lie within almost the same range. However, the space between two signals is very close for 6-levels-per-cell operation. The spacing becomes larger for 4-level and 3-level operation, and it is much more relaxed for 2-levels-per-cell operation. The generated cell voltage and reference voltage levels measured from the simulation results are close to the ideal values. Table 4-1 gives the voltage signals for 4-levels-per-cell operation.

Table 4-1. Comparison of voltage signals for 4-levels-per-cell operation

Cell voltage (ideal) (V)	Cell voltage (simulated) (V)	Reference voltage (ideal) (V)	Reference voltage (simulated) (V)	Critical attenuated voltage (ideal)	Critical attenuated voltage (simulated)
1.8	1.8			37.5 mV	35.62 mV
		1.5	1.494 1.483 1.468		
1.2	1.1896				
		0.9	0.882 0.886		
0.6	0.5903				
		0.3	0.309		
0	0.0015				

The smallest attenuated differential voltage signal measured from the simulation results is 35.62 mV for 4-level operation, which is very close to the calculated value. For 6-level operation, the smallest differential signal from the simulation results is only 2.53 mV, which is much smaller than the calculated value. From inspection of enlarged simulation waveforms, we observed that de-asserting the horizontal switches and vertical switches somehow caused this mismatch. Meanwhile, as shown in Table 4-1, from the superimposed sub-bitline signal simulation waveforms, there are clearly small offsets that create multiple reference values, where there should be only one value. The undesired split between the two reference signals in true and complement sub-bitlines becomes larger from section A to E. For example, for 6-levels-per-cell operation, there is no difference for the generated reference signals at all in section A; but the difference grows to 291 mV in section E.

All of the simulation results verify that the outputs read from the memory cell are identical with the inputs written earlier to the same cell. Notice that in the examples of simulation waveforms, an odd-numbered wordline was addressed in the 2-levels-per-cell mode (WLE<11>). In all other operating modes, only even-numbered wordlines were addressed. This is because all of the even-numbered wordlines are connected with the memory cells on the true sub-bitlines; and the odd-numbered wordlines are connected with the cells on the complement sub-bitlines. When an even-numbered wordline is addressed, data are directly written to (read from) the true sub-bitlines and stored in the cells connected with true sub-bitlines. Thus, the output data is exactly the same as the input data. However, when an odd numbered wordline is addressed, data are still written/read to/from the true sub-bitlines, while the complemented data are stored in the cells connected with the complement sub-bitlines. For 2-levels-per-cell operation, since there are only 2 levels, data can still be read directly from the odd-numbered wordlines. However, for the multiple-levels-per-cell operations, since there are more than two levels, data read out directly from the true sub-bitlines are not identical with the input data. For example, in 6-levels-per-cell mode, assume that the input data are "11000", and an odd-numbered wordline is addressed. Voltages "V_{DD}", "V_{DD}", "V_{SS}", "V_{SS}" and "V_{SS}" are written to the true sub-bitlines in sections A to E, respectively. Thus "V_{SS}", "V_{SS}", "V_{SS}", "V_{DD}" and "V_{DD}" are written to the complement sub-bitlines in the five sections. Therefore, instead of storing a 2/5 V_{DD} in a cell connected with true sub-bitlines, a 3/5 V_{DD} signal is stored in the addressed cell. During the sensing operation, each copy of the attenuated 3/5 V_{DD} signal voltage in the complement sub-bitline is compared with the attenuated 1/10 V_{DD}, 3/10 V_{DD}, ..., 9/10 V_{DD} reference voltage on the true sub-bitline in sections A to E, respectively. That will result in "11100" in the complement sub-bitlines and "00011" in the true sub-bitlines. Thus, it seems that input data '11000" is written to the cell and "00011" are read out from cell.

Therefore, for multiple-levels-per-cell operation, when an odd-numbered wordline is addressed, the output data can be verified according to the conversion table: Table 4-2. Or, a simple step can be done in the design stage to avoid the data conversion: when X address bit 0 is 1 (for odd-number wordlines), first invert the input data that is to be written to the data bus, and thus send the data to the complement sub-bitlines instead of the true sub-bitlines.

Table 4-2. Output data conversion table

Operating Mode	Input data	Output data
6-level per cell	00000	00000
	10000	00001
	11000	00011
	11100	00111
	11110	01111
	11111	11111
4-level per cell	000	000
	100	001
	110	011
	111	111
3-level per cell	00	00
	10	01
	11	11

It should be pointed out that while a cell at one Y-address in one section is being accessed, all other cells with the same X-address in the same section as the addressed cell are also being accessed in parallel using the same reference signals, but on different folded sub-bitlines. If two independent data buses are provided into the SAs, then the two cell accesses required in 1.5 and 2.5-bits-per-cell mode can be performed at the same time in parallel. This should be quite feasible in embedded DRAMs, where there is a wide available page of data.

Chapter 5 Test Chip Evaluation

Ten packaged dies and 28 loose dies were fabricated by TSMC through CMC. Five of the packaged chips were verified using an Agilent 81200 digital tester in 6 and 2-levels-per-cell operation. The loose dies have been kept in reserve for future packaging and more detailed characterization.

5.1 The Tester and the DUT Board Interface

The Agilent 81200 is a modular platform consisting of VXI-standard front-ends, modules, mainframes, test fixtures and user interfaces. Our tester configuration, shown in Figure 5-1, included one E1401A high power VXI mainframe, which has 13 slots. One HP E8491B IEEE-1394 PC Link to VXI card was installed in slot 0, providing a communication link between the PC (which is also used as the controller for the testing system) and the tester mainframe. The E4805B central clock module was installed in slot 1; it generated the system clock that was routed to all of the data generator/analyzer modules installed in the other slots. Six E4841A 667 MHz data generator/analyzer modules are installed. The E4841A data module has four free daughter card positions for generator/analyzer front-ends and it has 1 Mbit of built-in memory to store the test vector data [22]. Thirteen installed data generators were the E4846A dual channel, single-ended front ends with 200 Mb/s maximum frequency and NRZ (*Non-return-to-zero*) or DNRZ (*Delayed-non-return-to-zero*) as alternative data formats. Five E4843A single-channel, 667 Mb/s differential generator front ends were also installed. Therefore, the tester could provide up to 13 × 2 + 5 = 31 stimulus signals for the DUT (device under test) input pins. The tester also had four E4844A single channel, 667 Msample/s analyzer front ends and two E4847A dual channel, 50 Ohm/high-impedance selectable, 333 Msample/s analyzer front ends, which can handle up to eight output signals from the DUT.

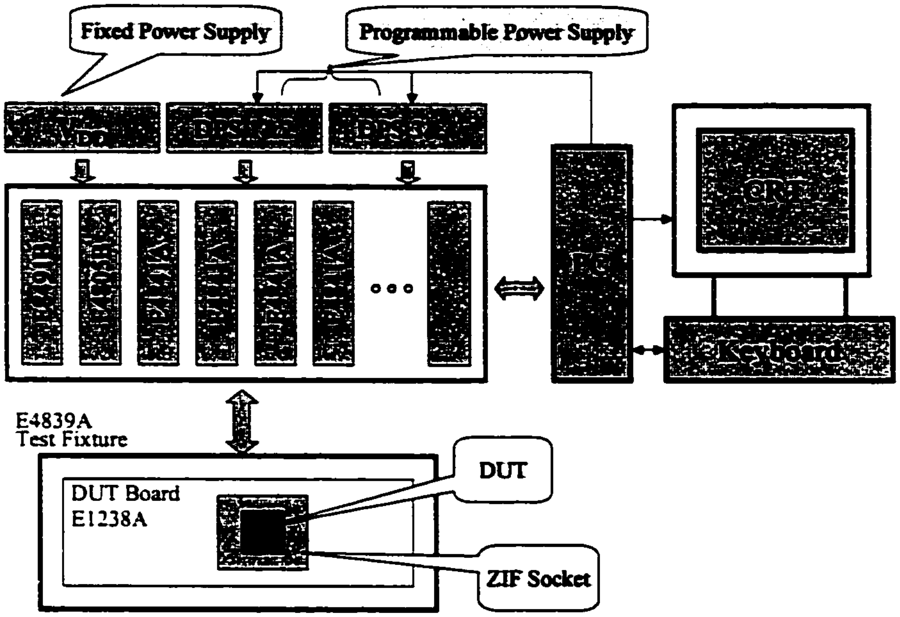


Figure 5-1. Agilent 81200 tester configuration

Individual generator channels can be assigned as either data or pulse ports. Pulse ports are independent of the data sequence and can be used as a clock source. The test vector memory can be divided into loopable segments, which can be one of the three types: pattern, pause, and PRBS/PRWS (Pseudo-random binary sequence/pseudo-random word sequence). One of four possible data pattern formats, NRZ, DNRZ, RZ (Return-to-zero) and R1 (Return-to-one), can be selected for each generator channel [22]. Since the CLK signal in ML5 is only used to latch the X addresses, it does not have to be a real clock signal, but can instead be a normal control signal just like all other input control signals for the chip. Therefore, data port and NRZ data format were selected for all the input pins during testing. With the NRZ format, signal transitions can only occur at the system clock rising edges. Thus each tester vector will specify a set of input signals that will be held steady for one system clock period.

The Agilent E4839A test fixture and the E1238A universal DUT board interfaced the test chip to the tester. Stimulus signals generated by the data generator front ends and output signals from the DUT under the special stimulus are connected between the tester front ends and spring-loaded pogo pins in the test fixture using shielded cables. The DUT board is flush latched to the test fixture so that the signals are connected to the DUT board firmly through the pogo-pins. The E4839A test fixture has 12 rows on both the left and right sides, with eight pins in each row. Thus it can connect up to 192 pins to/from the DUT. The E4839A test fixture supports operations up to 660 MHz. However, the E1238A universal DUT board can only handle up to 10 rows, with four pins in each row, of signals on both sides, i.e., 80 pins [23]. Signals are connected from the signal connection area to the header in the DUT connection area by traces on the DUT board. A 15 × 13 socket is wire-wrapped to the headers according to the pin list in Table 5-1. On top of the socket, a 13 × 13 PGA ZIF (Zero insert force) socket is used to hold the chip that is in a PGA68 package. Due to use of wire-wrapped connections and the limitations of the PGA68 package, the maximum operation speed for the test chips is 50 MHz [5] [23].

Up to four power supply voltages, which are called DPS1 to DPS4 (DPS denotes Device Power Supply) can be applied to the DUT board through the device power supply cable. Two power supply voltages P1 and P2 can be distributed in the pin grid of the DUT connection area. Since the ring and core power supplies are available along each side of the test chip, these two distributed power supply voltages were assigned to VDD_ring and VDD_core, which were connected to the DUT board directly from the external power supply source. The other four voltage sources for the test chip, VBB, VBLP, VCP and VPP, were connected to the four DSP power supplies. Two 100 nF decoupling capacitors were inserted between the VDD_core, VDD_ring power supply voltage and ground; and four 10 μF decoupling capacitors were connected between the four DSP power supplies and ground. The power supply connections are also shown in Table 5-1. Two Agilent E3647A dual output, programmable DC power supplies were used to provide power supplies for VBB, VBLP, VCP and VPP. Finally, one non-programmable BK PRECISION 1651 triple output DC power supply was used to provide power supply for VDD_ring and VDD_core.

Table 5-1. ML5 pin list

Pin #	Signal Name	Pogo Pin	C-M-C	Pin #	Signal Name	Pogo Pin	C-M-C
1	NC			35	Ref0_12	R4-14	C1M5C1
2	Vblp	DPS1		36	Ref0_23	R4-14	C1M5C1
3	Addr_0_	R1-01	C1M7C1	37	Ref0_all	R4-14	C1M5C1
4	Addr_1_	R1-03	C1M7C2	38	Ref1_12	R5-10	C1M5C2
5	Addr_2_	R1-05	C1M7C3	39	Ref1_23	R5-10	C1M5C2
6	Addr_3_	R1-07	C1M7C4	40	Ref1_all	R5-10	C1M5C2
7	Addr_4_	R2-01	C1M7C5	41	Write	R5-16	C1M5C3
8	Addr_5_	R2-03	C1M7C6	42	vdd_core.2	P1	
9	vdd_core	P1		43	vss_core.2	GND	
10	vss_core	GND		44	Sense	R6-01	C1M5C5
11	Addr_6_	R2-05	C1M7C7	45	xdec_en	R6-03	C1M6C1
12	Addr_7_	R1-10	C1M3C5	46	ydec_en	R6-05	C1M6C2
13	Addr_8_	R1-12	C1M3C6	47	Rgx1	R6-07	C1M6C3
14	Addr_9_	R1-14	C1M3C7	48	Rgx2	R5-01	C1M6C4
15	Addr_10_	R1-16	C1M3C8	49	vdd_ring.2	P2	
16	vdd_ring	P2		50	vss_ring.2	GND	
17	vss_ring	GND		51	NC		
18	NC			52	NC		
19	NC			53	NC		
20	swt0_all	R3-10	C1M4C1	54	NC		
21	swt0_bc	R3-10	C1M4C1	55	Gen	R5-03	C1M6C8
22	swt0_cd	R3-10	C1M4C1	56	Clk	R5-05	C1M6C5
23	swt1_all	R3-14	C1M4C2	57	Cnct	R5-07	C1M6C6
24	swt1_bc	R3-14	C1M4C2	58	Vbb	DPS4	
25	vdd_core.3	P1		59	Eqln	R4-01	C1M6C7
26	vss_core.3	GND		60	vdd_core.1	P1	
27	swt1_cd	R3-14	C1M4C2	61	vss_core.1	GND	
28	Din	R4-10	C1M4C3	62	Vcp	DPS3	
29	Rgx3	R4-12	C1M4C4	63	Vpp	DPS2	
30	vdd_ring.3	P2		64	d_out	R4-05	C1M3C1
31	vss_ring.3	GND		65	vdd_ring.1	P2	
32	NC			66	vss_ring.1	GND	
33	NC			67	NC		
34	NC			68	NC		

Note: Pin # denotes the pin number in the PGA68 package bonding diagram provided CMC.
Signal Name denotes the name of the input/output pins of the test chip, where "NC" means "not connected".
Pogo Pin denotes the pogo pin position in the DUT board. For example, "R3-10" denotes "row 3, position number10".
C-M-C denotes clock group number, module number and channel number.

Agilent's I/O libraries for instrument control and the Agilent 81200 user software was installed in a Windows NT 4.0 PC to provide the software environment for the tester.

5.2 Test Results

After the test chip had been sent for fabrication, I discovered an error in the row-section-select decoder (x_enable_dec) schematic which was used for creating the netlist of the final periphery layout. The xdec_en signal for section C should have been decoded as X address bits A6*A5 A4*. However, by mistake, it was decoded as A6*A5A4 instead, which is the same combination for decoding section D. Therefore, row addresses for section C can never select a wordline. When a wordline in section D is addressed, another wordline in section C with the same four bits LSB row address will be selected as well. Due to this mistake, cell accesses in section C and D had to be avoided. Fortunately, this design error does not prevent accesses to sections A, B and E.

The first chip tested was damaged with a short between Vdd_core and ground even before a function test had been applied. All of the hardware connections had been checked without any errors. The cause for the damage was never determined. However, it was found upon reflection that we did not apply the various power supply voltages in a proper order. A C program was written by Dan Leder to control the order of applying power supplies from the PC and so avoid power-up sequencing errors in future test runs.

The second chip was tested by applying the power supplies in the correct order. The memory array pwell substrate voltage V_{BB} was applied at first to avoid latch up. Then Vdd_ring and Vss_core were applied. Finally the rest of the power supplies were applied at the same time. As the first step, a basic functional test checking whether the data buses are working was applied. With input data toggling between logic one and logic zero, each of the 32 data buses was connected to the D_IN and D_OUT pin. Through the waveform viewer, we could observe that the input data were all set up correctly. But we could not observe the output data at the beginning. It seemed that D_OUT always got stuck-at-zero, even when we increased the sampling rate of the output data. By using the oscilloscope to directly observe the output pin signal at the DUT board header, a signal amplitude of 2.2 V for logic one was noted instead of 3.3 V. We concluded that the output pad in ML5 does not have enough drivability to drive the 667 Msample/s, single-ended (with 50 Ohm internal impedance) E4844A analyzer. So we changed the output pin analyzer front end from the E4844A to the E4847A and selected the high impedance option as the analyzer termination impedance. With these changes, we observed the output signal with near the expected 3.3 V amplitude.

5.2.1 Test Vector Generation

Test vectors were generated using C programs. The ASCII text files created from the C programs consisted of the tester vectors that could be imported into the tester in pieces called data segments. A *data segment* contains the data to be generated or expected for a certain data port, as shown Figure 5-2. The segment width is the number of input or output pins in a data port; the segment length is the number of vectors in a segment.

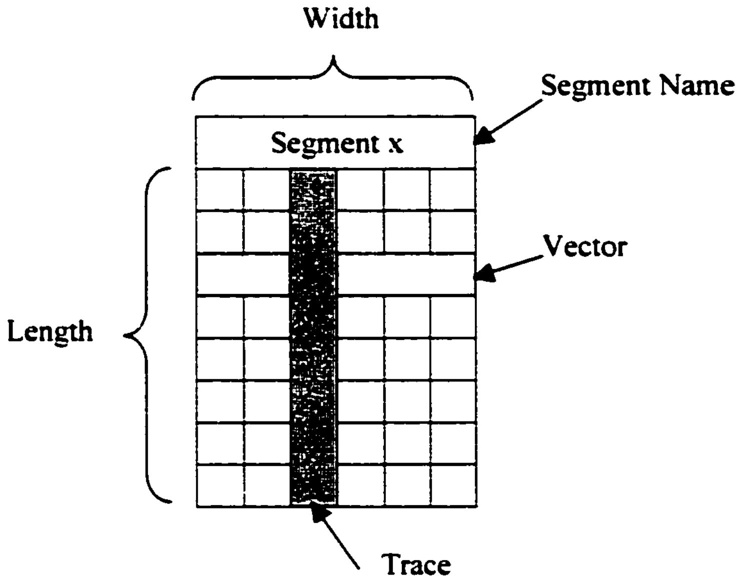


Figure 5-2. Data segment structure [22]

Real data segments contain either free programmable memorized data (a certain pattern), or the specification of PRBS or PRWS data (i.e., algorithmically generated data). *Pseudo data segments* are commands, such as Pause, Don't care, Expect 0, Acquire, and so on, which consume one word and thus can save memory [22]. The overall stream of generated and expected data is called the *data sequence*, which is based on sequence blocks. A block references segments and actions on events (a *event* is a signal condition that triggers a branch). It comprises all data ports. Single blocks and groups of blocks can be repeated using loops. A trigger pulse can be issued at the beginning of a block [22] as a condition for the sequence to go to a certain block. Figure 5-3 is a diagram of data sequence, blocks and segments.

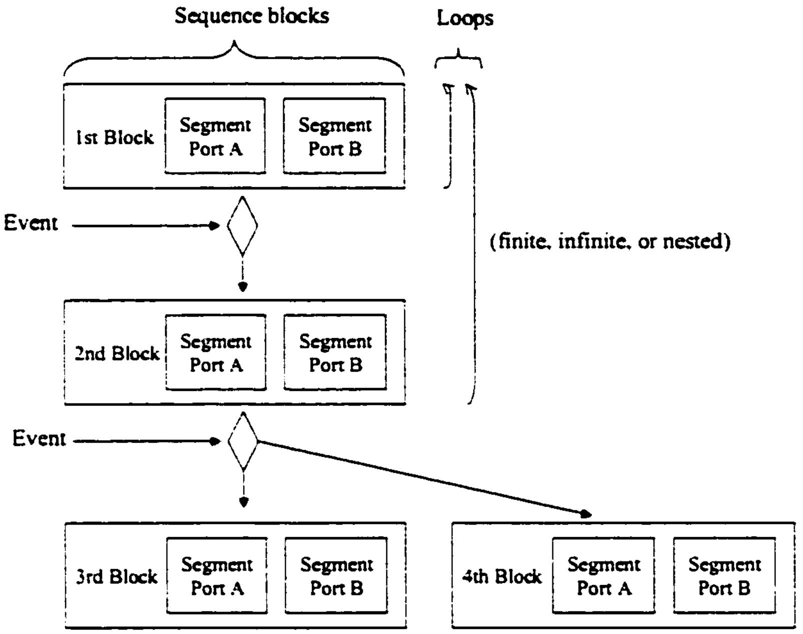


Figure 5-3. Data sequence, blocks and segments

No external trigger pulses (i.e., events) were used in the test sequences during testing. Test patterns were generated for various tests using different C programs. Pause 0 (pseudo segment for generator front ends, consisting only logic 0s for the specified block length) and Pause (pseudo segment for analyzer front ends, making the analyzer channels fall asleep for the specified block length.) were used in the retention time tests. It was found that the E4841A data module has a storage capacity (memory) of 64 Kwords, or 1 Mbits [22]. Therefore, all segments created can only contain up to 64 K test vectors and up to four sequence blocks that reference such data segments can be loaded into the memory at one time. Since the segment loading time is much greater than the retention time of the memory cells in the test chips, any written data must be read using the same segment load. This criterion has to be satisfied when creating the test patterns. In order to avoid the hassle of manually importing and executing many separate data segments into/from the tester, Dan Leder wrote a master C program to automate the whole testing process from sequencing the power supplies to loading the test vectors, capturing and parsing the result, and exporting an easy-to-read ASCII results file.

5.2.2 A Basic Functional Test

There have been few publications on MLDRAM testing. Redeker et al. proposed a fault model for Gillingham's MLDRAM scheme in [24]. A fault model for Birk's MLDRAM is under development now by other researchers at the University of Alberta. Therefore, we started with an ad hoc basic functional test for each cell in the 6-levels-per-cell operating mode, as shown in Figure 5-4. A second write of a different value to the cell in the same column follows the first write to the target cell before reading back the target cell to ensure that the data read is from the cell instead of a remnant bitline voltage. Since the read operation immediately follows the write operation, retention time problems affecting the cell at address (xaddr, yaddr) should have no effect.

```
for (yaddr = 0; yaddr < 250; yaddr++)
   for (xaddr = 0; xaddr < 12; xaddr++)
      for (phys_lev = 0; phys_lev < 6; phys_lev++)
      {
         write (phys_lev; xaddr; yaddr);
         write ((phys_lev +1)%6; xaddr+1; yaddr);
         read (phys_lev; xaddr, yaddr);
      }
```

Figure 5-4. Basic functional test for 6-levels-per-cell operation

Both the forward and reverse row and column address orders were used to run the basic functional test on three operational chips: chip number 2, 3 and 5. Similar functional test were run on these three chips for 2-levels-per-cell operation using only forward row and column address orders. The output data were collected in a database, from where data can be easily sorted and analyzed. Bitmaps were created based on the data from the database to facilitate the faulty cell localization and fault analysis. Figure 5-5 is a gray-scale bitmap created from the test results for section A in chip 2.

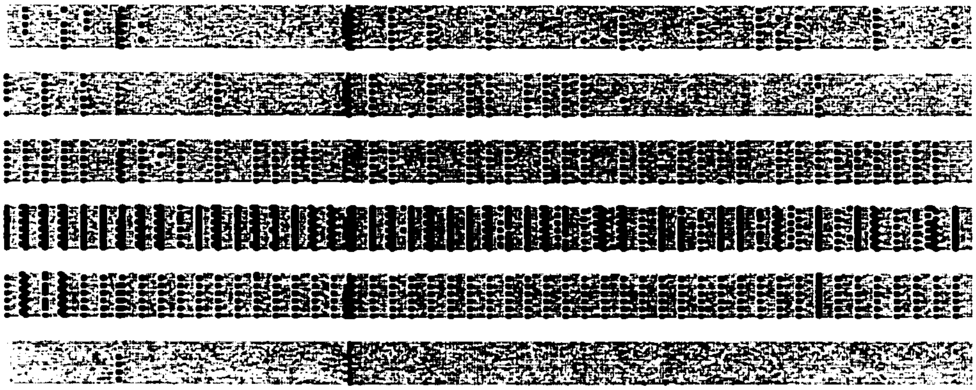


Figure 5-5. A gray-scale bitmap for section A in chip 2

Every pixel in the bitmap physically maps to a memory cell in the test chip. Each of the six strips in the bitmap represents the test result for writing one of the 6 cell voltage levels to all cells in section A of the chip. From top to bottom, cell voltage levels 5, 4, ..., 0 were written to the memory cells and then read out. Grey-coloured pixels represent good cells, i.e., the read data are identical to the written data. Black-coloured pixels represent faulty cells, where read data do not match written data. As shown in Figure 5-5, most of the cells in the chip were good cells. The faulty cells were distributed in a very regular patten. Detailed data giving the read values for the six written cell voltage levels on ten different basic types of sub-bitlines for the three functional sections of chip 2 are given in Appendix A. Bitmaps for chip 2, 3 and 5 are given in Appendix B.

By further analysis of the data from Appendix A and the patterns shown in the bitmaps of the three functional chips in Appendix B, it was found that the first column of every group of five columns was the most vulnerable column. The cell voltage read out from the first column drifted down one level for almost all faulty cells. Also shown in the bitmaps is that level 1 and 2 were the most critical cell voltages. As the cell voltage level became bigger, there were less faulty cells. The exception of level 0 could be because it is the smallest cell voltage level. Therefore the downward drift of the cell voltage could not be detected. It seems like the cells in every first column of a basic memory array block are extremely leaky; alternatively, the reference voltage levels could possibly be higher in these columns. Since similar patterns were found in all three functional chips and faulty columns appeared in every basic array block, it must be a design error in either the layout or the timing. However, the real cause for the faults is undetermined yet. Whether the cell or reference voltage offset is due to the known unbalanced sub-bitline charge injection in Birk's scheme is not yet determined. More simulations and circuit investigations must be done to diagnose the problem. Note that there were some irregular clustered faulty cells in section E chip 2 and section B chip 5. These faulty cells could be caused by manufacturing defects in a certain chip or soft errors.

Table 5-2 shows the percentage of good cells calculated from the results of the functional tests. The percentages of good cells for the 2-levels-per-cell operation of all three chips are 100%. The percentages of good cells for the 6-levels-per-cell operation range from 77.02% to 80.33%. We assume the reduced good cell percentage occurs from the degraded noise margins in the 6-levels-per-cell operation.

Table 5-2. Percentage of good cells

Chip Number	6-levels-per-cell operation		2-levels-per-cell operation
	Forward address order	Reverse address order	Forward address order
2	77.02%	77.59%	100%
3	79.82%	80.23%	100%
5	80.14%	80.33%	100%

5.2.3 A Cell Retention Time Test

Using the basic functional test described above, we were able to verify that many cells were working properly. A cell retention time test was then derived and applied to the test chips. Figure 5-6 illustrates the possible leakage current sources for a cell capacitor in the test chip. There are four possible leakage currents: I1 is the junction leakage currents from the storage node to the pwell memory array substrate, which is biased at V_{BB} to decrease this leakage; I2 is the sub-threshold leakage current from the storage node to the BL through the access transistor; I3 is leakage current between adjacent cell storage nodes; and I4 is the leakage current from the storage node to the cell plate due to tunneling effects. The final state of the data should be the result of the sum of these leakage currents. Depending on the strength of the leakages, the cell signal voltage stored in the cell capacitors will finally go to either V_{BB} through the junction leakage currents or to whatever the bitline voltage is.

By writing both the lowest and highest voltage levels to all functional cells and by observing the cell populations in the different levels over time, we tried to measure the retention time and to find the final cell voltage level so that we could conclude the dominant leakage source for the test chip. From test results we got, we found that most of the cells can hold their values correctly up to only about 150 μs. Compared with the 2 ms refresh rate for ATMOS' 0.18-μm SoC-RAM, this average retention time is too small. However, the final cell voltage levels of the cells are random data. One possibility is, like commercial DRAM, the cell access transistor leakage current I2 is the dominant leakage current. The undetermined voltage levels in the BLs (BL is precharged to 1/2 V_{DD}) finally created the random data for the cells.

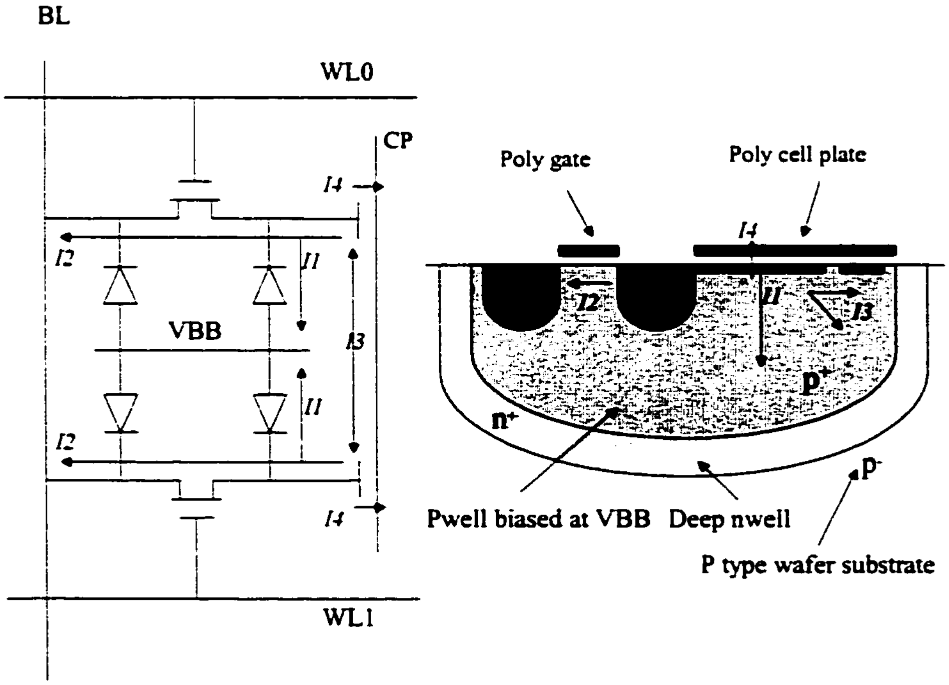


Figure 5-6. Diagram of cell leakage currents

5.3 Conclusions

The test evaluation proves that the test chip works since about 80% of the cells functioned properly in the basic functional test for the 6-levels-per-cell operation. By using the same length of bitlines, 100% cells were functional for all of the three functional chips in 2-levels-per-cell operation. In 6-levels-per-cell operation, most of the errors occurred for cell voltage levels 1 and 2. The errors are almost all located in the first column of a group of five columns in a basic array block. Reduced noise margins are definitely the source of the errors. However, further testing and investigation of the circuits are needed to explain the cause of degraded signal levels on the critical columns. Most of the cells can hold their value up to 150 μs but then the cell voltages appeared to drift to a random value. Further testing is needed to evaluate the 3 and 4-levels-per-cell operating modes and to find out the cause of the final random data in the cells. A more detailed characterization of the test chips is to be carried out by other students.

Chapter 6 Conclusions

The reduced noise margins in MLDRAM pose probably the greatest challenge preventing commercial acceptance of the technology. In this thesis project, an MLDRAM test chip with adjustable cell capacities using Birk et al.'s scheme was designed and implemented in TSMC's 0.18-μm technology. Test chips were fabricated through CMC and most of the cells were shown to work properly in silicon in the 2- and 6-levels-per-cell operating modes.

6.1 Chip Design and Simulation

The architecture of the test chip was designed based on Birk et al.'s MLDRAM scheme. The detailed schematic and layout design was based on elements from the previous chip designs ML2 and ML3, as well as ATMOS' SoC-RAM^{TM}. Having more control signals than ML3, the ML5 test chip can be operated in 2.5, 2, 1.5 or 1 bit(s)-per-cell operating modes. The memory core was a full custom layout with 15,000 memory cells; the periphery of the chip was placed and routed using standard cells provided by CMC. Due to the reduced noise margins and the lack of accurate values for such key parameters as the gate-substrate capacitance of NMOS transistors, the parasitic capacitance of bitlines, etc., the test chip was designed with a safe bitline to cell capacitance ratio by using large-sized cell capacitors and short sub-bitlines. Despite many precautions, there were still some errors and bugs in the design. As mentioned earlier, there is a decoding error in the row_enable_decoder that has the effect of disabling sections C and D in the test chips. Also, the column scrambling and the mirrored cell voltage levels for cells connected with even and odd-numbered wordlines was unnecessarily complex and this made the testing of the test chips overly complicated.

Simulation of the test chip was based on the schematics of the memory core because the cell capacitor layout could not be fully extracted by the available tools. Thus a cell capacitance and a bitline parasitic capacitance were added into the schematic of each memory cell using the calculated and estimated capacitance values, respectively. Every signal transition was simulated using a relaxed 30 ns inter-transition time. With the simulation based on the ideal schematics, it was found that even 1 ns inter-transition time between critical signals still allowed the chip to work properly in simulation. By using 1 ns non-overlapping inter-transition time for most irrelative signals and 10 ns WL asserting time, the read access time for 6-levels-per-cell operation was 15 ns and the cycle time was 30 ns.

6.2 Chip Evaluation

Five out of ten available packaged chips were tested using an Agilent 81200 tester. A control program, designed by Dan Leder, was used to control the tester and the programmable power supplies. Test vectors were generated from a number of C programs. Column de-scrambling and cell voltage level de-scrambling were integrated in these programs to ensure the proper physical positions of the cells and the analog cell voltages.

Three of the five tested chips were found to have functional cells. An average of 79.19% of cells on the whole chip worked fine for a basic functional test in the 6-levels-per-cell operating mode. 100% cells were found to be good in the 2-levels-per-cell mode. In the 6-levels-per-cell mode, it was found that most of the errors occurred on the first column of every group of five columns, and cell voltage levels 1 and 2 had the poorest noise margins. Further tests are needed to find out the dominant cell leakage current and the final cell voltage.

6.3 Future Work

With commercialization of MLDRAM as our final goal, the discovered problems in the current test chip and the remaining research work are summarized below.

The discovered problems in ML5 design and evaluation are listed as follows:

1. During the design process, since the memory cell capacitor could not be extracted from the layout, a full LVS check for the whole chip with the memory core extracted was not performed.
2. Test chip circuits were simulated mostly using the circuit parameters from schematics instead of circuit parameters extracted from the layout. Simulations were limited to a small memory core with only 10 columns

instead of the original 250 columns in the chip due to the simulator handling capability.

3. From the simulation results, it was found that there were voltage offsets in the generated reference voltages.
4. An address decoding error was found in the row address enable decoder.
5. Data written to the cells in true and complement sub-bitlines had to be scrambled during chip evaluation because a simple scrambling circuit was not considered in the design.
6. Column scrambling was much more complicated during chip evaluation due to the improper column address decoding.
7. A pattern of the faulty cells was found using the basic functional tests for 6-levels-per-cell operation.

Clearly, more detailed test chip characterization needs to be carried out next. The causes of the small offsets in the generated reference voltages seen in the simulation need to be determined. The distribution of the stored cell signals should be measured experimentally. This should be possible using a bump test, where the cell plate is used to inject controlled voltage offsets into the cells. A fault model for this MLDRAM scheme is under development now at the University of Alberta. Further characterization should be based on the fault model.

Although the initial test results of the test chip are promising, one important additional step to clarify the commercial value of this MLDRAM scheme is to design and implement a further test chip under more critical and realistic conditions. All of the found problems listed above should be avoided. Various sizes of cell capacitors and/or longer sub-bitlines should be included to explore the limits of the noise margins. Different types and different sizes of transistors should be included to implement the sense amplifiers in order to find the most suitable sense amplifiers which are robust to noise and still sensitive enough to sense the closely-spaced bitline signals. Bitlines should be tightly pitched to mimic the critical bitline coupling noise present in commercial DRAMs. Bitline twisting should be used to cancel out bitline-to-bitline coupling noise, like in a commercial part. The layout of the new chip design should be made as compact as possible to better evaluate the area overhead of the MLDRAM scheme.

References

[1] B. Prince, "Semiconductor Memories: A Handbook of Design Manufacturing & Applications," 2nd Ed., John Wiley & Sons, 1996.

[2] G. Birk, D. G. Elliott, B. F. Cockburn, U. S. patent application filed May 2000, "Improved Multilevel DRAM".

[3] A. Sheikholeslami and P. G. Gulak, "A survey of circuit innovations in Ferroelectric random-access memories," *Proceedings of the IEEE*, Vol. 88, No. 3, pp. 667-689, May 2000.

[4] B. Keeth and R.J. Baker, "DRAM Circuit Design: A Tutorial," IEEE Press, Piscataway, NJ, 2000.

[5] CMC webpage: http://www.cmc.ca.

[6] G. Birk, "Evaluation, Design, and Implementation of Multilevel DRAM," M.Sc. thesis, Dept. of Elect. and Comp. Eng., U. Alberta, August 1999.

[7] Website: http://www.eetimess.com. "Intel plans 4-bit-per-cell flash for mobile phones," by Anthony Cataldo, EE Times, March 5, 2001.

[8] M. Aoki et al., "A 16-Level/Cell Dynamic Memory," *IEEE J. Solid-State Circuits*, vol. 22, no. 2, pp. 297-299, April 1987. (Also in ISSCC Dig. Tech. Papers, pp. 246-247, 1985.)

[9] T. Furuyama et al., "An Experimental 2-bit/Cell Storage DRAM for Macrocell or Memory-on-Logic Application," *IEEE J. Solid-State Circuits*, vol. 24, no. 2, pp. 388-393, April 1989.

[10] P. Gillingham, "A Sense and Restore Technique for Multilevel DRAM," *IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 43, no. 7, pp. 483-486, July 1996.

[11] T. Okuda and T. Murotani, "A Four-Level Storage 4-Gb DRAM," *IEEE J. Solid-State Circuits*, vol. 32, no. 11, pp. 1743-1747, November 1997.

[12] D. A. Rich, "A Survey of Multivalued Memories," *IEEE Trans. Computers*, Vol. C-35, No. 2, pp. 99-106, February 1986.

[13] G. Birk, et al, "A Comparative Simulation Study of Four Multilevel DRAMs," *1999 IEEE Int. Workshop on Memory Tech., Design and Testing*, pp. 102-109, August 1999.

[14] A. Chan, "Design and Implementation of a Multilevel DRAM," M.Eng. report, Dept. of Elect. and Comp. Eng., U. Alberta, April 2000.

[15] D. Clein, "CMOS IC Layout, Concepts, Methodologies, and Tools," Newnus, Butterworth-Heinemann, U.S.A., 2000.

[16] B. Prince, "High Performance Memories, New Architecture DRAMs and SRAMs Evoluation and Function," John Wiley & Sons Ltd., England, 1996.

[17] S. S. Iyer and H.L. Kalter, "Embedded DRAM: Opportunities and Challenges," *IEEE Spectrum*, Vol. 36, No. 4, pp. 56-64, April 1999.

[18] C. S. Wang and E. CK Chen, "Embedded-DRAM technologies: comparisons and trade-offs," *EDN magazine*, pp. 113-120, September 28, 2000.

[19] ATMOS website: http://www.atmoscorp.com.

[20] EE 633 lecture notes, calculated by Dr. Filanovski based on the technology parameters provided by CMC documentation in 2000.

[21] K. S. Kundert, "The Designer's Guide to SPICE and SPECTRE," Kluwer Academic Publishers, Norwell, Massachusetts, 1995.

[22] Agilent 81200 Data Generator/Analyzer User Guide, revision 3.0, Agilent Technologies, 2000.

[23] Agilent E4839A Test Fixture, User's Guide, Agilent Technologies, July 2000.

[24] M. Redeker et al., "Fault Models and Test Strategies for a Two-bit per Cell DRAM," *IEEE Design & Test of Computers*, pp. 22-31, January - March 1999.

Appendix A Test Results for Chip 2

According to the physical positions, all of the sub-bitlines in an array are catalogued into one of the 10 sub-bitline types. The 250 sub-bitline pairs in an array are grouped into 50 basic blocks with each block consisting of 5 sub-bitline pairs. In each group, the true bitline of the first column is called Type 0 sub-bitline; the complement sub-bitline of the first column is called Type 1 sub-bitline. Similarly, the other 8 sub-bitlines are called Type 2 to Type 9 sub-bitline. Each section has 50 sub-bitlines for every sub-bitline type. The following table shows the test result of a basic functional test for the 10 different sub-bitlines in sections A, B and E of chip 2.

Type 0 sub-bitlines

Sec A	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	600	0	0	0	0	0	0	100.00
00001	18	582	0	0	0	0	0	97.00
00011	0	439	161	0	0	0	0	26.83
00111	0	0	2	598	0	0	0	99.67
01111	0	0	0	0	600	0	0	100.00
11111	0	0	0	0	0	600	0	100.00

Sec B	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	12	0	0	0	0	0	98.00
00001	70	530	0	0	0	0	0	88.33
00011	0	355	245	0	0	0	0	40.83
00111	0	12	16	572	0	0	0	95.33
01111	0	12	0	0	588	0	0	98.00
11111	0	12	0	0	0	588	0	98.00

Sec E	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	594	6	0	0	0	0	0	99.00
00001	50	550	0	0	0	0	0	91.67
00011	0	467	133	0	0	0	0	22.17
00111	0	6	1	593	0	0	0	98.83
01111	0	6	0	0	594	0	0	99.00
11111	0	6	0	0	0	594	0	99.00

Type 1 sub-bitlines

Sec A	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	600	0	0	0	0	0	0	100.00
00001	600	0	0	0	0	0	0	0.00

00011	0	587	13	0	0	0	0	2.17
00111	0	0	452	148	0	0	0	24.67
01111	0	0	0	156	444	0	0	74.00
11111	0	0	0	0	152	448	0	74.67

Sec B	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	595	5	0	0	0	0	0	99.17
00001	595	5	0	0	0	0	0	0.83
00011	7	576	17	0	0	0	0	2.83
00111	7	5	457	131	0	0	0	21.83
01111	7	5	0	163	425	0	0	70.83
11111	7	5	0	0	146	442	0	73.67

Sec E	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	595	0	0	0	0	0	5	99.17
00001	597	0	0	0	0	0	3	0.00
00011	3	583	11	0	0	0	3	1.83
00111	3	0	472	122	0	0	3	20.33
01111	3	0	0	159	435	0	3	72.50
11111	3	0	0	0	59	535	3	89.17

Type 2 sub-bitlines

Sec A	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	600	0	0	0	0	0	0	100.00
00001	0	600	0	0	0	0	0	100.00
00011	0	0	600	0	0	0	0	100.00
00111	0	0	0	600	0	0	0	100.00
01111	0	0	0	0	600	0	0	100.00
11111	0	0	0	0	9	591	0	98.50

Sec B	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	12	0	0	0	0	0	98.00
00001	0	600	0	0	0	0	0	100.00
00011	0	12	588	0	0	0	0	98.00
00111	0	12	0	588	0	0	0	98.00
01111	0	12	0	0	588	0	0	98.00
11111	0	12	0	0	11	577	0	96.17

Sec E	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	12	0	0	0	0	0	98.00
00001	0	600	0	0	0	0	0	100.00
00011	0	12	588	0	0	0	0	98.00
00111	0	12	0	588	0	0	0	98.00
01111	0	12	0	0	588	0	0	98.00
11111	0	12	0	0	0	588	0	98.00

Type 3 sub-bitlines

Sec A	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	600	0	0	0	0	0	0	100.00
00001	0	600	0	0	0	0	0	100.00
00011	0	0	600	0	0	0	0	100.00
00111	0	0	0	600	0	0	0	100.00
01111	0	0	0	0	600	0	0	100.00
11111	0	0	0	0	0	600	0	100.00

Sec B	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	590	4	0	0	0	0	6	98.33
00001	2	592	0	0	0	0	6	98.67
00011	2	4	588	0	0	0	6	98.00
00111	2	4	0	588	0	0	6	98.00
01111	2	4	0	0	588	0	6	98.00
11111	2	4	0	0	0	588	6	98.00

Sec E	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	591	0	0	0	0	0	9	98.50
00001	2	588	0	0	0	0	10	98.00
00011	2	0	588	0	0	0	10	98.00
00111	3	0	0	588	0	0	9	98.00
01111	2	0	0	0	588	0	10	98.00
11111	2	0	0	0	0	588	10	98.00

Type 4 sub-bitlines

Sec A	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	600	0	0	0	0	0	0	100.00
00001	0	600	0	0	0	0	0	100.00
00011	0	0	600	0	0	0	0	100.00
00111	0	0	0	600	0	0	0	100.00
01111	0	0	0	0	600	0	0	100.00
11111	0	0	0	0	0	600	0	100.00

Sec B	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	12	0	0	0	0	0	98.00
00001	0	600	0	0	0	0	0	100.00
00011	0	12	588	0	0	0	0	98.00
00111	0	12	0	588	0	0	0	98.00
01111	0	12	0	0	588	0	0	98.00
11111	0	12	0	0	0	588	0	98.00

Sec E	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	6	0	0	0	6	0	98.00
00001	0	594	0	0	0	6	0	99.00

00011	0	6	588	0	0	6	0	98.00
00111	0	6	0	588	0	6	0	98.00
01111	0	6	0	0	588	6	0	98.00
11111	0	6	0	0	0	594	0	99.00

Type 5 sub-bitlines

Sec A	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	600	0	0	0	0	0	0	100.00
00001	0	600	0	0	0	0	0	100.00
00011	0	0	600	0	0	0	0	100.00
00111	0	0	0	600	0	0	0	100.00
01111	0	0	0	0	600	0	0	100.00
11111	0	0	0	0	0	600	0	100.00

Sec B	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	0	0	0	0	4	8	98.00
00001	0	588	0	0	0	4	8	98.00
00011	0	0	588	0	0	4	8	98.00
00111	0	0	0	588	0	4	8	98.00
01111	0	0	0	0	588	5	7	98.00
11111	0	0	0	0	0	593	7	98.83

Sec E	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	596	0	0	0	0	0	4	99.33
00001	9	588	0	0	0	0	3	98.00
00011	9	0	588	0	0	0	3	98.00
00111	9	0	0	588	0	0	3	98.00
01111	9	0	0	0	588	0	3	98.00
11111	8	0	0	0	0	588	4	98.00

Type 6 sub-bitlines

Sec A	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	596	0	0	0	0	0	4	99.33
00001	0	596	0	0	0	0	4	99.33
00011	0	0	531	65	0	0	4	88.50
00111	0	0	0	596	0	0	4	99.33
01111	0	0	0	0	595	0	5	99.17
11111	0	0	0	0	0	595	5	99.17

Sec B	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	586	12	0	0	1	0	1	97.67
00001	0	598	0	0	0	0	2	99.67
00011	0	12	535	51	0	0	2	89.17
00111	0	12	0	586	0	0	2	97.67
01111	0	12	0	0	586	0	2	97.67
11111	0	12	0	0	0	586	2	97.67

Sec E	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	12	0	0	0	0	0	98.00
00001	0	600	0	0	0	0	0	100.00
00011	0	12	535	53	0	0	0	89.17
00111	0	12	0	587	1	0	0	97.83
01111	0	12	0	0	588	0	0	98.00
11111	0	12	0	0	0	588	0	98.00

Type 7 sub-bitlines

Sec A	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	595	0	0	0	1	4	0	99.17
00001	2	595	0	0	0	0	3	99.17
00011	2	0	595	0	0	0	3	99.17
00111	2	0	0	595	0	0	3	99.17
01111	2	0	0	0	595	0	3	99.17
11111	2	0	0	0	0	595	3	99.17

Sec B	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	586	7	0	0	0	3	4	97.67
00001	0	593	0	0	0	3	4	98.83
00011	0	7	586	0	0	3	4	97.67
00111	0	7	0	586	0	3	4	97.67
01111	0	7	0	0	586	3	4	97.67
11111	0	7	0	0	0	589	4	98.17

Sec E	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	590	0	0	0	0	0	10	98.33
00001	2	588	0	0	0	0	10	98.00
00011	2	0	588	0	0	0	10	98.00
00111	2	0	0	588	0	0	10	98.00
01111	2	0	0	0	588	0	10	98.00
11111	2	0	0	0	0	588	10	98.00

Type 8 sub-bitlines

Sec A	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	590	4	0	0	0	0	6	98.33
00001	23	571	0	0	0	0	6	95.17
00011	0	221	373	0	0	0	6	62.17
00111	0	4	0	590	0	0	6	98.33
01111	0	4	0	0	590	0	6	98.33
11111	0	5	0	0	0	589	6	98.17

Sec B	Read							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	10	0	0	0	0	2	98.00

00001	24	574	0	0	0	0	2	95.67
00011	0	158	440	0	0	0	2	73.33
00111	0	10	2	586	0	0	2	97.67
01111	0	10	0	0	588	0	2	98.00
11111	0	11	0	0	0	587	2	97.83

Sec E	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	586	13	1	0	0	0	0	97.67
00001	24	575	1	0	0	0	0	95.83
00011	0	240	360	0	0	0	0	60.00
00111	0	13	1	586	0	0	0	97.67
01111	0	13	1	0	586	0	0	97.67
11111	0	13	1	0	0	586	0	97.67

Type 9 sub-bitlines

Sec A	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	594	0	0	6	0	0	0	99.00
00001	5	589	0	6	0	0	0	98.17
00011	5	0	589	6	0	0	0	98.17
00111	5	0	0	595	0	0	0	99.17
01111	5	0	0	6	589	0	0	98.17
11111	0	5	0	6	0	589	0	98.17

Sec B	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	0	0	0	0	0	12	98.00
00001	0	588	0	0	0	0	12	98.00
00011	0	0	588	0	0	0	12	98.00
00111	0	0	0	588	0	0	12	98.00
01111	0	0	0	0	588	0	12	98.00
11111	0	0	0	0	0	588	12	98.00

Sec E	**Read**							
Write	00000	00001	00011	00111	01111	11111	Invalid	% Correct
00000	588	0	0	0	0	0	12	98.00
00001	2	585	0	0	0	0	13	97.50
00011	2	0	585	0	0	0	13	97.50
00111	2	0	0	585	0	0	13	97.50
01111	2	0	0	0	585	0	13	97.50
11111	2	0	0	0	0	586	12	97.67

As illustrated from the test data in the table, type 0 and type 1 sub-bitines, i.e., every first column of a group of five columns was the most vulnerable column. Meanwhile, the faulty data read were usually one level below the written data.

Appendix B Test Chip Bitmaps

The following bitmaps are the basic functional test results for sections A, B and E in chips 2, 3 and 5, respectively. In each bitmap, from top to bottom, the nominal voltage levels are 0, 1/5 V_{DD}, 2/5 V_{DD}, 3/5 V_{DD}, 4/5 V_{DD}, V_{DD}.

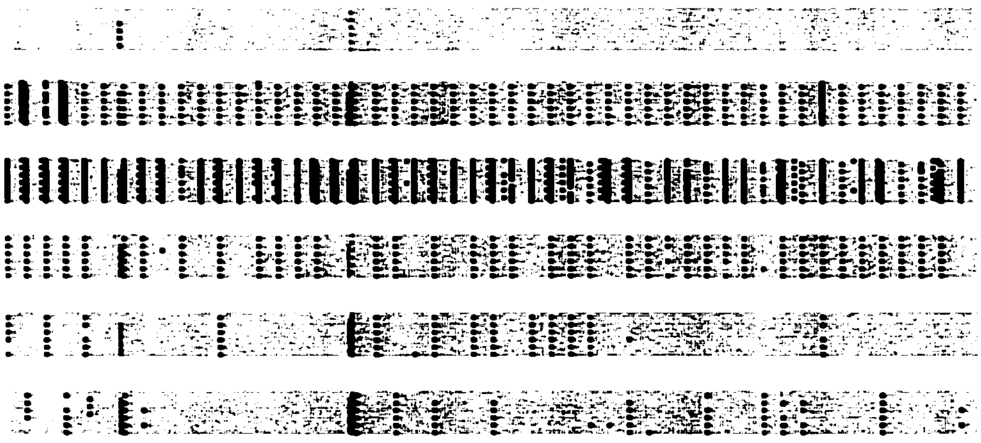


Figure B-1. Bitmap for chip 2 section A

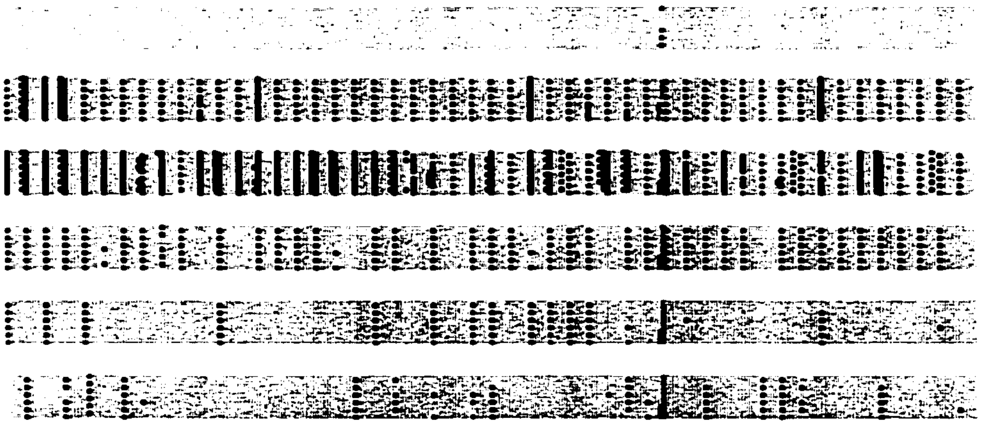


Figure B-2. Bitmap for chip 2 section B

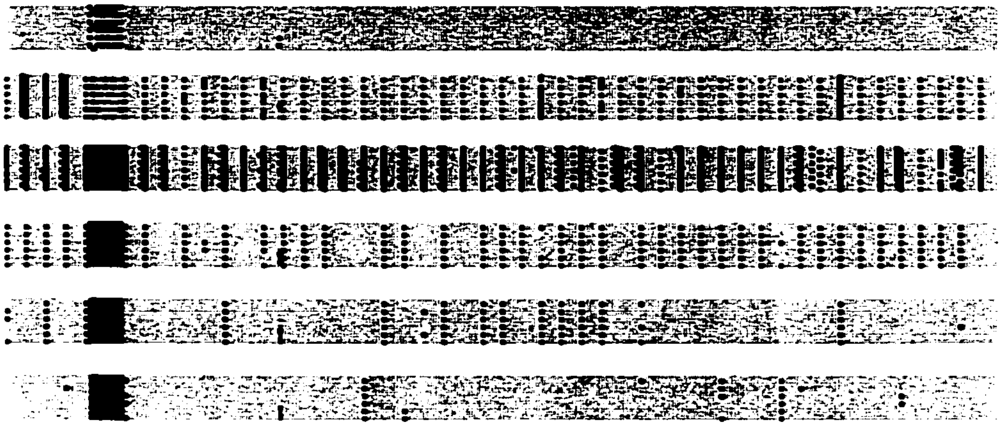


Figure B-3. Bitmap for chip 2 section E

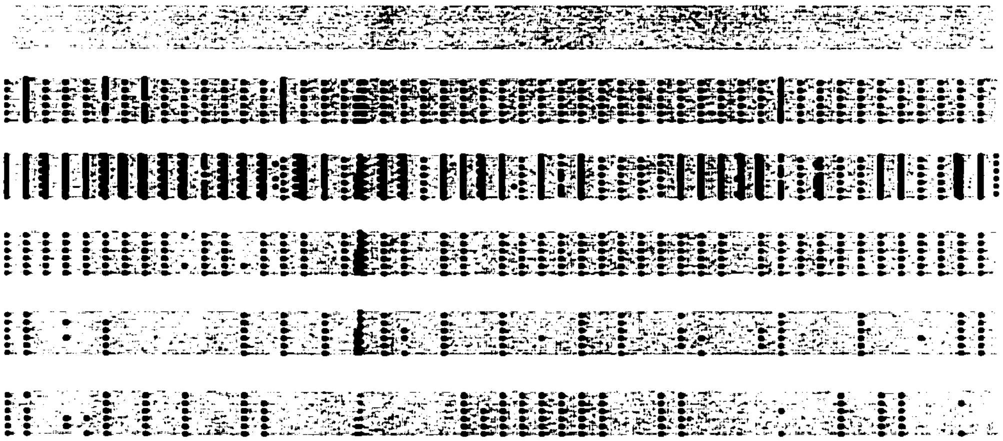


Figure B-4. Bitmap for chip 3 section A

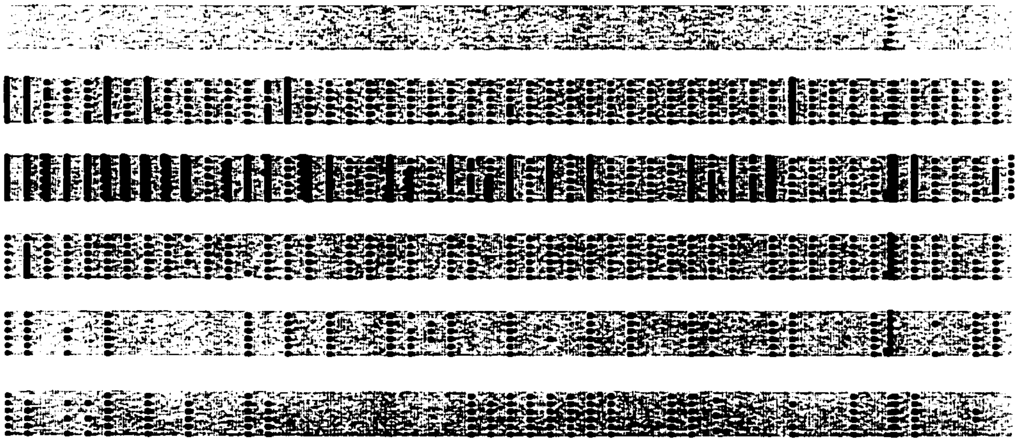


Figure B-5. Bitmap for chip 3 section B

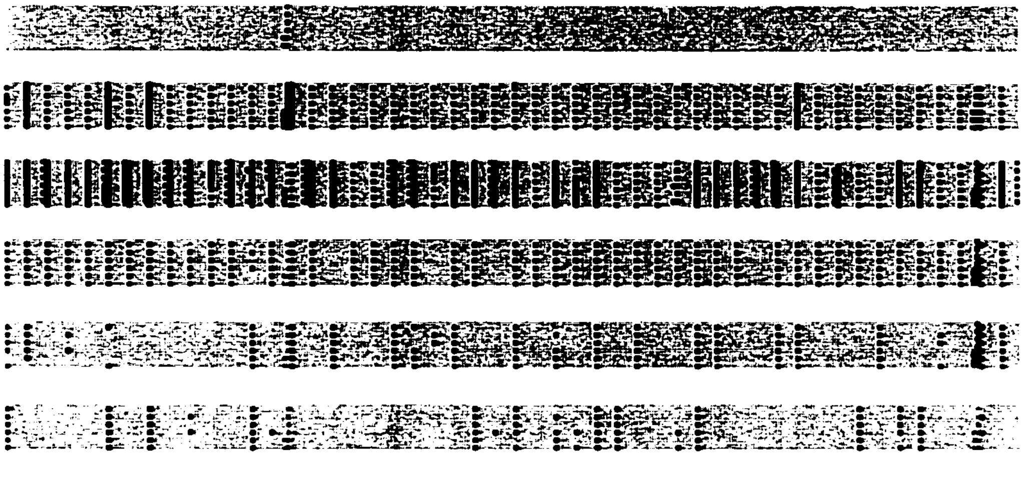


Figure B-6. Bitmap for chip 3 section E

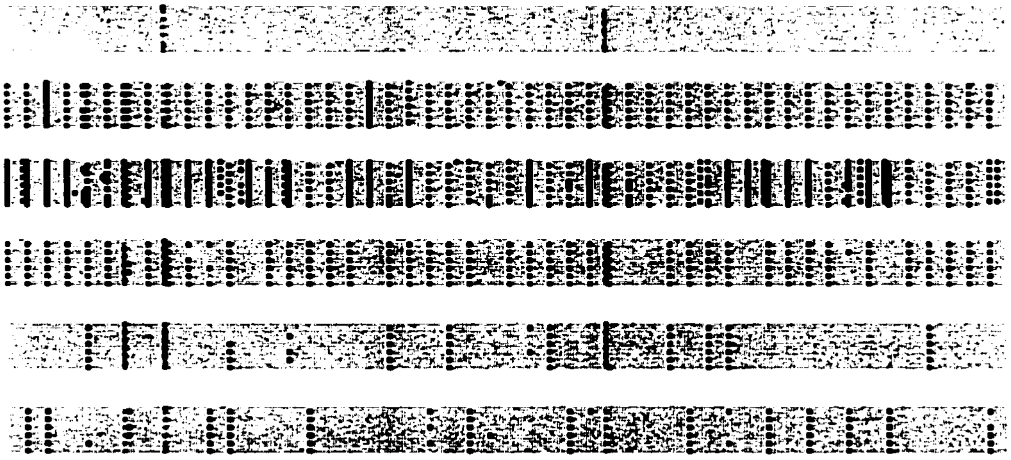


Figure B-7. Bitmap for chip 5 section A

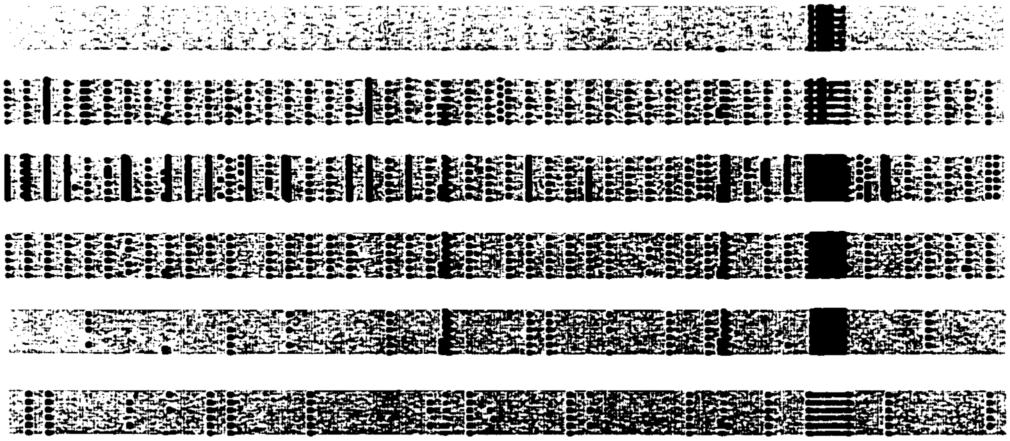


Figure B-8. Bitmap for chip 5 section B

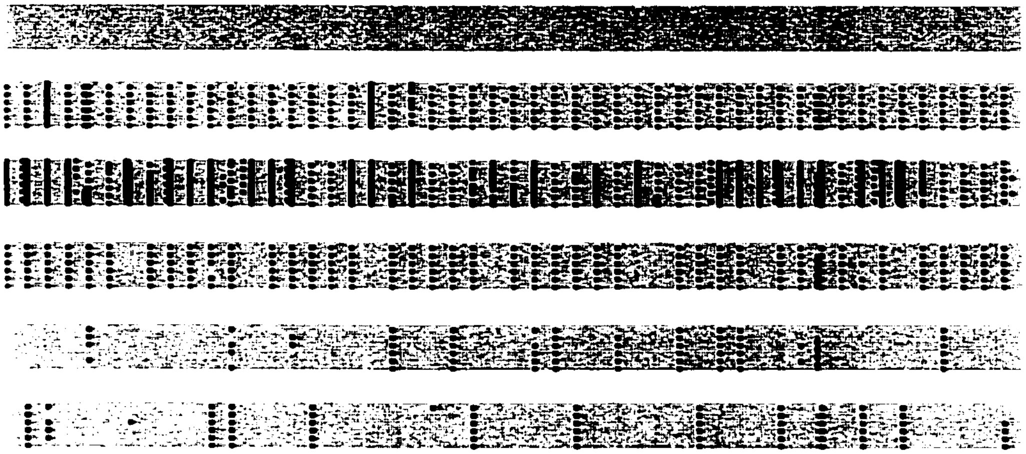


Figure B-9. Bitmap for chip 5 section E