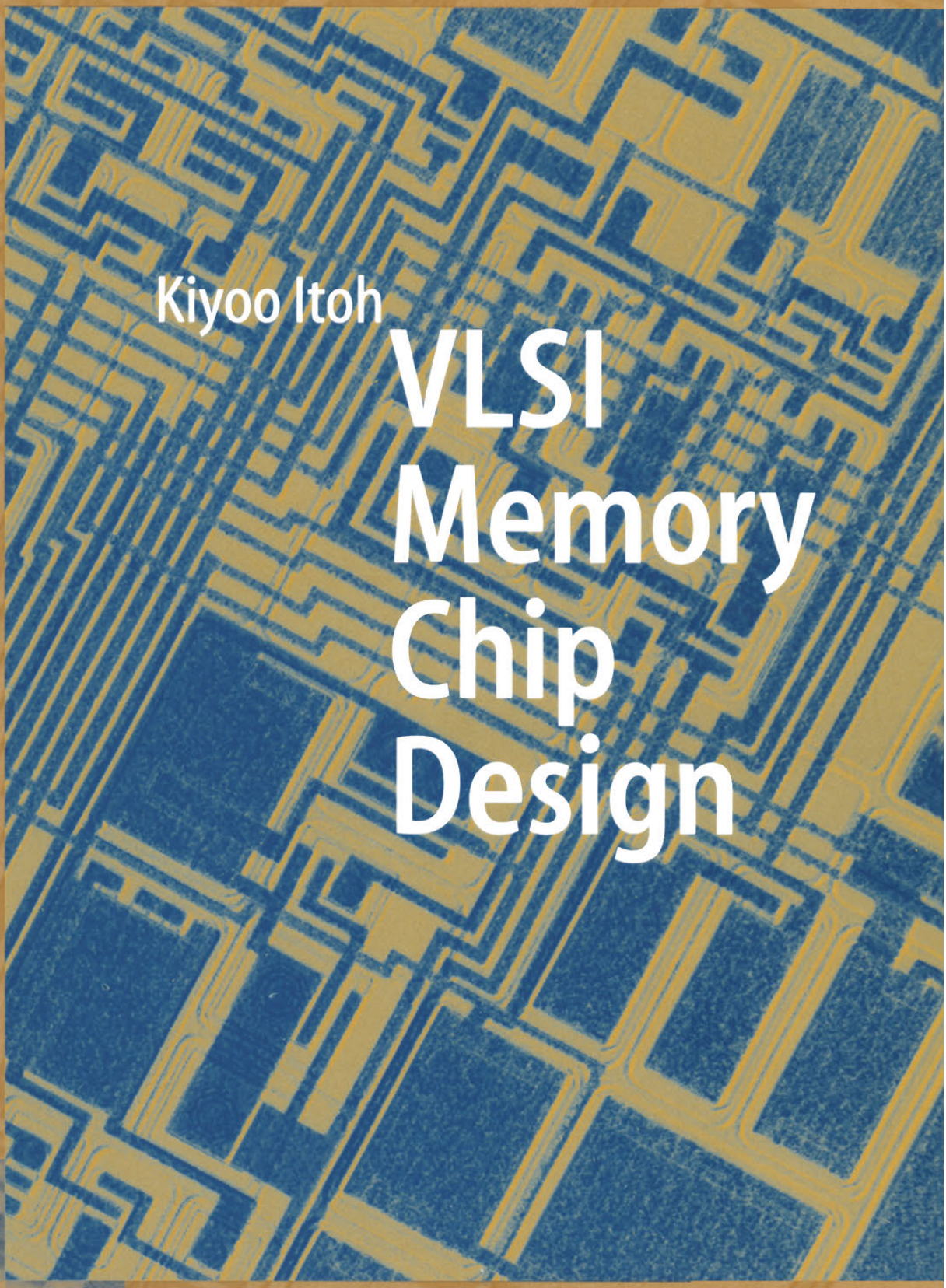
ADVANCED MICROELECTRONICS



Kiyoo Itoh

VLSI Memory Chip Design

Springer

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page (cover-1)

Springer Series in

ADVANCED MICROELECTRONICS 5

Physics and Astronomy ONLINE LIBRARY

Springer-Verlag Berlin Heidelberg GmbH http://www.springer.de/phys/

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page (cover-2)

Springer Series in

ADVANCED MICROELECTRONICS

Series editors: K. Itoh, T. Sakurai

The Springer Series in Advanced Microelectronics provides systematic information on all the topics relevant for the design, processing, and manufacturing of microelectronic devices. The books, each prepared by leading researchers or engineers in their fields, cover the basic and advanced aspects of topics such as wafer processing, materials, device design, device technologies, circuit design, VLSI implementation, and subsystem technology. The series forms a bridge between physics and engineering and the volumes will appeal to practicing engineers as well as research scientists.

1 **Cellular Neural Networks**
Chaos, Complexity and VLSI Processing
By G. Manganaro, P. Arena, and L. Fortuna

2 **Technology of Integrated Circuits**
By D. Widmann, H. Mader, and H. Friedrich

3 **Ferroelectric Memories**
By J. F. Scott

4 **Microwave Resonators and Filters for Wireless Communication**
Theory, Design and Application
By M. Makimoto and S. Yamashita

5 **VLSI Memory Chip Design**
By K. Itoh

Series homepage – http://www.springer.de/phys/books/ssam/

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page (cover-3)

Kiyoo Itoh

VLSI Memory Chip Design

With 416 Figures and 26 Tables

Springer

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page (cover-4)

Dr. Kiyoo Itoh
Hitachi Ltd., Central Research Laboratory
1-280, Higashi-Koigakubo
Kokubunji-shi
Tokyo185-8601
Japan
e-mail: k-itoh@crl.hitachi.co.jp

Series Editors:

Dr. Kiyoo Itoh
Hitachi Ltd., Central Research Laboratory
1-280 Higashi-Koigakubo
Kokubunji-shi
Tokyo 185-8601
Japan

Professor Takayasu Sakurai
Center for Collaborative Research
University of Tokyo
7-22-1 Roppongi, Minato-ku,
Tokyo 106-8558
Japan

Library of Congress Cataloging-in-Publication Data

Itoh, Kiyoo, 1941-
VLSI memory chip design / Kiyoo Itoh.
p. cm. -- (Springer series in advanced microelectronics ; 5)
Includes bibliographical references and index.

1. Semiconductor storage devices--Design and construction. 2. Integrated circuits--Very large scale integration--Design and construction. I. Title. II. Series.

TK7895.M4 I876 2001
621.39'732--dc21

00-068735

ISBN 978-3-642-08736-3 ISBN 978-3-662-04478-0 (eBook)
DOI 10.1007/978-3-662-04478-0

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

http://www.springer.de

© Springer-Verlag Berlin Heidelberg 2001

Originally published by Springer-Verlag Berlin Heidelberg New York in 2001.
Softcover reprint of the hardcover 1st edition 2001

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting by the author.
Data conversion and figure processing by LE-TEX Jelonek, Schmidt & Vöckler GbR, 04229 Leipzig.

Cover concept by eStudio Calmar Steinen using a background picture from Photo Studio "SONO". Courtesy of Mr. Yukio Sono, 3-18-4 Uchi-Kanda, Chiyoda-ku, Tokyo.
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11327046 57/3111/mf - 5 4 3 2 1

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page (cover-5)

Contents

1. An Introduction to Memory Chip Design 1
- 1.1 Introduction 1
- 1.2 The Internal Organization of Memory Chips 3
  - 1.2.1 The Memory Cell Array 3
  - 1.2.2 The Peripheral Circuit 5
  - 1.2.3 The I/O Interface Circuit 6
- 1.3 Categories of Memory Chip 6
- 1.4 General Trends in DRAM Design and Technology 11
  - 1.4.1 The History of Memory-Cell Development 11
  - 1.4.2 The Basic Operation of The 1-T Cell 15
  - 1.4.3 Advances in DRAM Design and Technology 19
- 1.5 General Trends in SRAM Design and Technology 24
  - 1.5.1 The History of Memory-Cell Development 24
  - 1.5.2 The Basic Operation of a SRAM Cell 26
  - 1.5.3 Advances in SRAM Design and Technology 29
- 1.6 General Trends in Non-Volatile Memory Design and Technology 31
  - 1.6.1 The History of Memory-Cell Development 31
  - 1.6.2 The Basic Operation of Flash Memory Cells 34
  - 1.6.3 Advances in Flash-Memory Design and Technology 46

2. The Basics of RAM Design and Technology 49
- 2.1 Introduction 49
- 2.2 Devices 49
  - 2.2.1 MOSFETs 49
  - 2.2.2 Capacitors 57
  - 2.2.3 Resistors 60
  - 2.2.4 Wiring and Wiring Materials 61
  - 2.2.5 Silicon Substrates and CMOS Latch-Up 65
  - 2.2.6 Other Devices 67
- 2.3 NMOS Static Circuits 67
  - 2.3.1 The dc Characteristics of an Inverter 68
  - 2.3.2 The ac Characteristics of an Inverter 70
  - 2.3.3 The Improved NMOS Static Inverter 74

2.4 NMOS Dynamic Circuits ... 76
2.4.1 The Dynamic Inverter ... 76
2.4.2 The Bootstrap Driver ... 77
2.5 CMOS Circuits ... 79
2.5.1 The dc Characteristics ... 80
2.5.2 The ac Characteristics ... 82
2.6 Basic Memory Circuits ... 83
2.6.1 The Inverter and the Basic Logic Gate ... 83
2.6.2 The Current Mirror ... 83
2.6.3 The Differential Amplifier ... 83
2.6.4 The Voltage Booster ... 87
2.6.5 The Level Shifter ... 88
2.6.6 The Ring Oscillator ... 88
2.6.7 The Counter ... 89
2.7 The Scaling Law ... 90
2.7.1 Constant Electric-Field Scaling ... 90
2.7.2 Constant Operation-Voltage Scaling ... 92
2.7.3 Combined Scaling ... 92
2.8 Lithography ... 93
2.9 Packaging ... 94

3. DRAM Circuits ... 97
3.1 Introduction ... 97
3.1.1 High-Density Technology ... 98
3.1.2 High-Performance Circuits ... 100
3.2 The catalog Specifications of the Standard DRAM ... 102
3.2.1 Operational Conditions ... 102
3.2.2 Modes of Operation and Timing Specifications ... 105
3.3 The Basic Configuration and Operation of the DRAM Chip ... 110
3.3.1 Chip Configuration ... 110
3.3.2 Address Multiplexing ... 111
3.4 Fundamental Chip Technologies ... 113
3.4.1 A Larger Memory Capacity and Scaled-Down Devices ... 113
3.4.2 High S/N Ratio Circuits ... 116
3.4.3 Low Power Circuits ... 117
3.4.4 High-Speed Circuits ... 123
3.4.5 The Multidivision of a Memory Array ... 128
3.5 The Multidivided Data Line and Word Line ... 131
3.5.1 The Multidivided Data Line ... 132
3.5.2 The Multidivided Word Line ... 139
3.6 Read and Relevant Circuits ... 141
3.6.1 The Address Buffer ... 141
3.6.2 The Address Decoder ... 144
3.6.3 The Word Driver ... 147
3.6.4 The Sensing Circuit ... 157

3.6.5 The Common I/O-Line Relevant Circuit 167
3.6.6 The Data-Output Buffer 172
3.7 Write and Relevant Circuits 174
3.8 Refresh-Relevant Circuits 175
3.8.1 Refresh Schemes 175
3.8.2 The Extension of Data-Retention Time in Active Mode 176
3.8.3 Current Reduction Circuits in Data-Retention Mode .. 176
3.9 Redundancy Techniques 178
3.9.1 Issues for Large-Memory-Capacity Chips 184
3.9.2 Intra-Subarray Replacement Redundancy 185
3.9.3 Inter-Subarray Replacement Redundancy 189
3.9.4 The Repair of dc-Characteristics Faults 191
3.10 On-Chip Testing Circuits 192

4. High Signal-to-Noise Ratio DRAM Design and Technology 195
4.1 Introduction 195
4.2 Trends in High S/N Ratio Design 195
4.2.1 The Signal Charge 197
4.2.2 Leakage Charge 204
4.2.3 The Soft-Error Critical Charge 208
4.2.4 The Data-Line Noise Charge 210
4.3 Data-Line Noise Reduction 210
4.3.1 Noise Sources and Their Reduction 210
4.3.2 Word-Line Drive Noise 213
4.3.3 Data-Line and Sense-Amplifier Imbalances 217
4.3.4 Word-Line to Data-Line Coupling Noise 230
4.3.5 Data-Line Interference Noise 237
4.3.6 Power-Supply Voltage Bounce 240
4.3.7 Variation in the Reference Voltage 241
4.3.8 Other Noises 244
4.4 Summary 247

5. On-Chip Voltage Generators 249
5.1 Introduction 249
5.2 The Substrate-Bias Voltage (V_{BB}) Generator 251
5.2.1 The Roles of the V_{BB} generator 251
5.2.2 Basic Operation and Design Issues 256
5.2.3 Power-On Characteristics 258
5.2.4 Characteristics in the High-V_{DD} Region 264
5.2.5 The V_{BB} Bump 266
5.2.6 Substrate-Current Generation 269
5.2.7 Triple-Well Structures 272
5.2.8 Low-Power V_{BB} Generators 273

5.3 The Voltage Up-Converter 276
5.3.1 The Roles of the Voltage Up-Converter 276
5.3.2 Design Approaches and Issues 278
5.3.3 High Boost-Ratio Converters 283
5.3.4 Low-Power, High Supply Current Converters 285
5.4 The Voltage Down-Converter 290
5.4.1 The Roles of the Voltage Down-Converter 290
5.4.2 The Negative-Feedback Converter and Design Issues . . 293
5.4.3 Optimum Design 297
5.4.4 Phase Compensation 301
5.4.5 Reference-Voltage Generators 316
5.4.6 Burn-In Test Circuits 323
5.4.7 Voltage Trimming 327
5.4.8 Low-Power Circuits 329
5.5 The Half-V_{DD} Generator 332
5.6 Examples of Advanced On-Chip Voltage Generators 333

6. High-Performance Subsystem Memories 339
6.1 Introduction 339
6.2 Hierarchical Memory Systems 341
6.2.1 Memory Hierarchy 341
6.2.2 Improvements in Memory-Subsystem Performance 344
6.2.3 Memory-Chip Performance 349
6.3 Memory-Subsystem Technologies 354
6.3.1 Wide-Bit I/O Chip Configurations 354
6.3.2 Parallel Operation of Multidivided Arrays 354
6.3.3 Multibank Interleaving 357
6.3.4 Synchronous Operation 358
6.3.5 Pipeline/Prefetch Operations 362
6.3.6 High-Speed Clocking Schemes 363
6.3.7 Terminated I/O Interfaces 363
6.3.8 High-Density Packaging 364
6.4 High-Performance Standard DRAMs 365
6.4.1 Trends in Chip Development 365
6.4.2 Synchronous DRAM 368
6.4.3 Rambus DRAM 380
6.5 Embedded Memories 383

7. Low-Power Memory Circuits 389
7.1 Introduction 389
7.2 Sources and Reduction of Power Dissipation in a RAM Subsystem 392
7.2.1 Wide-Bit I/O Chip Configuration 393
7.2.2 Small Package 394
7.2.3 The Low-Voltage Data-Bus Interface 396

7.3 Sources of Power Dissipation in the RAM Chip 402
7.3.1 Active Power Sources 402
7.3.2 Data-Retention Power Sources 405
7.4 Low-Power DRAM Circuits 406
7.4.1 Active Power Reduction 406
7.4.2 Data-Retention Power Reduction 412
7.5 Low-Power SRAM Circuits 413
7.5.1 Active Power Reduction 413
7.5.2 Data-Retention Power Reduction 423

8. Ultra-Low-Voltage Memory Circuits 425
8.1 Introduction 425
8.2 Design Issues for Ultra-Low-Voltage RAM Circuits 426
8.2.1 Reduction of the Subthreshold Current 426
8.2.2 Stable Memory-Cell Operation 432
8.2.3 Suppression of, or Compensation for, Design Parameter Variations 433
8.2.4 Power-Supply Standardization 435
8.3 Ultra-Low-Voltage DRAM Circuits 437
8.3.1 Gate Boosting Circuit 439
8.3.2 The Multi-V_T Circuit 440
8.3.3 The Gate–Source Back-Biasing Circuit 442
8.3.4 The Well Control Circuit 456
8.3.5 The Source Control Circuit 461
8.3.6 The Well and Source Control Circuit 462
8.4 Ultra-Low-Voltage SRAM Circuits 463
8.5 Ultra-Low-Voltage SOI Circuits 466

References 473

Index 489

1. An Introduction to Memory Chip Design

1.1 Introduction

Several essential inventions and innovations, and subsequent sustained efforts [1.1] toward high densities have paved the way to large-scale integrated circuit (LSI) memories, as shown in Fig. 1.1 [1.2]. Since two epoch-making announcements accompanying the start of LSI memory production in 1970 [the first extensive usage of a semiconductor memory chip for the IBM 370 mainframe computers, and the first sales of a 1-Kb dynamic random access memory (DRAM), named the 1103, from Intel], the increase in memory chip capacity has skyrocketed with the help of the ever-higher-density

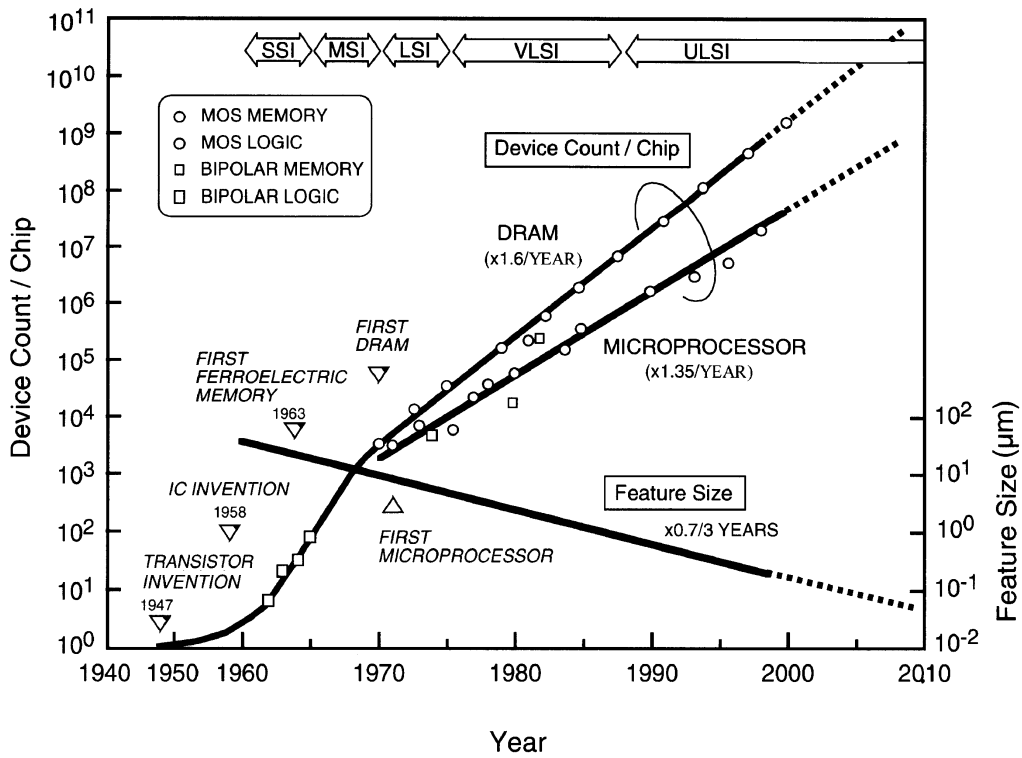


Fig. 1.1. The trend in the device count per chip [1.2]

MOS/CMOS design and technology. The resultant LSI memories have given computers, networks, and almost everything with electrical components the benefit of a dramatically reduced cost per bit and far superior performance. Data processors and data terminals, such as personal computers, workstations, and POS terminals, as well as telephone exchanges, digital televisions, and numerical control machines, could not have been produced without them.

The first priority of LSI development is high-density technology. In the past, technology development has been driven by memory because of its inherent features: a strong need for low cost, that is mainly realized by higher-density technology, and a regular structure of memory chips, that enables easy failure analysis as well as the application of redundancy techniques. Thus, even higher-density technology quickly develops into mass production with an acceptable yield. Fortunately, the vast level of chip production may pay off the tremendous effort involved. Currently, 0.2-μm CMOS technology is being used in the manufacturing of 64-Mb DRAMs, as well as over 400-MHz 32-bit microprocessors, with the numbers of transistors per chip in the 10^7-10^8 range. Even below 0.18-μm, CMOS technology has been used in experimental 1−4 Gb DRAMs [1.3] that incorporate over 2−4 billion electrical components, revealing a rapid increase in memory-chip capacity by more than six orders of magnitude (1 Kb−4 Gb) in the past 30 years, since the DRAM advent in 1970. Technologies necessary for 0.1 μm or less are also being investigated and reported at conferences. A typical example of the high-density technology contributing to these advances is fine pattern technology with shorter light-source wavelengths, as well as the larger wafer and related technology for ever-larger chips. High-density/low-power devices and circuits, such as the one-transistor, one-capacitor memory cell [1.4] and CMOS technology [1.5], are also contributors, as shown in Chap. 2. The resultant reduction in the bit cost of DRAM is shown in Fig. 1.2 [1.7]. Cost has been cut dramatically, at an annual reduction of around 24%, proving of great benefit to system designers. Sometimes, however, excessive competition aiming at a bigger slice of the pie has made the market less profitable for all, with drops in cost, as exemplified by the 256-Kb and 16-Mb generations. The drops has long been experienced as the "silicon cycle", which is fundamentally caused by the relation between supply and demand since the advent of DRAM. Note that the annual bit-cost reduction is around 27%, if the saturated cost of each DRAM generation is assumed to be the same. The difference between the 24% above and 27% implies an ever-increasing chip cost in each successive generation.

SRAM (Static Random Access Memory) using flip-flop memory cells, and non-volatile memories such as ROM (Read-Only Memory) and Flash memory, have advanced with almost the same technology as DRAM. In particular, the invention of non-volatile memories utilizing a floating gate structure is noteworthy [1.6]. Sales of these memory chips have increased annually, as shown in Fig. 1.3 [1.8].

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 2

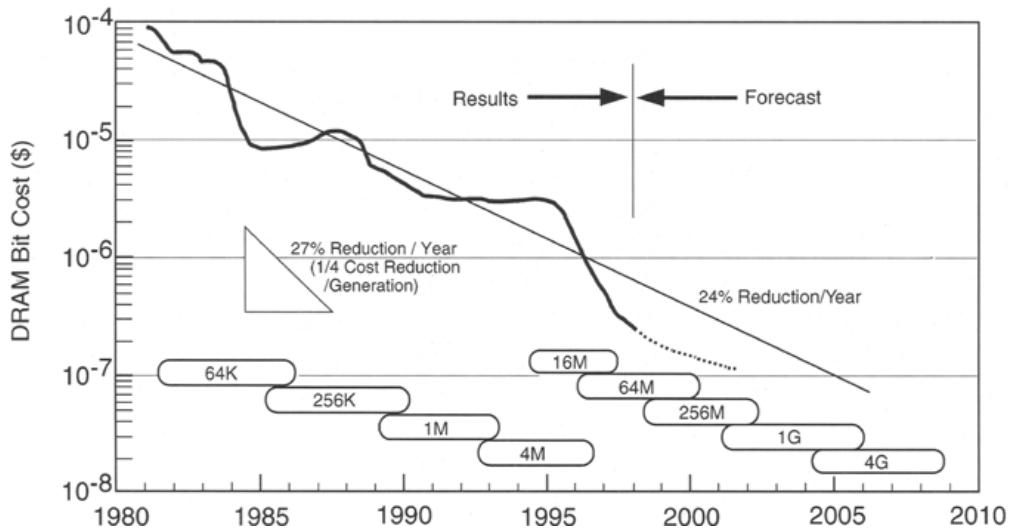


Fig. 1.2. The reduction in the bit cost of DRAM [1.7]

In general, the semiconductor chip is called by the following names, depending on the number of components integrated in the chip: IC (integrated circuit for less than 10^3), LSI (large-scale IC for 10^3-10^5), VLSI (very-large-scale IC for 10^5-10^7), ULSI (ultra-large-scale IC for more than 10^7). Sometimes LSI, VLSI, and ULSI are comprehensively referred to as LSI or VLSI.

In this chapter, the fundamentals of MOS memory-chip technology are described. First, the category of MOS memory chip, the internal organization of the chip, and technology trends in memory chips are discussed. Next, the individual trends of DRAM, SRAM, and non-volatile memory technologies are investigated.

1.2 The Internal Organization of Memory Chips

A memory chip is composed of three blocks [1.9]: a memory cell array, a peripheral circuit, and an input/output (I/O) interface circuit, as shown in Fig. 1.4.

1.2.1 The Memory Cell Array

A memory cell array comprising a matrix of 2^N rows and 2^M columns can store binary information of 2^{N+M} bits. For example, if $N+M=20$, a memory chip can store 1 Mbit of information, and is called a 1 Mbit (or simply 1 M or 1 Mb) memory chip. Here M denotes 1024 K, with K = 1024. If $N+M=30$, we call it a 1 Gbit (1 G or 1 Gb) memory chip, with G = 1024 M. Any cell can be accessed at random with the same speed by selecting both the

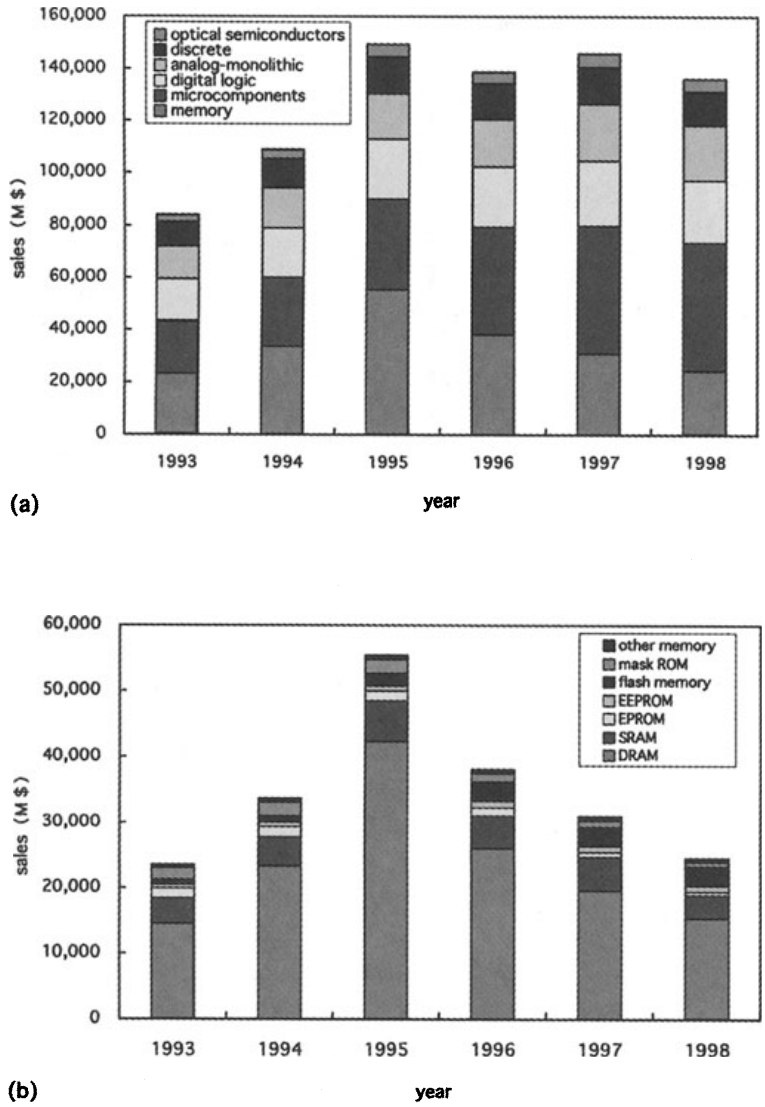


Fig. 1.3. The amount of sales of LSIs (**a**) and various memory chips (**b**) [1.8]

corresponding row and column. Sometimes, the memory cell array is called memory cell matrix, a memory array, or simply an array. The row is also called the X line or word line, while the column is called the Y line, bit line, or data line. Note that a matrix arrangement minimizes the number of driving circuits of memory cells: the number is $(2^N + 2^M)$, which is a minimum at $N = M$ for the matrix (two-dimensional) arrangement, while it is 2^{N+M} for a one-dimensional arrangement, as shown in Fig. 1.5. For a 1 Mb chip, the matrix arrangement reduces the number from about one million to about two thousand.

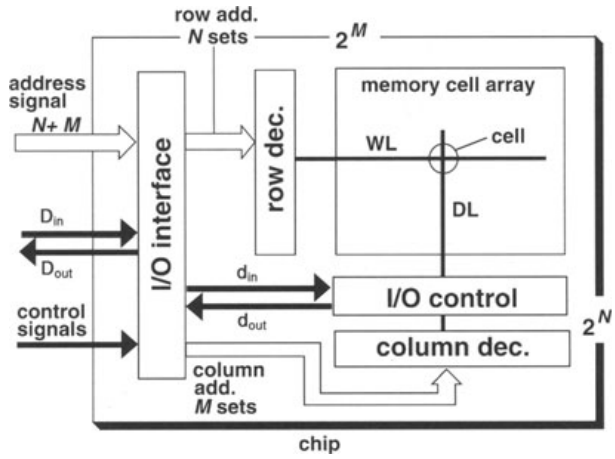


Fig. 1.4. The memory chip configuration

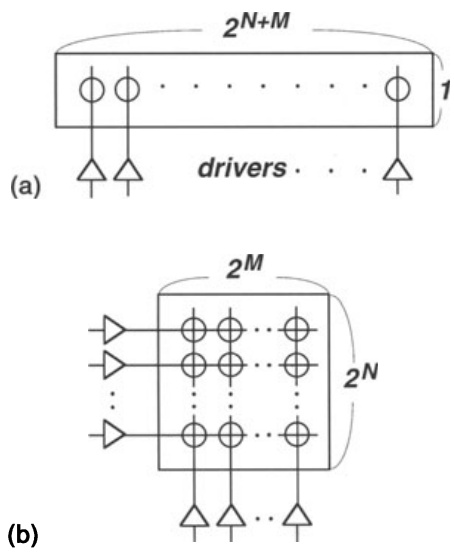


Fig. 1.5. Memory cell arrangements [1.9]. (**a**) One-dimensional; (**b**) two-dimensional

1.2.2 The Peripheral Circuit

The block bridges between the memory array and the I/O interface circuit so that they can communicate with each other. It sends write data to a memory cell in the memory array under the control of the interface circuit, or sends read data from the memory cell to the interface circuit. A typical circuit here is a decoder, which is composed of many logic circuits. It selects a logic circuit corresponding to one row or one column, based on a logical set of N or M address signals from address buffers in the interface circuit. Then, a row or column is driven by a driver, which is connected to the output of the logic circuit. Here, an I/O control circuit in the figure controls the transfer of the write or read data in the memory array.

1.2.3 The I/O Interface Circuit

This converts external signals, such as addresses, clocks, control signals, and data inputs, to the corresponding internal signals that activate the peripheral circuit. In addition, it outputs read data from the array as the data output of the chip. Address buffers in the I/O interface circuit generate N and M sets of complementary row and column address signals, respectively, through the use of $(N + M)$ external address signals. Data input and output buffers, a write control buffer, and control clock circuits are also typical components of the I/O interface circuit. Note that, in general, ROM does not need data input buffers. Here, access time is defined as the time from the start of chip activation by a read control signal to the outputting of the resultant data from the chip. The cycle time is defined as the time from the start of one chip activation for a read or write operation to the start of the next chip activation.

1.3 Categories of Memory Chip

The semiconductor memory [1.9] now widely used is categorized as RAM (Random Access Memory) and ROM (Read Only Memory), as shown in Fig. 1.6. RAM permits random write and read operations for any memory cell in a chip. The stored data are usually volatile when the power-supply voltage is turned off. RAM is further classified into DRAM (Dynamic RAM) and SRAM (Static RAM). Due to the advantage of low cost, despite medium speed, DRAM is widely and extensively used for the main memory in personal computers, mainframe computers, and engineering workstations. SRAM, which features high speed and ease of use, despite high cost, is also used for the main memory of supercomputers, the cache memory in mainframe

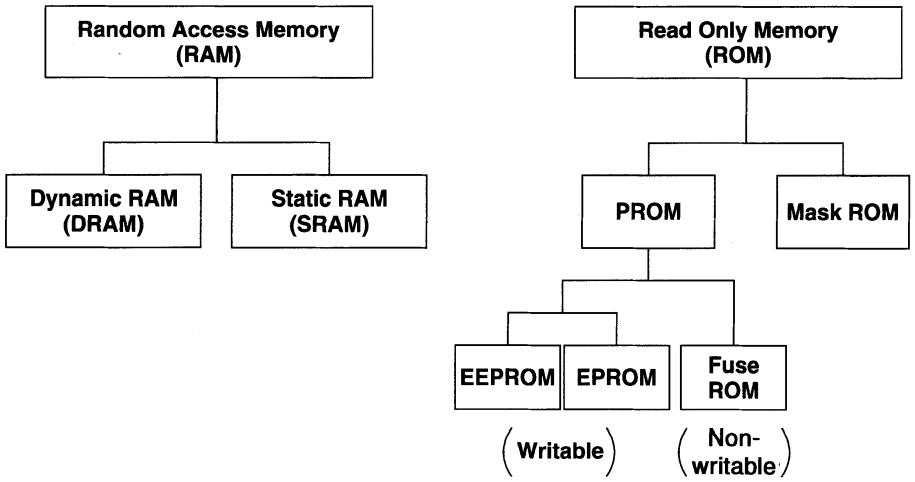


Fig. 1.6. Categories of semiconductor memories widely used [1.9]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 6

computers, engineering workstations, and microprocessors, and memory in handheld equipment.

ROM, dedicated to read operation with non-volatile stored data, is classified into two categories, depending on the process of the write procedure: PROM (Programmable ROM), in which data are written after chip fabrication; and Mask ROM, in which data are written during chip fabrication by the use of a photo mask that contains the write data. PROM is further categorized into two, depending on its erasable or non-erasable characteristics; EPROM (Erasable and Electrically PROM) and EEPROM (Electrically Erasable PROM), and Fuse ROM. EPROM erases the data by exposing the memory cells to ultraviolet rays, while it writes the data by electrical means. In EEPROM, the erase and write operations are both performed by electrical means. There are some drawbacks: the write speed is two to three orders slower than that of RAM, and there is an upper limit on the number of write operations of $10^4 - 10^5$. Note that Flash memory – a kind of EEPROM – is beeing intensively developed and has emerged in the market offering the potential of high density and low cost, in some usages, it may replace magnetic disk memory. In Fuse ROM, users cannot rewrite data once they are written, because the fuses will be blown. ROM, which is generally cheaper than RAM, is used as memory for character-fonts, games machines, fax machines, telephones, engine-controls, and so on.

In addition to the well-known serial memories such as shift registers, non-volatile RAMs utilizing the hysteresis characteristics of ferroelectric and magnetic materials, various application-specific memories such as video memories, and merged RAM and logic LSIs are also beeing intensively developed.

Figure 1.7 shows schematic circuits of memory cells, which store binary information, "1" or "0", on the above-described memory chips. N-channel MOSFETs (NMOSFETs) are used, due to ease of explanation. A DRAM cell [1.4] comprises a MOSFET, that works as a switch, and a capacitor for storing charges. For example, non-existence of charges (electrons for NMOSFET) at the capacitor corresponds to "1", while existence of charges corresponds to "0". In other words, for the voltage expression, a high stored

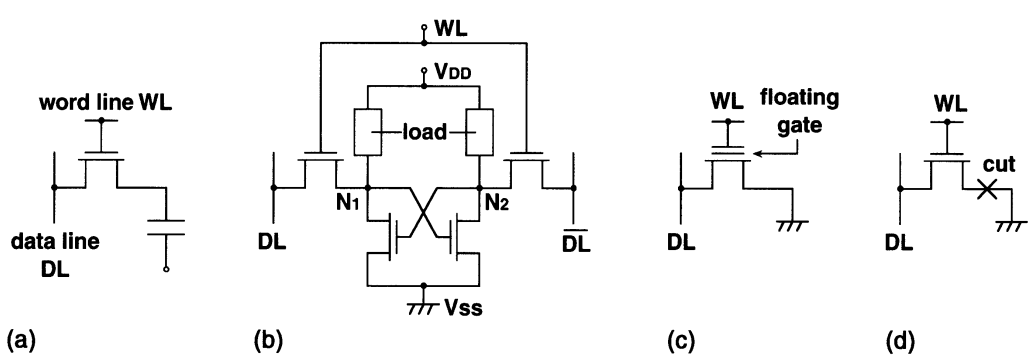


Fig. 1.7. Memory cell circuits [1.9]. (**a**) DRAM; (**b**) SRAM; (**c**) EPROM and EEPROM; (**d**) Mask ROM and Fuse ROM

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 7

voltage corresponds to "1" while a low stored voltage corresponds to "0". The write operation is performed by turning on the switch and applying a voltage corresponding to the write data from the data line, DL, to the capacitor. Here, the switch is turned on by applying a sufficiently high voltage to the word line, WL. The read operation is performed by turning on the switch. A resultant signal voltage developed on the data line, depending on the stored data at the capacitor, is discriminated by a detector on the data line. In principle, the cell holds the data without power consumption. Actually, however, a leakage current at the p–n junction in the storage node degrades an initial high stored voltage, finally causing the loss of information. This loss can be avoided by a "refresh" operation: the cell is read before the stored voltage has become excessively decayed, and then it is rewritten by utilizing the resultant read information, so that the voltage is restored to its initial value. A succession of read–rewrites operation at a given time interval retains the data. The time interval, which is determined by the leakage current, is about 2–64 ms. The name DRAM is derived from the fact that data is dynamically retained by refresh operations, which differs from SRAM.

An SRAM cell consists of a flip-flop circuit that is constructed between a power supply voltage (V_{DD}) and the ground (V_{SS}), and two switching MOSFETs. Data is communicated between a pair of data lines, DL and $\overline{DL}$, and a flip-flop by turning on two FETs. The write operation is performed by applying a differential voltage between a high voltage (H) and a low voltage (L) to a pair of data lines and thus to the storage nodes, N_1 and N_2. For example, "1" is written for a polarity of H at DL (N_1) and L at $\overline{DL}$ (N_2) while "0" is written for an opposite polarity of L at DL and H at $\overline{DL}$. The read operation is performed by detecting the polarity of a differential signal voltage developed on the data lines. No refresh operation is needed, because the leakage currents at N_1 and N_2, if any, are compensated by a static current from the power supply, as long as V_{DD} is supplied. Thus it allows ease of use, although the use of more FETs in a cell increases the memory cell area to more than four times that of DRAM.

Unlike RAMs, a PROM cell needs an additional erase operation, because it must be initialized to a state of non-existence of electrons by extracting electrons from the floating gate (i.e. the storage node). The succeeding write operation is achieved by either injecting electrons to the floating gate or not doing so. For example, "0" is written for the injection of electrons, while "1" is written for the non-injection state, that is, the erased state. The read operation is performed by capacitive coupling between the word line and the floating gate. For "0", the transistor is kept off even with the help of a positive pulse coupled from WL to the floating gate, because electrons at the floating gate prevent the FET from turning on. For "1", however, the FET turns on. Thus, a detector on the data line can differentiate the currents to discriminate the information. Note that, in principle, the electrons injected never discharge because they are stored at the floating gate and surrounded

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 8

by pure insulators. Data retention is ensured even when the power supply is off, thus realizing a non-volatile cell. In an EPROM chip, the stored data of all of the cells are simultaneously erased by means of exposing the memory cells on the chip to ultraviolet rays through the crystal glass lid of the chip package. In an EEPROM chip, the data are electrically erased by a tunnel current. The erase operation is normally done for every unit of 8 bits. In particular, an EEPROM in which all the data in a chip are simultaneously erased is called Flash memory. In EEPROM, there are two kinds of write-operation mechanisms: the injection of hot electrons, generated by avalanche breakdown phenomena at the drain of the cell FET; and the injection of electrons generated by a tunnel effect. In a Fuse ROM cell, the write operation is accomplished by the blowing of program devices, such as fuses or p–n diodes connected to the cell FET. Thus the on or off state of the cell FET can be programmed according to the write data. The resultantly destroyed program devices, however, permit only one write operation, as described before. In a Mask ROM, a mask pattern that programs the on or off state of each cell FET is used. It offers the smallest cell area and ease of fabrication, allowing the largest memory capacity and the lowest cost, despite its limited function.

Figure 1.8 shows trends in memory cell area for various memories [1.9, 1.44] which have been presented at major conferences. Both DRAM and SRAM cells have been miniaturized at a pace of about one-fiftieth per 10 years. Recently, however, there has been a saturation due to the ever-difficult process of device miniaturization. Flash memory is about to catch up with DRAM, by using the one-transistor, one-capacitor cell (the 1-T cell). Figure 1.9 shows trends in the memory capacity of VLSI memories at the

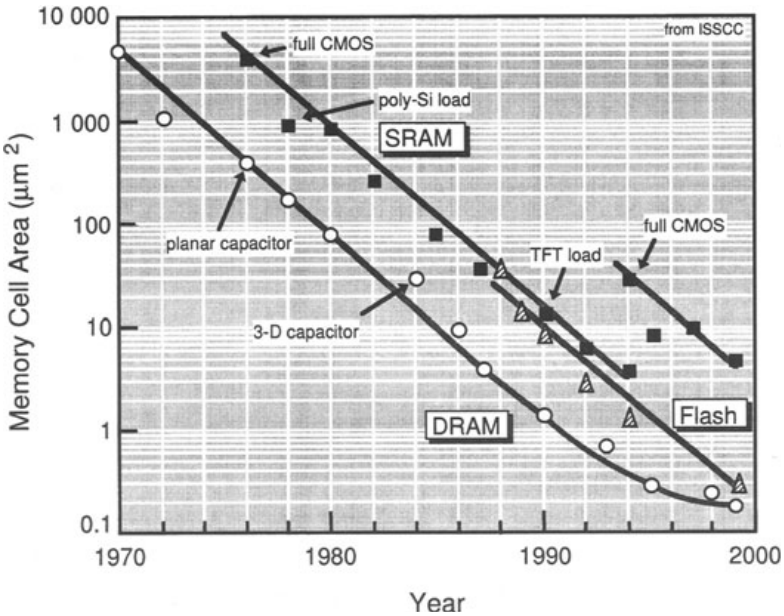


Fig. 1.8. Trends in the memory-cell area of VLSI memories [1.9, 1.44]

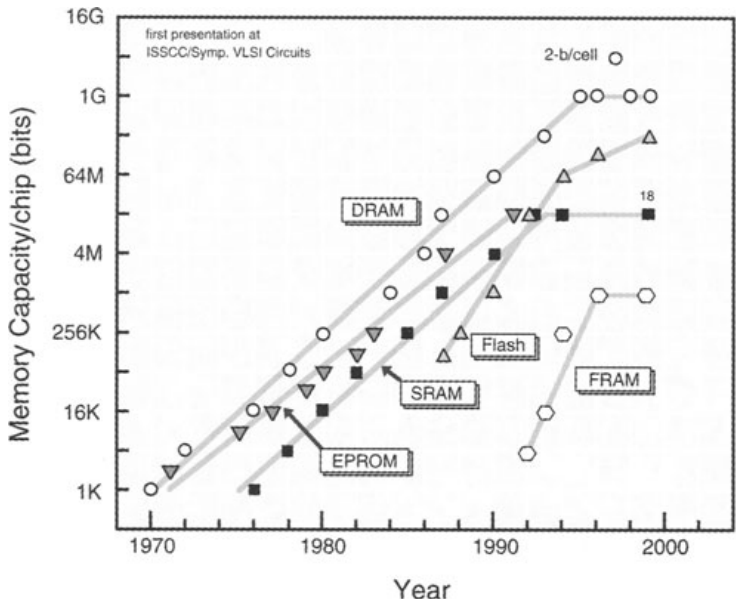


Fig. 1.9. Trends in the memory capacity of VLSI memories [1.9, 1.44]

R&D level [1.9, 1.44]. DRAM has quadrupled its memory capacity every two and a half years, although at the production level it has quadrupled every three years. As a result, standard (commodity or stand-alone) DRAMs have reached 1−4 Gb at the R&D level, and 64−256 Mb at the volume-production level. In addition, the throughputs have been boosted, as exemplified by the 1.6 Gb/s 1 Gb chip [1.45, 1.46], by new memory-subsystem architectures. An 8 Mb embedded DRAM [1.47] using the standard 1 Gb DRAM technology has also revealed an address access time as fast as 3.7 ns with a 1 GHz clock. However, standard DRAM technology tends to be saturated at the 1 Gb generation. Note that the 4Gb [1.3] in the figure employed a risky 2-bit-per-cell scheme that degrades the inherently low signal-to-noise ratio of the 1-T cell. The saturation is caused by both the ever-difficult process of memory-cell miniaturization (as discussed above) and the ever-prevailing chip-shrinking approach, rather than the traditional memory-capacity quadrupling approach, as discussed in Chap. 3. In the early days, the development of SRAM chips was focused on low-power applications, especially with very low standby and data-retention power, while increasing memory capacity with high-density technology. Nowadays, however, more emphasis has been placed on high speed rather than large memory capacity, a trend primarily led by cache applications in high-speed microprocessors. Consequently, on-chip SRAM caches have reached a 0.55 ns, 43 W, 1 Mb BiCMOS macro [1.48] and 1.8−3.4 ns, 1−7 W, 8−18 Mb CMOS macros [1.49], while commodity SRAM has stayed at around 16 Mb since 1992. The market for Flash memories has expanded, due to their exceptional feature of non-volatility, despite small memory-chip capacities of 16 Mb or less. Front-end commercial chips of 64

and 256 Mb [1.50, 1.51] will soon help to expand the market as Flash memories almost catch up with DRAMs in memory capacity, with their inherently small cell structures. The pace of increase in memory capacity of Flash memory will eventually follow the trend in DRAM, due to the limitations of their common lithography technology, although it has skyrocketed since the advent of Flash memory. Non-volatile ferroelectric RAM (FRAM) is still limited to small memory capacities of less than 256 Kb for IC-card applications, although experimental 4 Mb chips have been reported [1.52].

Table 1.3 shows a chip comparison [1.9] between 64 Mb DRAM [1.10], 16 Mb SRAM [1.11], and 64 Mb Flash memory [1.12], all of which use 0.4 μm CMOS technology. With regard to the area ratio of the memory array to the chip, SRAM is the largest because it has the largest memory cell. On the other hand, Flash memory is the smallest due to the necessity for additional control circuits, which is the main reason why it has a large chip area despite having a small memory cell. In read operation, SRAM has the fastest speed, while DRAM and Flash memory are almost the same. In write operation, Flash memory is extremely slow and suffers from a limited number of write cycles. In power dissipation, DRAM is inferior to Flash memory, supposedly due to the large current that is dissipated when many data lines are simultaneously charged and discharged.

1.4 General Trends in DRAM Design and Technology

1.4.1 The History of Memory-Cell Development

The one-transistor, one-capacitor (1-T) cell [1.4] has been universal since the mid-1970s because it offers the highest density. This is a result of tremendous efforts aimed at miniaturization of the memory cell, through reducing the number of components and wirings necessary to construct a cell, as discussed below. The features of the cells proposed so far are characterized by three issues: gain or no gain; destructive or non-destructive readout characteristics; and refresh operation schemes. Here, there are two refresh operation schemes, depending on the number of cells that are simultaneously refreshed by only one refresh operation; all of the cells in a chip or all of the cells along a selected word line.

Figure 1.10 shows flip-flop cells comprising four FETs (4-T cells). Despite the number of components necessary in a cell, both cells develop a quite large readout signal on the pair of data lines (DLs) because of gain cells. The cells enable non-destructive readout characteristics because the stored voltages in the cell are almost maintained even during a read operation. Furthermore, they feature low noise and high speed due to their differential-mode operations. Cell (a) was incorporated in an actual 1 Kb product in the early 1970s. The operations are similar to the SRAM cell operations discussed before. "1" is written by application of a combination of a high voltage (H)

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 11

Table 1.1. Performance comparison of memory chips [1.9]

	Word organization	V_{DD} (V)	Cell area (μm^2)	Chip area (mm^2)	Access/cycle (read) (ns)	Write cycle	Erase (s)	P/E cycles[a]	Current[b] (active/stand-by)(mA)
64 Mb DRAM	4 MW × 8 b	3.3	1.7	211 (0.52)	38/100	100 ns	–	∞	85/0.2
16 Mb SRAM	1 MW × 8 b	3.3	8.4	226 (0.59)	14/14	14 ns	–	∞	70/0.001
64 Mb Flash	4 MW × 16 b	Single[b] 3.3	1.7	257 (0.42)	50/100	6.4 µs	0.8 (chip)	10^4–10^5 (assumed)	30/0.1

[a] P/E, programming erasing cycles.
[b] DRAM, cycle time 100 ns, 8 K refresh cycles; SRAM, cycle time 33 ns; flash, cycle time 100 ns.

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 12

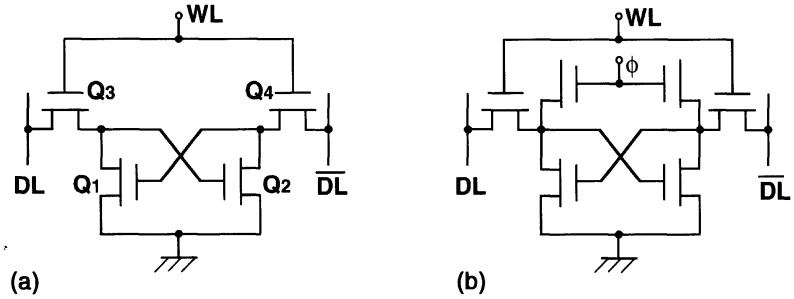


Fig. 1.10. The four-transistor cell [1.9]. (**a**) Standard; (**b**) charge pumping

and a low voltage (L, i.e. 0 V) to the gates of Q_1 and Q_2, respectively, from the data lines when the transfer FETs (Q_3, Q_4) are turned on by activation of the word line (WL). "0" is written by an opposite polarity combination. The read operation is performed only by activation of the word line. As a result, either Q_1 or Q_2 is turned on and discharges the corresponding data line while another data line, is almost quiescent. The polarity of the differential voltage between a pair of data lines corresponds to the read data. Obviously, a high voltage is held at the gate capacitance of the FET for a long time. A degraded high voltage due to leakage currents is refreshed by activation of the word line after a high supply-voltage is applied to both DL and $\overline{\mathrm{DL}}$, so that Q_3 and Q_4 work as the loads of a flip-flop. Thus, a compensation current is supplied through either Q_3 or Q_4. Note that a simultaneous activation of all the word lines enables a simultaneous refreshing of all of the cells in a chip, although it involves an excessive spike current. It is also possible to refresh all of the cells on each of the word lines in order. Cell (b) [1.13] features additional charge-pumping capacitances which work as refresh-function devices. Charges to compensate for leakage charges are given by cyclic clock (ϕ) drivings of the capacitances. All of the cells in a chip are simultaneously refreshed, but the charge-pumping device is difficult to design.

Figure 1.11 shows various three-transistor (3-T) cells. They are categorized by the number of lines necessary to compose a cell. Here, a ground wiring, which is shared with an adjacent cell, is counted as a half in number. These cells possess a gain that gives a large signal voltage on the data line, non-destructive readout characteristics, and a refresh operation that is performed for each word line. Cell (a) was used in a 1 Kb product in the early 1970s. The write operation is achieved by applying the data voltage (H or L) from the write data line (WD) to the Q_2 gate through Q_1 after activating the write word line (WW). The information is read out on the read data line (RD) by activating the read word line (RW). If the Q_2 gate voltage is high, RD that has been precharged to a high level is discharged to 0 V because both Q_2 and Q_3 are on. If the Q_2 gate voltage is low, RD is maintained at the high level because Q_2 is off. The stored data is discriminated by detecting the resulting RD voltage. The Q_2 gate capacitance is responsible for data retention. A refresh operation needs a succession of read and rewrite

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 13

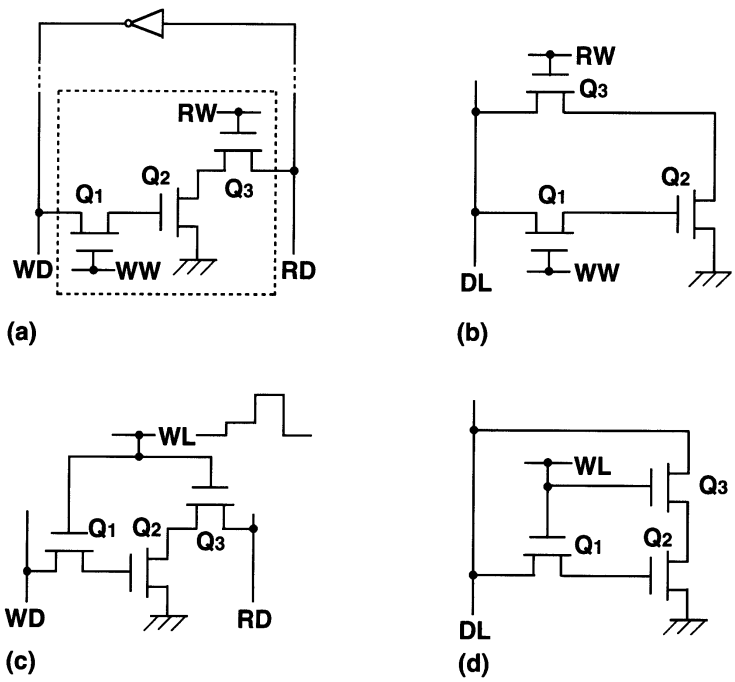


Fig. 1.11. The three-transistor cell [1.9]. (**a**) 4.5 lines; (**b, c**) 3.5 lines; (**d**) 2.5 lines

operations, with the help of an inverter at the end of the RD. The operation is successfully performed as long as a degraded Q_2 gate voltage can discharge the RD. The resultant low RD voltage is inverted to a sufficiently high voltage by the inverter so that a high enough voltage is rewritten to the Q_2 gate. Cell (b) [1.14] was also used in a 4 Kb product in which the read and write data lines of cell (a) were merged into one data line. The read or refresh operation differs from that of cell (a). When reading, an inverted voltage of the Q_2 gate voltage is developed on the data line, which has been precharged to a high voltage. The resultant data-line voltage is directly rewritten into the Q_2 gate, making the Q_2 gate voltage switch from high to low or from low to high at every read operation. Thus, a monitor is connected at the end of each word line. Exclusive OR of the output of the monitor and the data-line voltage makes the read and write operations logically consistent. Cells (c) [1.15] and (d) [1.16] merge the separated word lines (RW and WW) of cells (a) and (b) in one word line (WL), respectively, in order to reduce the cell area. The read operation is performed with an intermediate WL voltage to make Q_3 turn on but to make Q_1 turn off. A succeeding sufficiently high WL voltage enables the rewrite operation. Here, the intermediate level must be controlled precisely so as to be between V_T and $2V_T$ (V_T is the threshold voltage of the FET). When reading the low-level data at the Q_2 gate, the Q_2 gate can be raised by a current through Q_1 from the data line, that stays at a high level after precharging. Even in this case, the resultant Q_2 gate voltage must be below V_T to ensure the low level, thus the WL voltage must be below $2V_T$.

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 14

On the other hand, the WL voltage must exceed V_T to turn Q_3 on. Another drawback of cell (d) is that a large dc current flows from the data line to ground, since both Q_2 and Q_3 are turned on when a high level is written.

Figures 1.12 and 1.13 show the two-transistor, two-capacitor (2-T or twin) cell and the one-transistor, one-capacitor (1-T) cell, respectively. Neither cell has any gain. Each transistor works as a switch, and each capacitor stores the information charge. The read operation of both cells, which is preceded by a precharge operation that leaves the data line at a high floating voltage, has the following two features. The cell signal voltage developed on the data line is very small, and thus a large signal-voltage component that was stored at the cell node is destroyed by the read operation (the so-called destructive readout characteristics), as shown later. Thus, to successfully amplify the small cell signal and rewrite (restore) the cell with the amplified signal, a sense amplifier and a rewrite circuit are connected to the data line. The refresh operation is almost same as the above read–rewrite operation. The 2-T cell features a stable operation because of a completely differential operation, as described previously, despite having a larger cell area.

1.4.2 The Basic Operation of The 1-T Cell

Figure 1.14 shows a conceptual 1-T cell array of n rows by m columns, and an actual data-line configuration. Plural memory cells, a precharge circuit and equalizer (PE), and a latch-type CMOS sense amplifier (SA) are connected to each pair of data lines (DLs) which communicate with a pair of common data input/output lines (I/O and $\overline{\text{I/O}}$) through a column switch. The 1-T cell operation comprises read, write, and refresh operations. All operations entail common operations: precharging (i.e. initializing) all pairs of data lines to a floating voltage of a half V_{DD} by turning off the precharge circuit and equalizer, and then activating a selected word line.

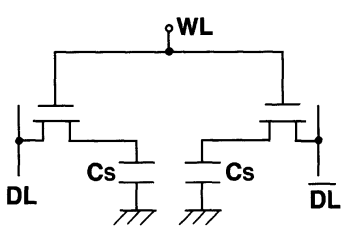


Fig. 1.12. The two-transistor cell (twin cell)

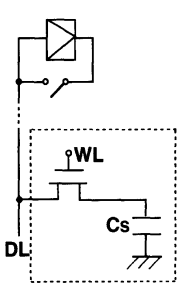


Fig. 1.13. The one-transistor cell

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 15

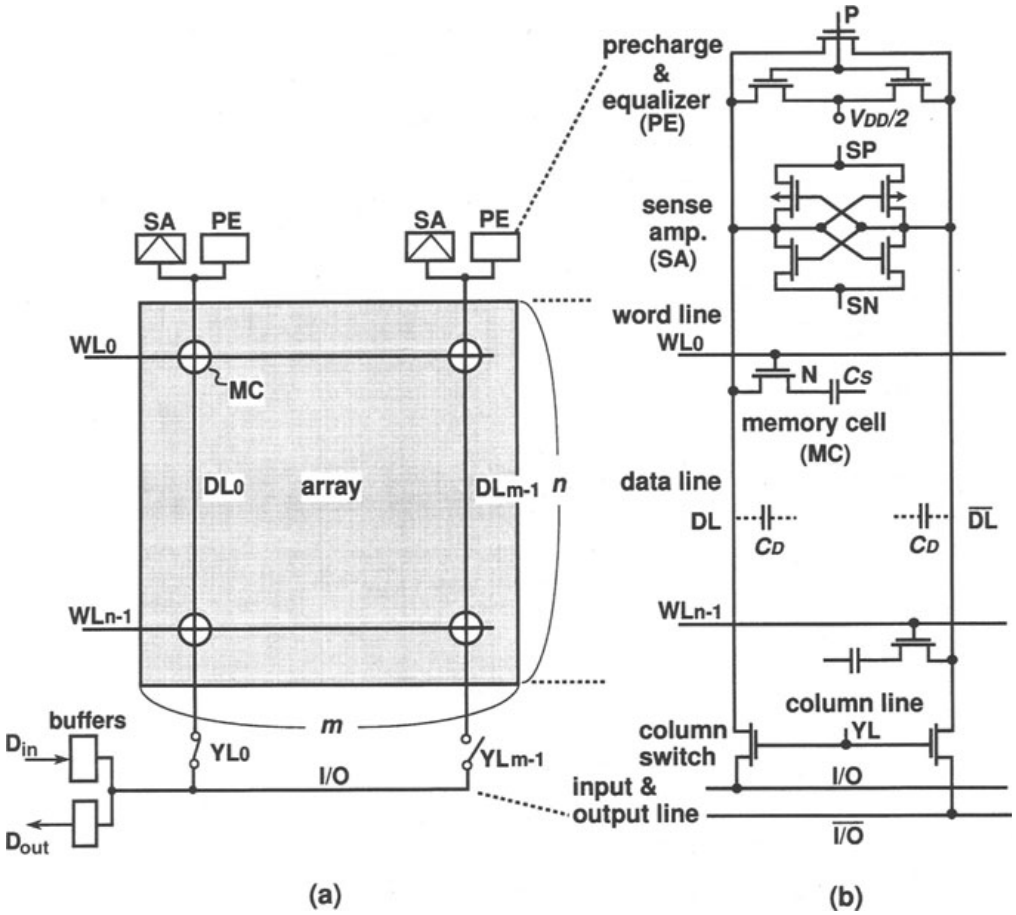


Fig. 1.14. A conceptual DRAM array (**a**), and an actual data-line configuration (**b**)

In the read operation (Fig. 1.15), a stored data voltage, V_{DD} ("1") or 0 V ("0"), at the cell node (N) of each cell along the word line is read out on the corresponding data line. As a result of charge sharing, the signal voltage ($\pm\nu_{\mathrm{S}}$) developed on the floating data line (for example, DL) is expressed by

$$\nu_{\mathrm{S}} = \frac{V_{\mathrm{DD}}}{2} \cdot \frac{C_{\mathrm{S}}}{C_{\mathrm{D}} + C_{\mathrm{S}}} \ .$$

Unfortunately, ν_{S} is inherently small (100−200 mV) because the data-line parasitic capacitance (C_{D}) is much larger than the cell storage capacitance (C_{S}). A small C_{S} and a large C_{D} result from the need for a small cell area and for connecting a large number of cells to a data line, respectively. Hence the original large signal component ($V_{\mathrm{DD}}/2$, usually 1−2.5 V) at the storage node collapses to ν_{S}. The destructive readout characteristics necessitate successive amplification and restoration for each of cells along the word line. This is performed by a latch-type differential CMOS sense amplifier on each data line, with the other data line ($\overline{\mathrm{DL}}$) as a reference. Then, one of the amplified

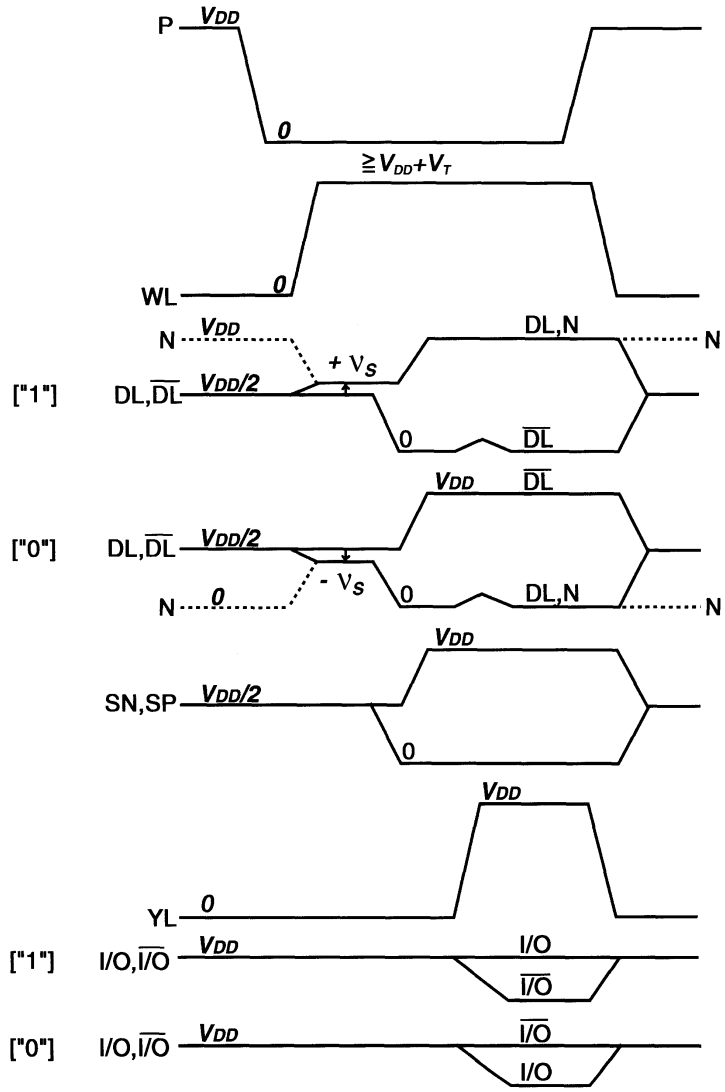


Fig. 1.15. The read operation

signals is outputted as a differential voltage to the I/O lines by activating a selected column line, YL.

The write operation (Fig. 1.16) is always accompanied by a preceding read operation. After almost completing the above amplification, a set of differential data-in voltages of V_{DD} and 0 V is inputted from the I/O lines to the selected pair of data lines. Hence, the old cell data are replaced by the new data. Note that the above read operation (i.e. amplification and restoration) is done simultaneously for each of the remaining cells on the selected word line to avoid loss of information.

The stored voltage of each cell degraded by the leakage current is restored by a refresh operation that is almost the same as for the read operation, except that all YLs are kept inactive. This is done by reading the data of cells on the word line and restoring them for each word line so that all of

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 17

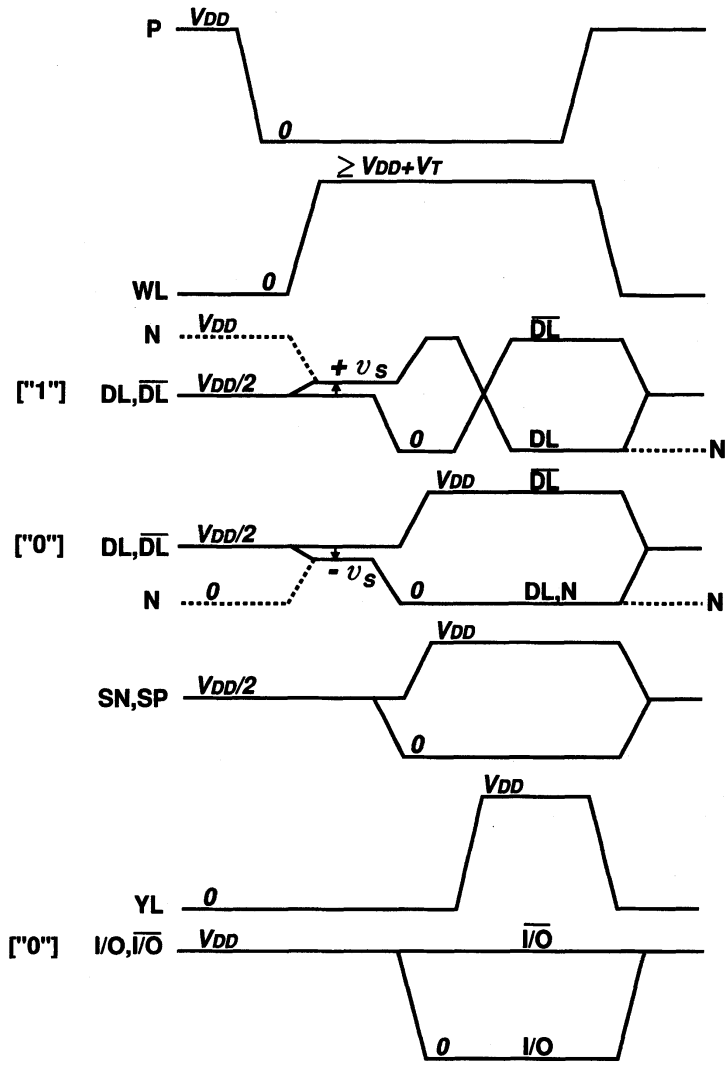


Fig. 1.16. The write operation for low-voltage ("0") binary information

the cells retain the data for at least t_{REFmax}. Here t_{REFmax} is the maximum refresh time for the cell, which is guaranteed in catalog specifications, and is exemplified by $t_{\mathrm{REFmax}} = 64\,\mathrm{ms}$ for a 64 Mb chip. Thus, each cell is periodically refreshed at intervals of t_{REFmax}, as shown in Fig. 1.17, although each cell actually has a data-retention time longer than t_{REFmax}.

The fundamental circuit design issues of the 1-T cell can be summarized as the signal-to-noise ratio (S/N), power dissipation, and speed, because of the following inherent 1-T cell characteristics:

1. A small read signal, and relatively large levels of noise. Thus, the read operation is unstable unless a high S/N design is achieved. A small read signal is caused by the cell having no gain. There are many noise sources during the read operation. The difficulty in accommodating the sense amplifier and precharge circuit within a small layout pitch of data lines

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 18

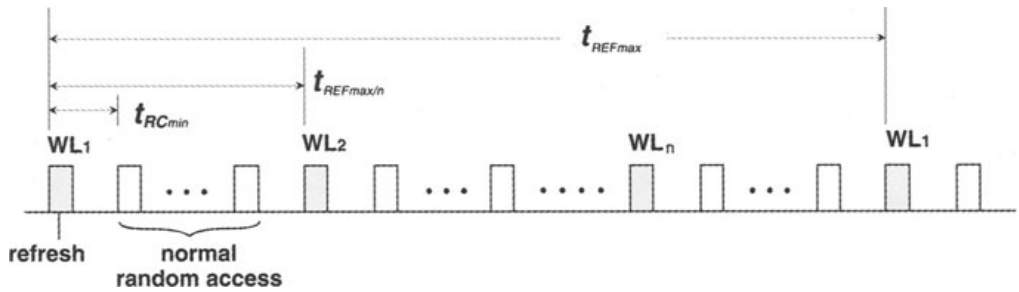


Fig. 1.17. The refresh operation. t_{RCmin}, minimum cycle time

tends to generate electrical imbalances to a pair of data lines. A large number of sense amplifiers results in a large deviation of the threshold-voltage mismatch (the offset voltage) between pairs of MOSFETs in the sense amplifier. Simultaneous charging and discharging of many heavily capacitive data lines with a high voltage invariably introduces many kinds of noise. In addition, cell leakage currents and α-particle hitting which degrade the stored charges (as discussed in Chap. 4) effectively work as noises.

2. The above charging and discharging of data lines also causes a high power dissipation.
3. Slow sense-amplifier operation. The relatively poor driving capability of the sense amplifier, which stems from the need for a small area, and operation based on a low voltage of a half-V_{DD}, makes the sense amplifier operation slow. Thus, the sensing time is the largest component of the access time for the chip.

1.4.3 Advances in DRAM Design and Technology

The density of DRAMs has quadrupled every three years since their advent in the early 1970s. The development of this higher density has made them cheaper than other types of RAM. Low power dissipation and a fast access time have been obtained even with an increase in chip area for every successive generation, as shown in Fig. 3.1. This has also contributed to the advantages of DRAMs. As a result, 64 Mb DRAMs have now reached maturity in production, and are being followed by the development of 256 Mb and 1 Gb DRAMs. The resultant DRAM technology has been applied to other successful products, such as video RAMs, ASIC memories, and the merged logic and DRAM LSIs. Even 1.5 V or less DRAMs for battery-based applications have been proposed. Their rapid evolution [1.17–1.24] has been achieved mainly because of progress in the 1-T cell and its supporting circuit technologies, as well as in photolithography. The 1-T cell still surpasses gain cells such as the 3-T and 4-T cells, which were common to the 1 and 4 Kb generations, with respect to chip area. A smaller chip area and the resulting

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 19

lower cost due to the 1-T cell justify the complexity of the fabrication process in forming the cell capacitor.

Table 1.2 shows circuit innovations [1.17] that have contributed to progress in DRAM development. With regard to a high S/N design, word voltage bootstrapping, a shared amplifier, and multidivided data lines are essential to increase the cell signal voltage. Differential sensing, a folded data-line arrangement, and transposed data lines are representative techniques for reducing noise. To reduce power dissipation, in addition to reducing the external power supply V_{DD}, dynamic amplifiers, dynamic drivers, CMOS circuits, half-V_{DD} data-line precharge, a shared I/O combined with multidivided data lines, and on-chip voltage down-converters have played important roles. To increase speed, multidivision of poly-Si word lines shunted by aluminum wiring and small signal chip-to-chip interfaces, as well as multilevel metal wiring and device scaling, are important. They contribute to reducing RC delay in a chip. There is no doubt that, in addition to the refreshing scheme, address multiplex, chip coatings against soft error, a substrate bias (V_{BB}) generator to realize a single power supply, and redundancy circuits are also important. Besides, the testing issue is becoming increasingly important with increasing memory capacity. On-chip multibit (parallel) testing, which enables parallel testing for multiple data read from subarrays, plays an important role in drastically reducing the testing time. Moreover, ECC circuits and an on-chip microcoded test pattern generator are expected to be promising circuits.

Examples of the Advances. Advances in high-density and large memory-capacity technology are explained in detail here. Table 1.3 shows a comparison [1.9] of technology between a 64 Kb chip [1.25, 1.26], presented in 1980, and a 64 Mb chip [1.27, 1.28], presented in 1990. Although the 64 Kb chip was in the early stages of production while the 64 Mb chip was still at the research level, the memory capacity increased one thousand-fold over about 10 years. The MOSFET and the memory cell were reduced in size by about one-tenth. Figures 1.18 and 1.19 show comparisons of memory cell structure and chip configuration. Figures 1.20 and 1.21 show microphotographs of a cross-section of the memory cell, different facets from that in Fig. 1.18, and the 64 Mb chip. Note the use of multilevel wirings, vertical structures, and new materials. A low-resistance metal (i.e. the first-level Al) line straps a highly resistive poly-Si word line for reducing the line delay, which becomes more prominent as the chip area increases. Hence, a through-hole (or contact) with a large aspect ratio is needed to connect the lower poly-Si line to the upper metal line, with a small pitch on the hilly surface. Another metal line (i.e. the second-level Al) running along each pair of data lines works as a column selection line. It confines the increase in chip area, which is caused by the use of a multidivided data-line structure, and reduces the number of bundles of column decoders, as discussed in Chap. 3. There are many possible improvements to obtain a larger cell capacitance, C_S. The capacitor

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 20

Table 1.2. Circuit innovations for DRAM [1.9, 1.17]

Capacity	Power supply (V)	Memory cell	Circuits
1–4 Kb	> 12	3-T	–Differential sensing –Adress multiplexing
16 Kb	12	1-T	–Dynamic amplifier –Dynamic driver
64 Kb	5	1-T	–Folded data (bit) line –Word bootstrapping –Substrate bias generator
256 Kb	5	1-T	–Shared amplifier –Metal-strapped poly-Si word line –Redundancy
1 Mb	5	1-T	–CMOS peripheral circuits –Half-V_{DD} precharge –Multidivided data line –Shared I/O –On-chip parallel test
4 Mb	5	1-T (3-D capacitor)	
16 Mb	5	1-T	–On-chip voltage down-converter –Transposed data line
64 Mb	3.3	1-T	–Multidivided word line –Small signal interface

features a cylindrical stacked capacitor formed on the data (bit) line, the so-called COB (Capacitor Over Bit) line. The COB maximizes C_S, and reduces the interference noise from adjacent data lines a resultant shielded data-line structure. Moreover, a high-permittivity (ε) Ta_2O_5 film is used to increase C_S. Instead, it necessitates a tungsten (W) electrode that is an oxidation-immune material. A poly-Si electrode would be oxidized, and the resulting thin SiO_2 film, with a lower permittivity, that would be formed between the poly-Si electrode and the Ta_2O_5 film would destroy the high-ε benefit provided by Ta_2O_5.

Regarding circuit technology, the number of array divisions has increased remarkably, from eight to 256. Each data line is divided into 64 for the 64 Mb while it is divided into 2 for the 64 Kb. The increased number of division is necessary to increase the cell signal voltage with a reduction in C_D, and to reduce the data-line power dissipation through partial activation of the resultant subdata lines. The increased number of word-line divisions reduces the line delay even for the Al-strapped poly-Si word line. A large chip area, increased by a factor of eight, and a large memory capacity need a redundancy technique to improve yield, and a built-in parallel test circuit to cope with

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 21

Table 1.3. Advances in DRAM chip technology [1.9]

		64 Kbit	64 Mbit
Process		4in wafer 3 μm NMOS Double poly-Si, single metal	8in wafer 0.3 μm triple-well CMOS Double poly-Si, single polycide, W electrode, triple metal
MOSFET (L/t_{OX})		2 μm/50 nm	0.35 μm/6.5 nm
Memory cell	Structure	144 μm^2, planar C_S V_{DD} electrode	1.3 μm^2, stacked C_S $V_{DD}/2$ electrode
	C_S/C_D	38 fF/515 fF	44 fF/250 fF, shielded DL
Circuit		– – V_{DD} precharge Two-divided DL	Redundancy Multibit test (parallel test) $V_{DD}/2$ precharge 64-divided DL
Arrangement of peripheral circuit		Two edges of chip	Cross-shaped area in chip
Chip	Area	26 mm^2 (3.43×7.52)	198 mm^2 (9.74×20.28)
	V_{DD}	5 V	1.5 V or 3.3 V (down-converter)
	Access time	90 ns	50 ns
	Power	220 mW (cycle 230 ns)	44 mW (cycle 180 ns, 1.5 V)
	Organization	64 KW × 1 b	16 MW × 4 b
	Refresh	2 ms/128 cycles	64 ms/8192 cycles

the ever-increasing testing time. The location of the peripheral circuit has changed from the two outer edges of the 64 Kb chip to the cross-shaped layout area at the center of the 64 Mb chip. This is due to equalizing the delay from the peripheral circuit to each divided array as much as possible, which is especially important for a larger chip. Note that a higher speed and a lower power dissipation are realized even for a memory capacity as large as 64 Mb. The 64 Mb chip works even at 1.5 V, although it also works at 3.3 V, combined with an internal power supply of 1.5 V that is generated from an on-chip voltage down-converter. With regard to the refresh operation, all of the cells in the 64 Kb chip can be refreshed by 128 refresh cycles for a 2 ms refresh time (t_{REFmax}). As for the 64 Mb chip, however, 8192 refresh cycles for 64 ms t_{REFmax} are needed to reduce the data-line power dissipation, as

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 22

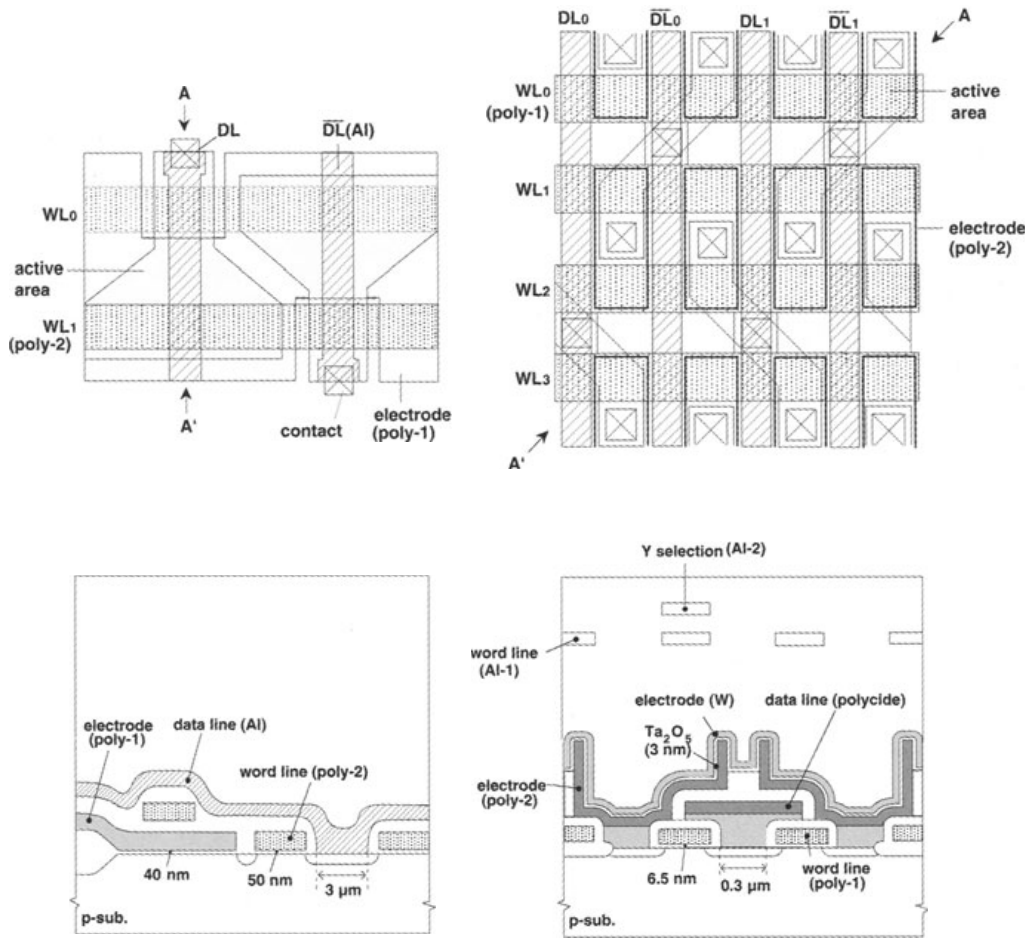


Fig. 1.18. A comparison between a 64 Kb memory cell (*left*) and a 64 Mb memory cell (*right*) [1.9]. *Top*, layout; *bottom*, cross-section A-A′. The word line Al-1, the Y-selection line Al-2, and the electrode W for the 64 Mb cell are omitted for ease of understanding

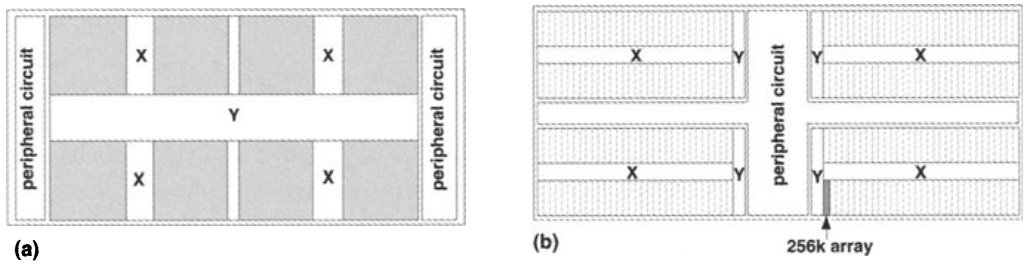


Fig. 1.19. A comparison between the configurations of a 64 Kb chip (**a**) and a 64 Mb chip (**b**) [1.9]

discussed in Chap. 3. Hence, a 32-fold reduction of the cell leakage current is needed to extend t_{REFmax} from 2 ms to 64 ms.

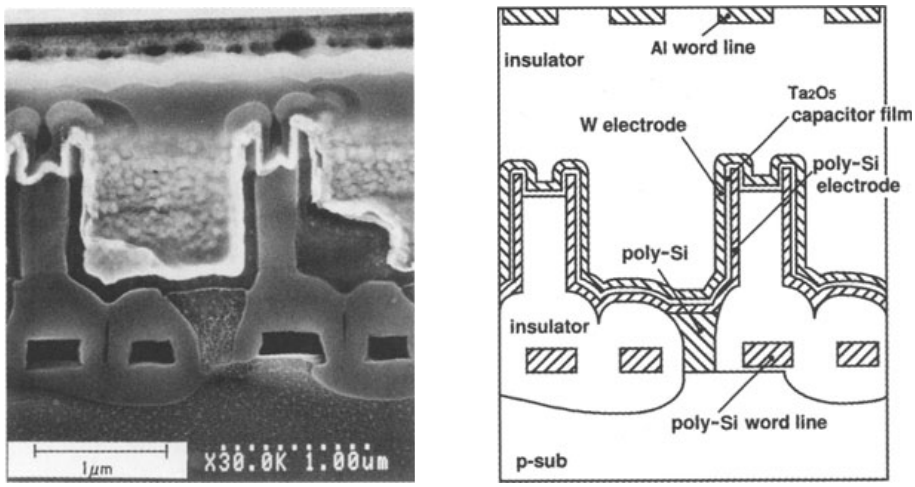


Fig. 1.20. An actual cross-section of a 64 Mb memory cell [1.9]

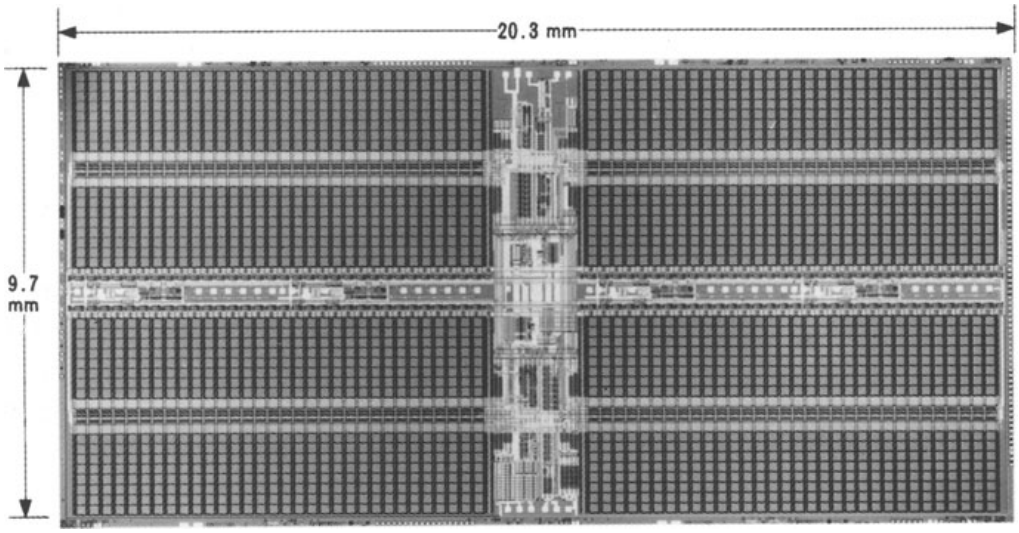


Fig. 1.21. A microphotograph of the 64 Mb DRAM chip in Fig. 1.19 [1.28]

1.5 General Trends in SRAM Design and Technology

1.5.1 The History of Memory-Cell Development

An SRAM cell consists of two inverters, and two transfer MOSFETs that are connected to a pair of data (or bit) lines, as shown in Fig. 1.22. To achieve a flip-flop, the input and output of one inverter are connected to the output and input of the other inverter, respectively. The SRAM cell has enabled high performance through circuit and structural improvements of the inverter loads [1.18, 1.29, 1.30].

Cells (a) and (b) are MOSFET-load cells, featuring the use of enhancement MOSFETs for cell (a) and depletion MOSFETs for cell (b). Cell (a)

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 24

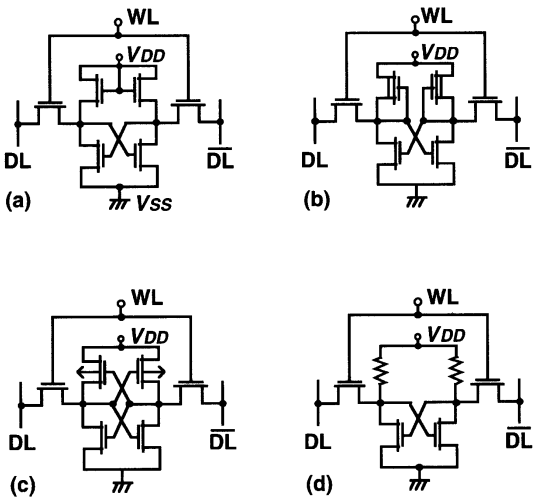


Fig. 1.22. Flip-flop SRAM memory cells [1.9]. (**a**) Enhancement MOS load; (**b**) depletion MOS load; (**c**) PMOS load (full CMOS); (**d**) polysilicon load

suffers from a narrow voltage margin, a slow speed, and a large area, which come from V_T drop of the enhancement MOSFET and a ratio-circuit operation, as discussed in Chap. 2. A dc current flow, which is a data-retention current, at either of the inverters is an additional problem. This causes an obstacle to increasing the memory capacity, because of the ever-increasing power dissipation. Cell (b) partly solves the above problems with a full swing of V_{DD} at the inverter output node, although the dc current still remains. Cell (c), called the full CMOS cell, inherits the advantages of the CMOS inverter that cuts the dc current completely and has a wide voltage margin. However, the cell area is quite large unless a tight isolation technology between NMOS and PMOS is available. Cell (d) is called a poly-silicon load cell, and is suitable for higher-density SRAMs. Poly-silicon (poly-Si) facilitates an extremely high resistance load with a small area, thus allowing a high density and a small dc current. It has brought about an advance in cell miniaturization by the use of a vertical structure with double poly-silicon technology; that is, the first poly-silicon for the MOSFET gates, and the second poly-silicon, which overlays the first poly-silicon, for the loads. For 4-Mb or higher density SRAMs, however, the poly-silicon load cell starts to be replaced by a TFT (Thin Film Transistor) load cell since polysilicon is unable to keep the data-retention current low enough. The TFT load cell is a kind of full CMOS cell, in which PMOSFETs are built into the second poly-silicon layer.

Recently, a full CMOS cell (c) has become increasingly important for use in cache memories [1.30]. This is because it offers a stable operation – that is, a wide voltage margin and noise immunity even for high-speed operation – and ease of fabrication, with a process compatible with CMOS logic in microprocessors. Although a cache SRAM does not need a density as high as that of a DRAM, the memory cell area is still a prime concern to

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 25

reduce the fabrication cost, especially for microprocessors that have a denser on-chip cache. Denser cells have been realized that uses the following technologies [1.30]: two-level local interconnection to form a flip-flop, source–drain silicidation, trench isolation fabricated by a chemical mechanical polishing (CMP) process, and thin TiN local interconnection for fine patterning.

1.5.2 The Basic Operation of a SRAM Cell

Figure 1.23 shows a conceptual full CMOS cell array. The read operation starts with activation of a selected word line after equalizing all pairs of data lines to around V_{DD}, as shown in Fig. 1.24. Each cell along the selected word line develops a small signal voltage, ν_S, on one of the data lines, depending on the stored cell information. For example, for "1" stored information, where cell node N_1 is at a low voltage and cell node N_2 is at a high voltage, a small signal voltage is developed on the data line (DL) as a result of the ratio of a data line load (Z), a transfer MOSFET (Q_5), and a driver MOSFET (Q_1). Another data line ($\overline{DL}$) remains at the equalized voltage with both Q_2 and Q_4 cut off, as long as the resultantly raised N_1 voltage is smaller than the Q_2 threshold voltage (V_T). A read data from one pair of data lines is transferred to a pair of I/O lines by turning on the column switch selected by a column

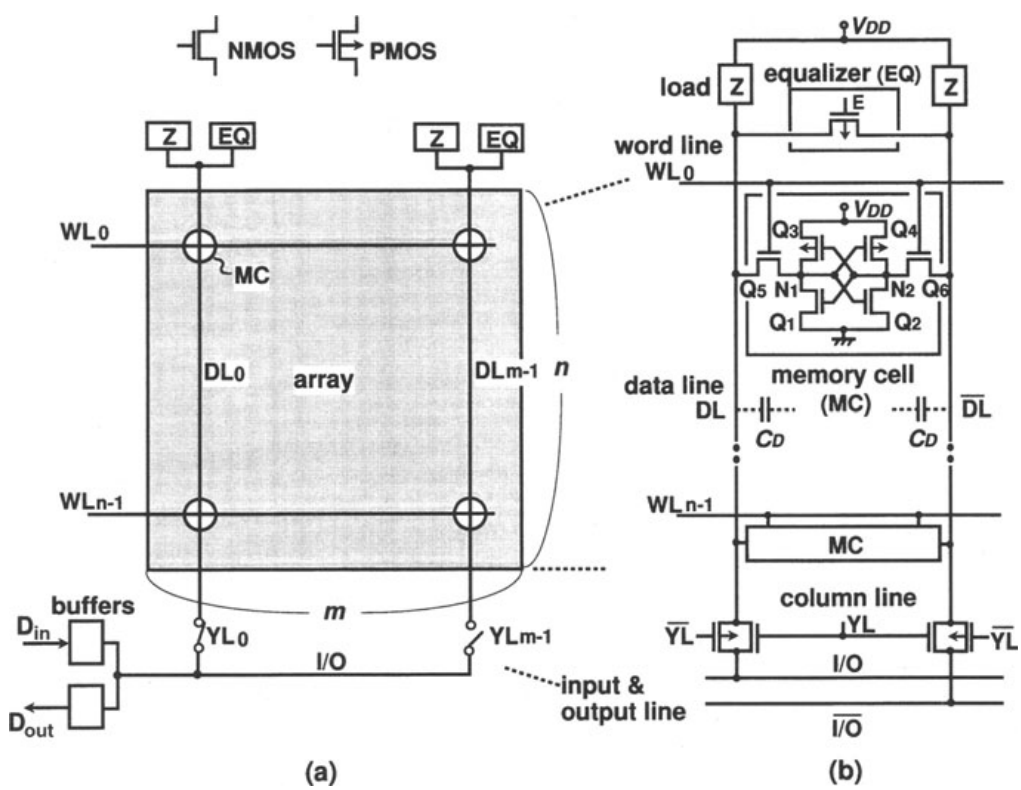


Fig. 1.23. A conceptual SRAM array (**a**) and an actual data-line configuration (**b**)

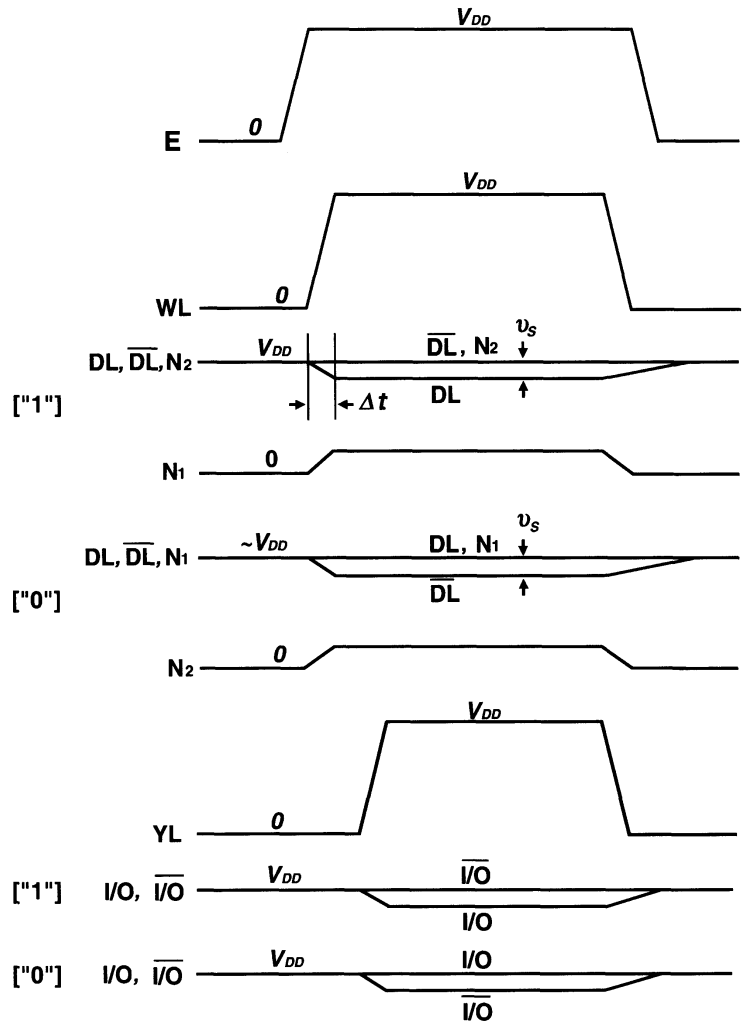


Fig. 1.24. The read operation

line (YL) activation. The polarity of the differential signal voltage is detected by a differential amplifier in a data output buffer, so as to be outputted as the data output (D_{out}) from the chip. For "0" stored information, the same operation is carried out, except that the opposite voltage polarity is applied.

The write operation is carried out by applying a differential input, whose polarity corresponds to the write binary information, to the I/O lines through the data input buffer, as shown in Fig. 1.25. The full differential voltage is quickly transferred to the data line and then to the cell nodes (N_1, N_2) if the conductance of the column switch MOSFETs is sufficiently higher than that of the data-line loads. That is why each column switch is composed of parallel-connected N- and PMOSFETs. An NMOSFET can fully discharge the data line with its high gate voltage, while a PMOSFET can fully charge up the data line with its low gate voltage. The write operation is completed by turning off the word line so that the differential voltage is held as stored

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 27

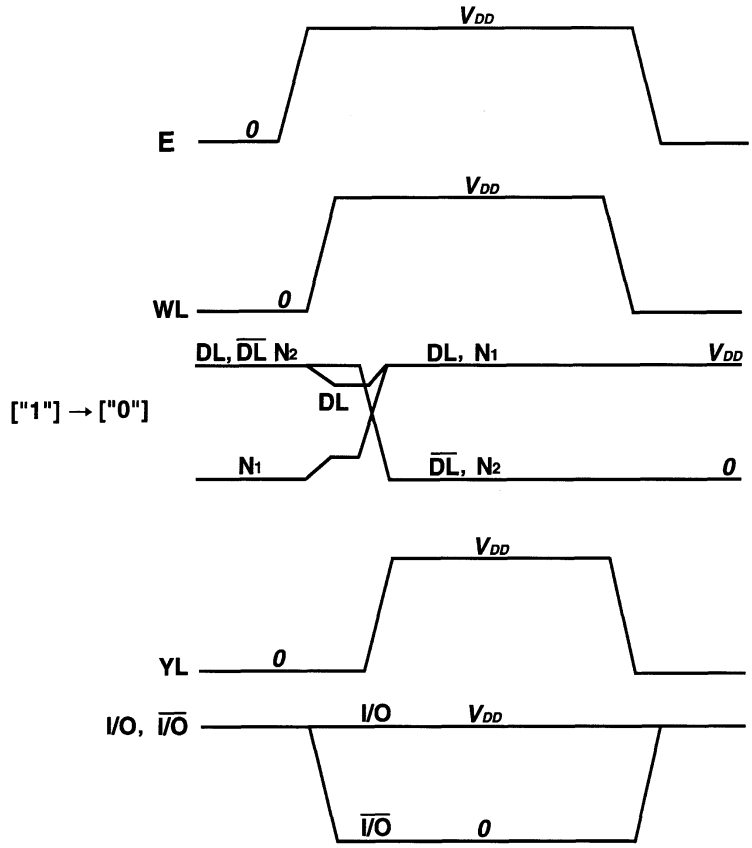


Fig. 1.25. The write operation

information. Note that each of the remaining cells on the selected word line continues to develop a small signal voltage on the corresponding data line, in the same manner as in the read operation.

The SRAM cell has the following features which differ from those of the DRAM cell:

1. The small signal voltage (ν_S) developed by the read operation offers a fast access time. The resultant ratio current (i) discharges the data-line capacitance (C_D) according to the following equation:

   $$i = C_D(\nu_S/\Delta t) .$$

   Hence, the discharge time (Δt) is shortened by reducing ν_S. Here, to achieve non-destructive readout characteristics the Q_5 conductance is set to be the smallest: it is 1.5−2 times smaller than the Q_1 conductance, which is much smaller than the Z conductance. Thus, i is eventually governed by the cell transfer FET, and ν_S is almost iZ.
2. Non-destructive readout characteristics allow the elimination of a sense amplifier on each pair of data lines. A possibly degraded voltage difference between the two nodes (N_1, N_2), that is developed during the read operation, finally recovers to a full V_{DD} after the word-line voltage is

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 28

turned off, with the help of the cells feedback loop. Hence, a SRAM cell does not need a rewrite operation. This is the origin of the faster cycle time compared with a DRAM cell.

3. Instead, a ratio current continues to flow during word-line activation for each of the cells along the word line. This causes a high power dissipation. In addition, a small ν_S needs a highly sensitive differential amplifier on the I/O line.

1.5.3 Advances in SRAM Design and Technology

Circuit innovations [1.18, 1.21, 1.29, 1.30] in the development of SRAMs are summarized in Table 1.4. For the peripheral circuit of SRAM chips, a PMOS

Table 1.4. Circuit innovations for MOS SRAM [1.9]

Capacity	Power supply (V)	Memory cell[a]	Circuits
256 bit–1 Kb	5	E-load, D-load, full CMOS	–PMOS –NMOS
4 K	5	Polysilicon load	–CMOS periphery –CMOS differential sense
16 Kb	5		–Address transition detection –Data-line equilibration
64 Kb	5		–Double-ended full differential sense –Multidivided word line –BiCMOS periphery
256 Kb	5		–Redundancy –Multistage sense amplifier
1 Mb	5		–I/O line equilibration –I/O line sense amplifier
4 Mb	3.3–5	TFT load	–On-chip voltage down-converter –Positive feedback sense –Synchronous/pipeline

[a] E, enhancement; D, depletion.

circuit was first used, and then quickly followed by an NMOS circuit to improve the speed. After that, a CMOS peripheral circuit combined with a poly-silicon load NMOS memory cell was introduced in the 4 Kb generation, to reduce the power dissipation. A typical example is a CMOS differential amplifier with a PMOS current mirror. In the 16 Kb generation, address transition detection (ATD), which has been used hereafter for SRAM and DRAM, was developed to generate short pulses at every address-signal transition. This enables high-speed and low-power dynamic operations of specific peripheral circuits on which static operations have been imposed, despite the CMOS circuit era. For example, it allows each pair of data lines to be quickly equalized for the next activation, and it cuts the cell ratio current, with a narrowed word pulse. Moreover, it achieves automatic power-down of the differential amplifiers on the I/O lines. In the 64 Kb generation, the double-ended full differential CMOS sensing technique, partial activation of the multidivided word line (i.e. a hierarchical word line) that reduces the ratio current while reducing the number of activated cells, and high-speed BiCMOS circuits were developed. In addition to improvements in amplifier and equilibration techniques to increase the circuit speed, redundancy and pipeline architectures were proposed in the 256 Kb generation and beyond.

Example of Advances. Figure 1.26 shows the structure of a high-speed and high-density full CMOS cell [1.31, 1.32]. A 0.25 μm CMOS process featuring two-level poly-Si, one-level TiN local interconnection, and a three-level metal process provides a 6-T cell in $2.6 \times 3.15\,\mu m^2$. Poly-2 power lines (V_{DD} and V_{SS}) are strapped at every 64 cells along the word line with metal 1, and each data

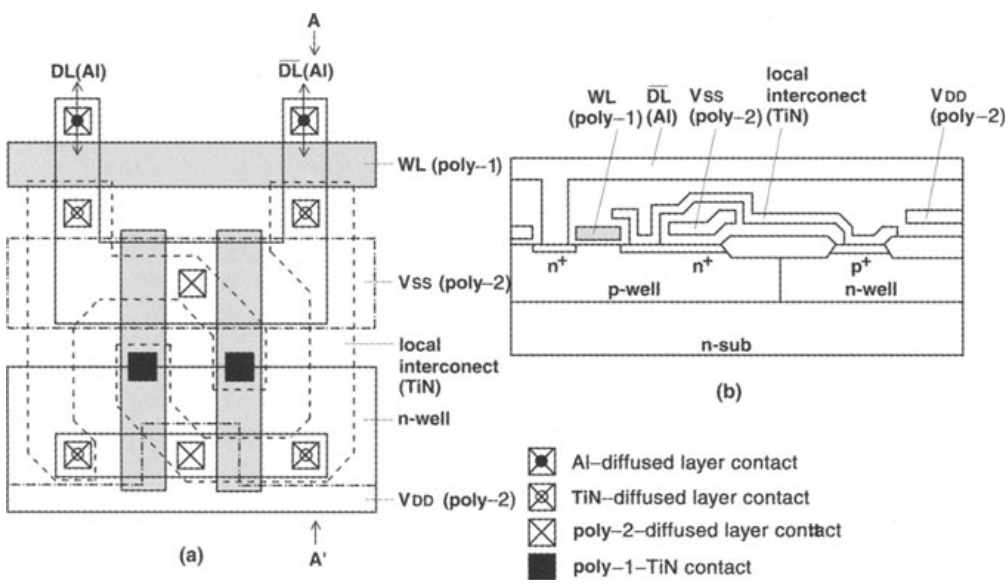


Fig. 1.26. A schematic of a full-CMOS SRAM-cell layout (**a**) and a cross-section along line A-A′ (**b**) [1.31]. The Al wirings for data lines in (**a**) are omitted for ease of understanding

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 30

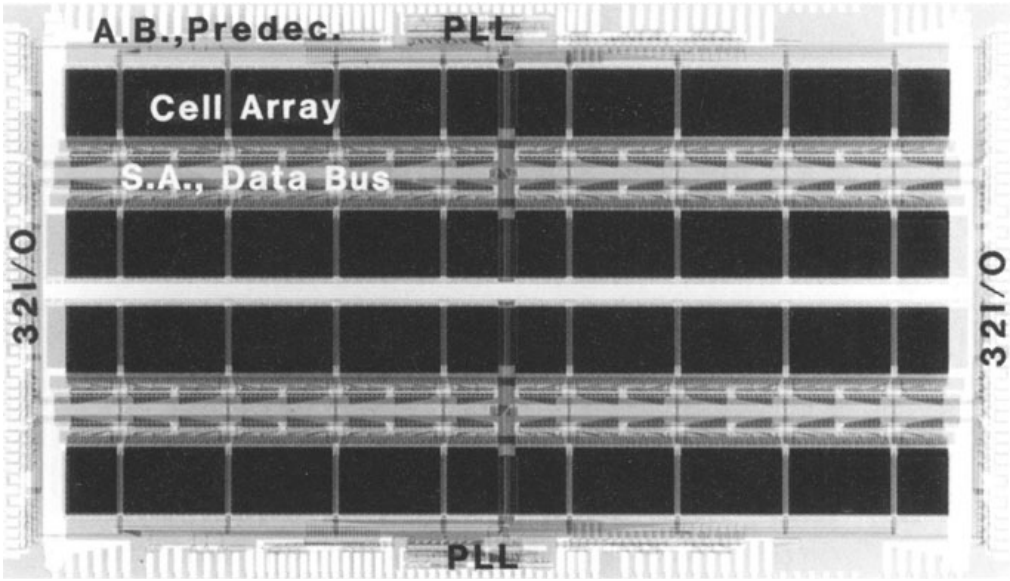


Fig. 1.27. A microphotograph of a 4 Mb CMOS SRAM [1.32]

line is composed of metal 2. To construct a hierarchical word line, a metal 3 main word line runs for every four poly-1 subword lines. Here, these metal (Al) lines are omitted for ease of understanding in the figure. P- and NMOSFETs use a gate length of 0.25 μm and a gate oxide thickness of 6 nm. A double well adjusts the values of V_T for both the N- and PMOSFETs. Figure 1.27 shows the chip micrograph for a 4 Mb CMOS SRAM [1.32], using the above full CMOS cell. The chip measures $8 \times 14.1\,\mathrm{mm}^2$, and it is divided into four quadrants, each of which is composed of 16 small blocks and thus 16 I/O pads. Each block has cells in a 128×512 row-and-column format. An access time of 9 ns and a 300 MHz data-rate read operation can be achieved at 2.5 V V_{DD}. A pipeline scheme, discussed in Chap. 6, is responsible for such a high data rate.

1.6 General Trends in Non-Volatile Memory Design and Technology

1.6.1 The History of Memory-Cell Development

Table 1.5 summarizes various non-volatile memories [1.9]. The mask ROM cell is the smallest if a NAND structure (i.e. series-connected cells) is employed. The EPROM cell is the second smallest, while the EEPROM cell is the largest because it needs two transistors. Except for Mask ROM, non-volatile memories suffer the drawback of a limited endurance; that is, a limited number of programming and erasing cycles, of $10^2 - 10^5$. Note that a write

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 31

Table 1.5. Comparision between non-volatile memories [1.9]

	Cell[a]	Cell area (ratio)	Mechanism		External power supply[b]		P/E cycles[c]	Program on board
			Erase	Write	Write	Read		
Mask ROM	1-T (NAND)	0.35–0.5	–	–	–	V_{DD}	0	–
EPROM	1-T	1	Ultraviolet exposure	Hot electrons	V_{PP}	V_{DD}	~ 100	Impossible
EEPROM	2-T	3–5	Tunnel current	Tunnel current	V_{PP} (int.)	V_{DD}	10^4–10^5	Possible
Flash-memory	1-T	1–2	Tunnel current	Hot electrons	V_{PP}	V_{DD}	10^4–10^5	Possible
			Tunnel current	Tunnel current	V_{PP} (int.)	V_{DD}	10^4–10^5	Possible

[a] T, transistor.

[b] V_{DD} =5 V or 3.3 V, V_{PP}=12 V or 12.5 V.

[c] P/E cycles, programming/erasing cycles.

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 32

(i.e. programming) operation is always accompanied by a preceding erase operation.

The concept of the non-volatile memory cell, in which charges are injected into a floating gate, was first proposed for a P-channel EPROM cell, called the FAMOS (Floating-gate Avalanche-injection Metal Oxide Semiconductor) cell [1.6]. Once charges are injected by a write operation, they are held at the gate surrounded by insulator even if the power supply is turned off. In this cell, the erase and write operations are performed by ultraviolet exposure and avalanche injection, respectively. Since then, the development of EPROMs [1.18] has centered on the simplification of the memory cell and the reduction of the programming voltage (V_{PP}), which is the control-gate voltage for the write operation, as discussed later. Consequently, the above-mentioned FAMOS cell was replaced by N-channel stacked gate cells. The original V_{PP} of around 25 V was reduced to 12.5 V by tailoring impurity profiles. EPROMs, which were developed up to the 16 Mb generation, were eventually replaced by Flash memories. Note that data in EPROM chips are not erased as long as the chips are mounted on a memory board (card) in a system. To erase the data by exposing the memory cells to ultraviolet rays, the package must be removed from the board. Thus both a special package with a glass lid for penetrating the rays and a socket are necessary. This is a problem in terms of cost and ease of use. EEPROM solves the problem. To achieve better electrical erasability in EEPROMs, a variety of memory cells [1.18] have been developed. However, EEPROM has failed to attract a large market because of its large cell area and the standardization problem, although it was produced up to the 1 Mb generation. Thus, EEPROM was also replaced by Flash memory. Note that a small memory-capacity EEPROM is still used in a system LSI chip, where an EEPROM and a large logic cell are merged on to a chip. This is because the peripheral circuit for EEPROM is simple despite its large cell, thus making the overall chip design simple. Flash memory is based on the technology of either EPROMs or EEPROMs, and in price and functionality it falls somewhere between the two. Flash memory is suitable for applications that require a high storage density and reprogrammability without removing memory chips from the system, as discussed previously. Flash memory, as well as EEPROM, also need a high voltage when writing data into the cells (i.e. programming). In general, if the write operation utilizes a hot-electron injection mechanism, V_{PP} must be supplied directly from an external supply voltage. An inherently large hot-electron current in a cell transistor requires an external V_{PP}, as shown later. However, if it utilizes a tunneling mechanism, V_{PP} can be supplied internally from an on-chip V_{PP} generator that boosts V_{DD} to V_{PP} using a charge pump, thus enabling a single external V_{DD}. A quite small tunneling current allows the use of the charge pump which is inherently poor in current driving capability. Thus, it is essential to utilize a tunneling mechanism for both the write and the erase operations.

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 33

1.6.2 The Basic Operation of Flash Memory Cells

Since the concept of the Flash memory was first reported in 1984 [1.33], several approaches to Flash memory cells have been proposed [1.34–1.36]. However, their operations are based on the floating-gate concept, in which the threshold voltage of a transistor can be changed repetitively from a high to a low state, corresponding to the two states of the memory cell, i.e. the binary values ("1" and "0") of the stored bit. Cells can be "written" into the "1" or "0" states by either "programming" or "erasing" methods. One of the two states is called "programmed" and the other "erased". In some kinds of cells, the low-threshold state is called "programmed"; in others, it is called "erased". Although this may cause some confusion, the different terms are related to the different organizations of the memory array. Their read operation is performed by applying a gate voltage that is between the above two threshold voltages, and by sensing the current flowing through the transistor.

The simple model shown in Fig. 1.28 helps in understanding the electrical behavior of the Flash memory cell. Consider the case when the charge, Q_{FG}, is stored in the floating gate (FG). Then,

$$\begin{aligned} Q_{\mathrm{FG}} = {} & C_{\mathrm{FC}}(V_{\mathrm{FG}} - V_{\mathrm{CG}}) + C_{\mathrm{S}}(V_{\mathrm{FG}} - V_{\mathrm{S}}) + C_{\mathrm{D}}(V_{\mathrm{FG}} - V_{\mathrm{D}}) \\ & + C_{\mathrm{B}}(V_{\mathrm{FG}} - V_{\mathrm{B}}) \,, \end{aligned} \tag{1.1}$$

where V_{FG} is the potential at the FG, V_{CG} is the potential at the control gate, and V_{S}, V_{D}, and V_{B} are potentials at the source, drain, and bulk (substrate), respectively. Hence, the potential at the FG due to capacitive coupling is given by

$$\begin{aligned} V_{\mathrm{FG}} &= \frac{Q_{\mathrm{FG}}}{C_{\mathrm{T}}} + \frac{C_{\mathrm{FC}}}{C_{\mathrm{T}}} V_{\mathrm{CG}} + \frac{C_{\mathrm{S}}}{C_{\mathrm{T}}} V_{\mathrm{S}} + \frac{C_{\mathrm{D}}}{C_{\mathrm{T}}} V_{\mathrm{D}} + \frac{C_{\mathrm{B}}}{C_{\mathrm{T}}} V_{\mathrm{B}} \,, \\ C_{\mathrm{T}} &= C_{\mathrm{FC}} + C_{\mathrm{S}} + C_{\mathrm{D}} + C_{\mathrm{B}} \,. \end{aligned} \tag{1.2}$$

If the source and bulk are both grounded, (1.2) can be rearranged as

$$V_{\mathrm{FG}} = \frac{Q_{\mathrm{FG}}}{C_{\mathrm{T}}} + \frac{C_{\mathrm{FC}}}{C_{\mathrm{T}}} V_{\mathrm{CG}} + \frac{C_{\mathrm{D}}}{C_{\mathrm{T}}} V_{\mathrm{DS}} \,. \tag{1.3}$$

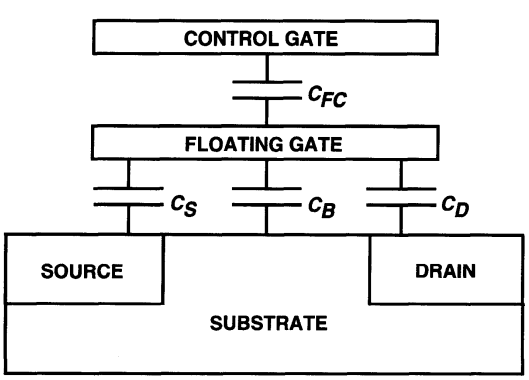


Fig. 1.28. A schematic of the cross-section of a Flash memory cell [1.35]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 34

Here, we define $V_{\mathrm{T}}(\mathrm{FG})$ as the threshold voltage of the FG transistor, the gate of which is the FG. If V_{FG} is lower than $V_{\mathrm{T}}(\mathrm{FG})$ the FG transistor turns off, while if it is higher than $V_{\mathrm{T}}(\mathrm{FG})$ it starts to turn on. Consider another transistor (i.e. the CG transistor) with the same source and drain, the gate of which is the control gate. Obviously, the CG transistor also starts to turn on when $V_{\mathrm{FG}} = V_{\mathrm{T}}(\mathrm{FG})$. This point is just the treshold voltage, $V_{\mathrm{T}}(\mathrm{CG})$, of the CG transistor. Thus, $V_{\mathrm{T}}(\mathrm{CG})$ is derived from (1.3) with $V_{\mathrm{FG}} = V_{\mathrm{T}}(\mathrm{FG})$ and $V_{\mathrm{CG}} = V_{\mathrm{T}}(\mathrm{CG})$, as follows:

$$V_{\mathrm{T}}(\mathrm{CG}) = \frac{C_{\mathrm{T}}}{C_{\mathrm{FC}}} V_{\mathrm{T}}(\mathrm{FG}) - \frac{Q_{\mathrm{FG}}}{C_{\mathrm{FC}}} - \frac{C_{\mathrm{D}}}{C_{\mathrm{FC}}} V_{\mathrm{DS}} \,. \tag{1.4}$$

Therefore, $V_{\mathrm{T}}(\mathrm{CG})$ which is the memory-cell threshold voltage, depends on Q_{FG}. The $V_{\mathrm{T}}(\mathrm{CG})$ shift, ΔV_{T}, is thus given by

$$\Delta V_{\mathrm{T}} = -\frac{\Delta Q_{\mathrm{FG}}}{C_{\mathrm{FC}}} \,, \tag{1.5}$$

where ΔQ_{FG} is the shift of Q_{FG}. This equation shows that the role of the injected charge is to shift the I–V curves of the cell (i.e. the CG transistor). If the reading biases are fixed (usually at $V_{\mathrm{CG}} \sim 5\,\mathrm{V}$ and $V_{\mathrm{DS}} \sim 1\,\mathrm{V}$), the presence of charge greatly affects the current level used to sense the cell state. Figure 1.29 shows two curves: curve A represents the "0" state, while curve B is for the "1" state with a $V_{\mathrm{T}}(\mathrm{CG})$ shift. The current is approximately

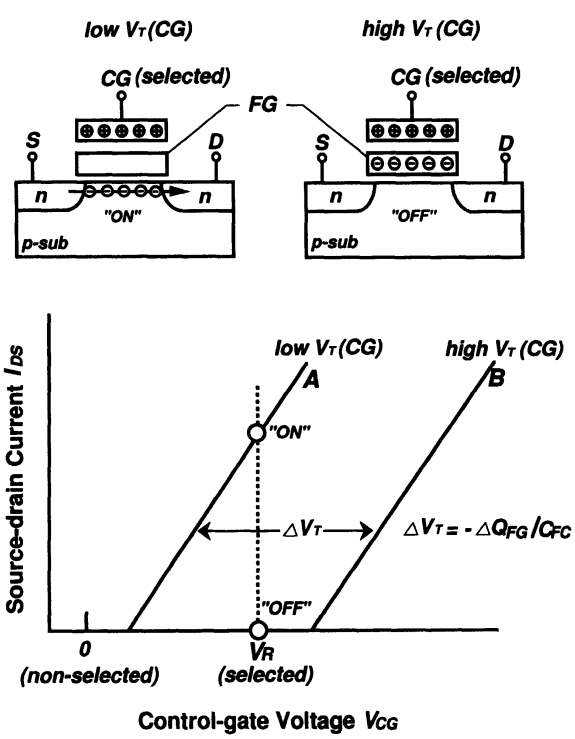


Fig. 1.29. The principle of the read operation of a Flash memory cell [1.36]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 35

100 μA for "0" while it is 0 for "1", when V_{CG} is chosen to be an appropriate voltage between two threshold voltages. It is indispensable for the normally off transistor to start to turn on only by the application of V_{CG}; otherwise, the cells connected to the same data line may turn on even without the application of a word pulse. Thus, to ensure a successful read operation the following relationships must be satisfied:

$$V_R > \text{low}\, V_T(CG) > 0\ ,$$
$$V_R > \text{high}\, V_T(CG)\ .$$

There are two typical mechanisms to transfer electric charges from and into the FG. They are the hot-electron injection (HEI) mechanism and the Fowler–Nordheim (FN) tunneling mechanism [1.35].

Hot-Electron Injection. Electrons traveling from the source to the drain are "heated" by a high lateral electric field (between source and drain), causing avalanche breakdown phenomena in the vicinity of the drain. The resultant impact ionization generates hole–electron pairs in the drain region. The electrons of the pairs are injected into the floating gate through the oxide by a transverse electric field (between the channel and the control gate), while the holes of the pairs flow to the substrate as the substrate current. Figure 1.30 shows the relationship between the gate current (I_G) and the gate voltage (V_{GS}) of an FG transistor. The gate current starts to flow in accordance with the start of the channel (source–drain) current flow when

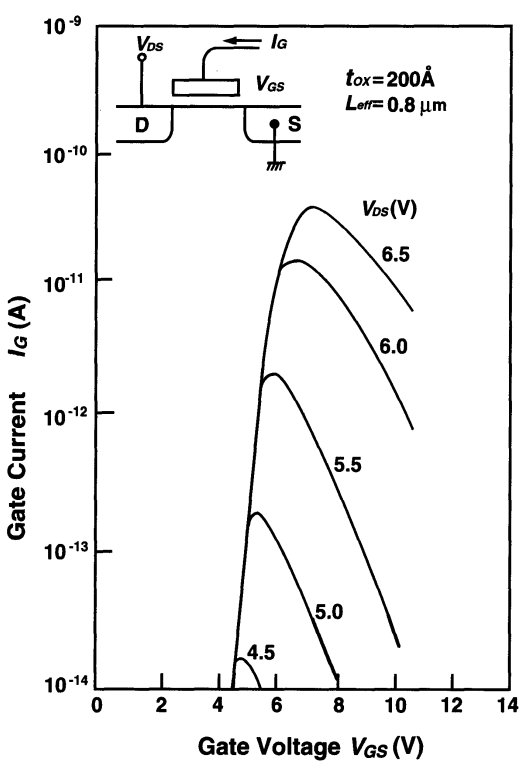


Fig. 1.30. The gate current caused by channel hot-electron injection [1.35]

V_{GS} is increased. It increases with V_{GS} because of the increase in the channel current. An excessive V_{GS} for a fixed V_{DS}, however, decreases I_G because of a reduced drain–gate potential difference. Thus, if the bias condition for the drain and floating gate is set so that the gate current is maximized, this realizes the fastest injection speed. This injection always entails a large channel current, causing a high power dissipation. For example, the current is as high as about 0.5 mA at a control-gate voltage of 12 V.

Fowler–Nordheim Tunneling. The current density (J) of the tunneling current is obtained in the simplest form as

$$J = AE^2 \exp(-B/E) \tag{1.6}$$

where A and B are almost constants, and E is the electric field. Figure 1.31 shows $\log J$ versus E. There is a large variation, of about seven orders of magnitude, in the tunnel current when the field is changed from 7 MV cm^{-1} to 10 MV cm^{-1}. Since the field is roughly the applied voltage divided by the oxide thickness, a reduction in the oxide thickness for a fixed applied voltage produces a rapid increase in the tunneling current. An optimum thickness, of about 10 nm with 10 MV cm^{-1}, however, is chosen in present-day devices as the result of trade-off between performance constraints and reliability concerns. Thin oxides reduce the erasing speed. However, they increase power consumption and degrade the reliability of the oxide instead. The above exponential dependence of the tunnel current on the oxide – electric field calls for very good process control. Otherwise, some critical problems are caused. For example, a very small variation in the oxide thickness among the cells in a memory array produces a great difference in the erasing or programming currents, thus spreading the threshold voltage distribution in both logical states. Moreover, the tunneling currents may become important in device reliability at low fields, either in the case of poor-quality tunnel oxides or when

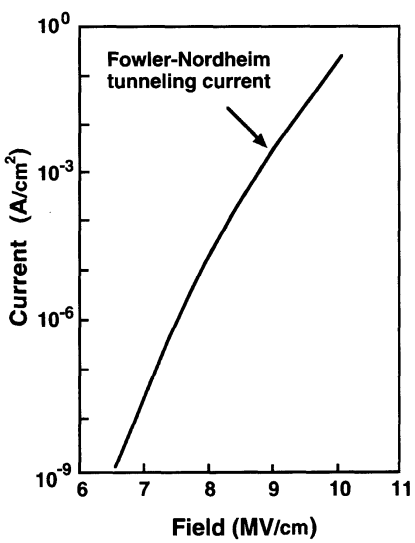


Fig. 1.31. The Fowler-Nordheim tunneling current, as a function of the electric field [1.35]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 37

thin oxides are stressed many times at high voltages. In fact, poor-quality oxides are rich in interface and bulk traps, and trap-assisted tunneling is made possible, since the equivalent barrier height seen by electrons is reduced, and thus tunneling requires a much lower oxide field than $10\,\mathrm{MV\,cm^{-1}}$. The oxide defects, whose density increases with decreasing oxide thickness, must be avoided to control the erasing and programming characteristics, and to ensure good reliability.

The following are the typical types of Flash cell. Some of them differ in their charge-transfer mechanisms, their cell structures, and the logical connections of their cells.

The NOR Cell. The name of the cell is derived from the logical connection of the cells. Figures 1.32 and 1.33 show the industry-standard one-transistor stacked-gate Flash memory cell [1.37] and its operation [1.9]. The tunnel oxide under the floating gate is about 10 nm thick. To have a junction that can sustain the high applied voltages without breaking down, the source junction is carefully designed by using the process to achieve a lighter and deeper junction. Therefore, the source diffusion is realized differently from the drain diffusion, which does not undergo such high bias conditions. Oxide/nitride/oxide (ONO) interpoly dielectrics are used to realize a high value of C_{FC}. The cell uses the FN tunneling current for erasing, and the electron injection for programming (write).

The erase operation is performed by a combination of a 0 V control-gate voltage and a high source-voltage, so that electrons, if any, at the floating gate are ejected to the source by the tunneling effect of the thin oxide. All cells are erased simultaneously, since a high voltage is applied concurrently to the sources of all cell transistors. Here, different initial values of the cell threshold voltage and different gate oxide thicknesses of the FG transistors may cause a variation in the threshold voltage at the end of the erase operation. A higher threshold voltage after erasing results in a slower read operation. Thus, in practice, before applying the erase pulse, all of the cells in the array/block

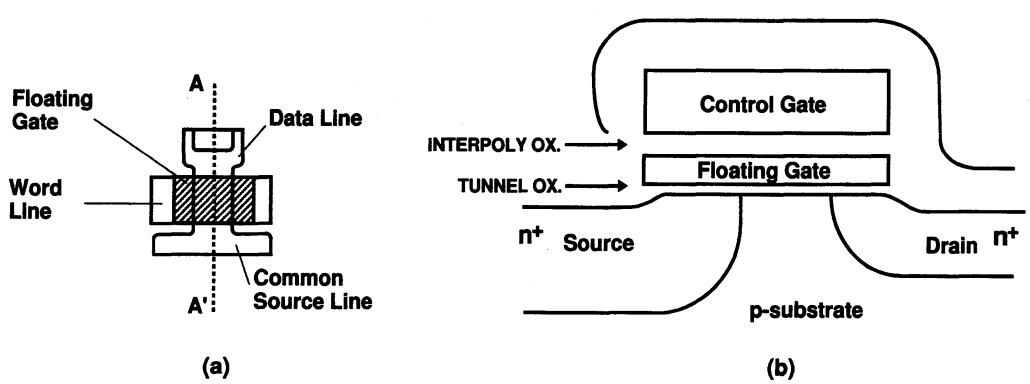


Fig. 1.32. The NOR cell [1.37]. (**a**) The layout of a typical double-polysilicon stacked gate cell; (**b**) a schematic cross-section along line A-A′

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 38

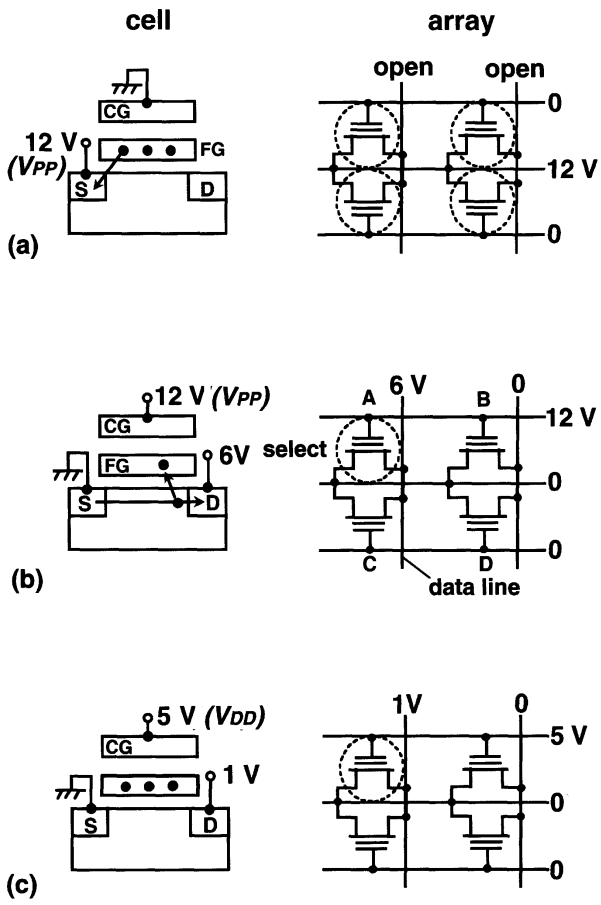


Fig. 1.33. The basic operation of the NOR cell [1.9]. (**a**) Erase; (**b**) write; (**c**) read

are programmed so that all of the thresholds start at approximately the same value. After that, an erase pulse that has a controlled width is applied. After an erase pulse, however, there may be typical bits and fast erasing bits due to gate-oxide thickness variation. Therefore, subsequently, the whole array/block is read to check whether or not all of the cells have been erased. If not, another erase pulse is applied and another read operation follows. This algorithm is applied until all of the cells have threshold voltages that are lower than the erase verify level. A similar algorithm is also useful to avoid "over-erase" that may cause normally on transistors. Typical erasing times are in the range 100 ms–1 s. As for the write operation (i.e. programming), a high-voltage pulse is applied to the control gate so that the capacitive coupling makes the transistor turn on, with a floating-gate voltage raised from a voltage that was set by the preceding erase operation. If the write data corresponds to a high voltage (for example, "1") applied to the drain (i.e. the data line), a high source–drain current of about 0.6 mA flows. As a result, hot electrons generated by the avalanche breakdown phenomena in the vicinity of the drain are injected into the floating gate. Consequently, after turning off the word pulse, the FG transistor is deeply cut off with a negative gate voltage. If the write data is for a drain voltage of 0 V ("0"),

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 39

the floating gate remains at the previous erasing state of having no electrons. Thus, after programming, the cell-threshold voltage, $V_T(CG)$, becomes high or low. The shift in the threshold voltage, ΔV_T, depends upon the width of the programming pulse. To have $\Delta V_T = 3-3.5\,V$, a pulse width with typical values in the $1-10\,\mu s$ range must be applied. Here, the drain is made open to prevent a turn-on current in the transistor, caused by capacitive coupling from the source to the floating gate. The read operation is done by the application of 5 V ($V_R = V_{DD}$) to the control gate and 1 V to the drain, as described before. An access time of 50−100 ns is obtained. Note that a higher drain voltage may cause "soft-write" for the cell during the read operation, although it offers faster sensing with a larger read current.

The reliability issue regarding the programming/erasing cycles and retention characteristics is highlighted for Flash memories. The following is an example of programming disturbances: there are two major disturbances when programming the selected cell (A) in Fig. 1.33. One is due to the high voltage (12 V) applied to the control gates of the non-selected cells (B) on the same word line. The other is due to the medium high voltage (6 V) applied to the drains of the non-selected cells (C) on the same data line. For cell B there might be tunneling of electrons from the FG to the control gate through the interpoly oxide if the FG is filled with electrons. This induces a charge loss, reducing the margin for the high threshold voltage. There might also be tunneling of electrons from the substrate to the FG if the FG is "empty". This induces a charge gain, reducing the margin for the low threshold voltage. For cell (C) electrons tunnel from the FG through the gate oxide to the drain, reducing the margin for the high threshold voltage. Disturbances similar to the above are present even during read operations. That is why a drain voltage as low as 1 V is applied, to avoid the soft-write that can occur during read operations. The influence of disturbances becomes more and more prominent when increasing the number of reading–programming or programming–erasing cycles.

Flash memory needs multipower supplies internally, calling for on-chip voltage converters. A voltage down-converter, which converts the high external supply voltage to a low internal supply voltage, allows the resultant low-voltage generator to drive the internal load with a quite large output current. On the other hand, a voltage up-converter, by means of charge pumps, suffers from a small output current. The details will be discussed in Chap. 5. In any case, V_{PP} must supply a large erasing current, because all of the cells in a block as large as 64 KB (i.e. 512 K cells) are simultaneously erased. Thus, an external V_{PP} supply is indispensable. In addition, the drain supply voltage (6 V) in the write operation must manage a large avalanche breakdown current that is necessary when a number of cells (for example, 16 cells) are written simultaneously. Thus, the drain voltage must be generated from the external V_{PP} (12 V) through an on-chip voltage down-converter. An on-chip voltage up-conversion from the external V_{DD} (5 V) to 6 V by using a charge

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 40

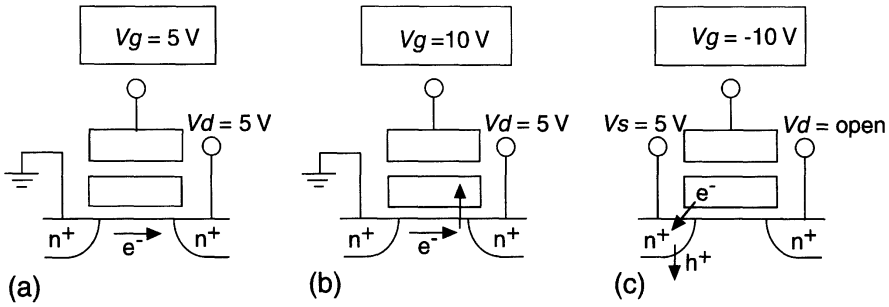


Fig. 1.34. The negative word-line voltage scheme [1.38]. (**a**) Read, 60 µA/cell; (**b**) program, 0.5 mA/cell (8 mA for 16 I/Os); (**c**) erase, 10 nA/cell (5 mA for 64 Kb block)

pump fails to supply a high enough current. A read voltage of 1 V is also generated from V_{DD}, using another voltage down-converter. Thus, the cell needs two external power supplies, V_{PP} and V_{DD}.

A negative word-voltage erasing scheme [1.38] allows even the NOR cell to operate at a single external V_{DD}, as shown in Fig. 1.34. Here, there are minor changes, caused by an advanced device, in the values of V_{PP} and currents, from those described so far. The scheme features a voltage setting of a negative control-gate voltage and a V_{DD} source voltage in the erase operation. This setting gives almost the same tunneling current as that for the conventional setting shown in Fig. 1.33. The V_{DD} at the source can supply a large tunneling current of 5 mA, which is necessary for a block erase. An on-chip negative voltage (−10 V) converter from V_{DD}, which comprises a charge pump, manages to drive its load despite the charge pump because of pure capacitive CG load. Even for write and read operations, single V_{DD} operation is realized: The external V_{DD} supplies a large enough hot-electron current of 8 mA at the V_{DD} drains, and generates a high CG voltage of 10 V through a voltage up-converter. Note that reduction of the source voltage simplifies the source structure.

The NAND Cell. In this type of cell the elementary unit is not composed of the single three-terminal cell, but of more FG transistors connected in a series (eight or 16), which constitutes a chain connected to the data line and to ground through two selection transistors, as shown in Fig. 1.35 [1.39]. This organization eliminates all contacts between word lines, which can be separated by their minimum design rule, thus reducing the occupied area by 40%. Figure 1.36 shows the cross-section of an 8-bit elementary block for a NAND array with peripheral circuits [1.39]. The FN tunneling current is used for both erasing and programming. The erase voltages are 20 V to the n-substrate, the p-well 2, and the drain and source of the chain, and 0 V to all the control gates of the chain when the selection transistors (Q_{S1}, Q_{S2}) are turned on. This biasing induces electron tunneling toward p-well 2, resulting in a low threshold voltage for all of the cells. Note that a low

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 41

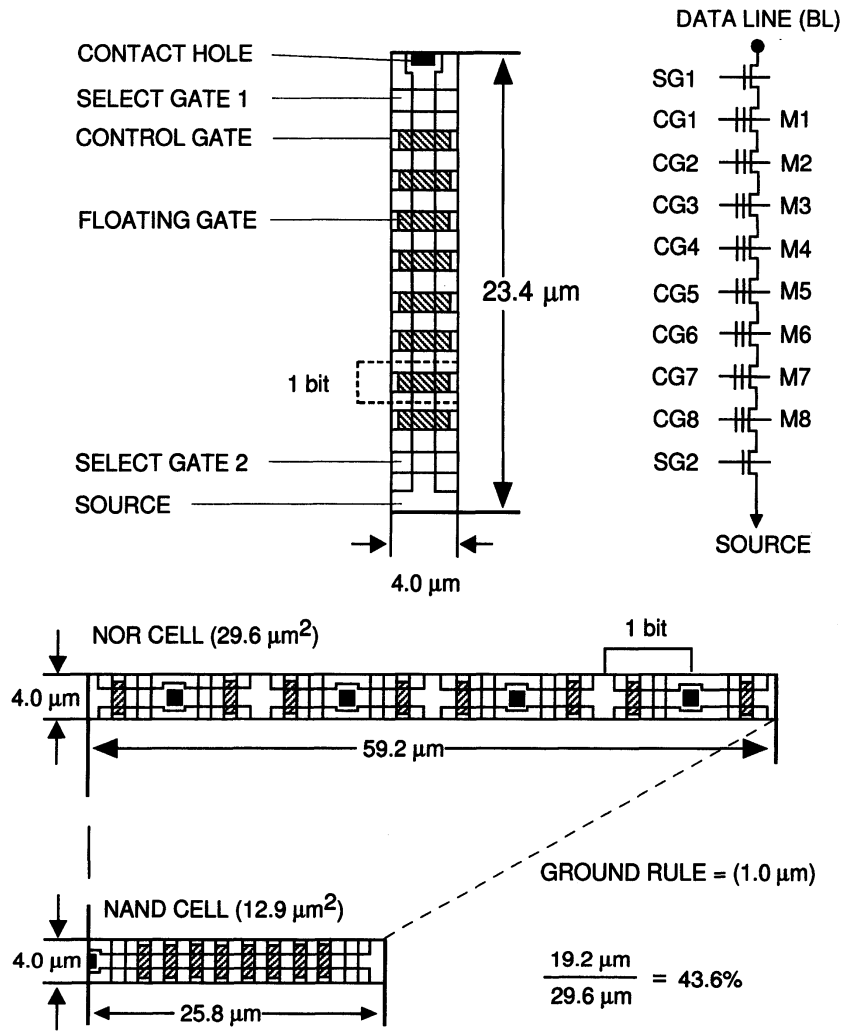


Fig. 1.35. NAND architecture. The dimensions of a NAND array are compared to those of a NOR array [1.39]

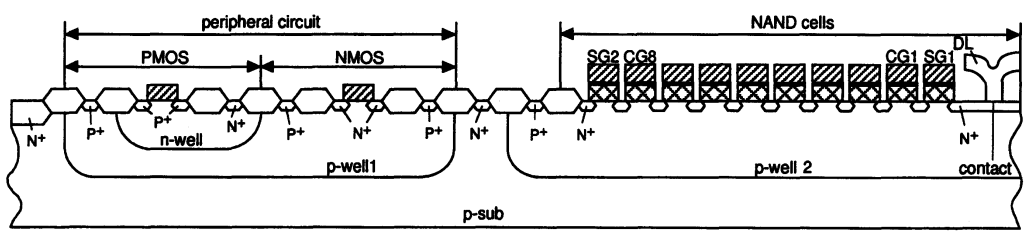


Fig. 1.36. A cross-section of NAND cells and the peripheral circuit [1.39]

threshold voltage is set to be negative (i.e. normally on), which differs from the NOR cell. There is no voltage difference between the drain of each cell transistor and the p-well, so that there is no breakdown of the junction. The programming voltages are 20 V to the control gate of the selected cell and

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 42

10 V to all of the remaining control gates of non-selected cells, with p-well 2 grounded. The selection transistors are biased to connect the chain to the data line and isolate it from the ground. If a "0" is to be stored, the data line is grounded. Hence, the sources, drains, and channels of the cell transistors are grounded, and only the selected cell transistor has an electric field in the oxide to induce electron injection from the channel (substrate) into the FG, causing a high threshold voltage. This is because the FG voltage of the selected cell is raised to a sufficiently high tunneling voltage of about 10 V, while the other FG voltages are raised to a sufficiently low voltage, of about 5 V. This is justified by an exponential relationship between the tunneling current and the electric field, as expressed in (1.6). If a "1" is to be stored, the data line is biased at 10 V. However, there is no tunneling for the selected cell because of a small voltage difference between the FG and the substrate, which keeps the threshold voltage negative. Obviously, to transfer the data-line voltage to any selected cell in the chain, the non-selected cell transistors must always be conductive, independent of their stored information. Thus, the FG voltages of the non-selected cells must exceed the high threshold voltage sufficiently. The read operation is performed by applying 5 V (V_{PASS}) to all of the control gates except the selected one, which is grounded. The selection transistors are turned on to connect the chain to the ground and data line. Thus, the data line, which has been precharged, is discharged if the stored data is "1" (normally on). It holds the precharged voltage if the stored data is "0" (a threshold voltage high enough to cut off the transistor). Figure 1.37 shows the voltage relationships for the selected and non-selected cells. A read operation succeeds when the following conditions for the selected cell and for the non-selected cells are satisfied, respectively:

$$\mathrm{low}\, V_{\mathrm{T}}(\mathrm{CG}) < 0 < \mathrm{high}\, V_{\mathrm{T}}(\mathrm{CG})\ ,$$
$$V_{\mathrm{PASS}} > \mathrm{high}\, V_{\mathrm{T}}(\mathrm{CG})\ .$$

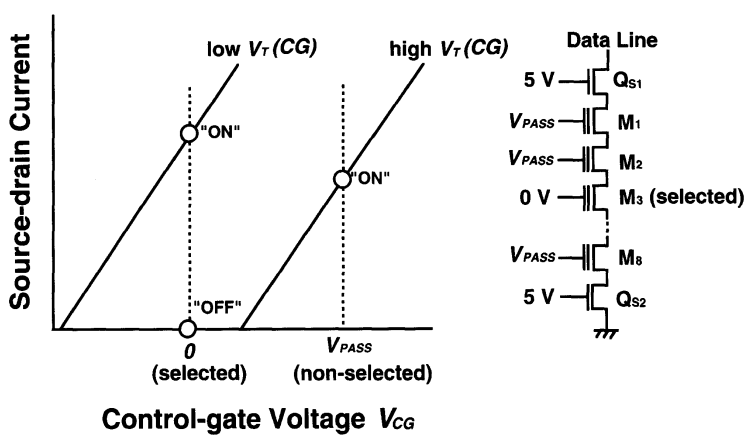


Fig. 1.37. The principle of the read operation of the NAND cell [1.39]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 43

Since tunneling is more efficient than hot-electron injection, currents are smaller, and different supply voltages can be internally generated by on-chip charge pumps through the use of a single external power supply. The small currents allow an increase in the number of parallel programmed cells without increasing power consumption. Thus, the programming time per byte is as fast as 200–400 ns. Electron tunneling from whole channel region is uniform through the oxide. This makes erasing fast, exemplified by 6 ms per block and around 100 ms per chip. However, the access time is slow because of a small cell current of around 1 μA. A chain structure is responsible for this current.

The DINOR (Divided Bit-line NOR) Cell. This features the high-speed random access of the NOR cell and the low power of NAND cell. FN tunneling is used for both erasing and programming. Figure 1.38 shows the cross-section and equivalent circuit [1.40]. Each data (or bit) line is composed of a hierarchical structure of poly-Si subdata lines, each of which connects 32 or 64 cells, and metal-2 main data lines. Thus, a main data line can "see" only one selected subdata line, which enables a reduced capacitance and thus a high speed. The hierarchical structure also confines the array area, which gives rise to disturbances, within the resultant small segment. The cell area is reduced by using self-aligned drain contacts without double-diffused drains. Metal-1 word-line strapping is used for high speed. In DINOR cells, the programming and erasing operations are opposite to those in the conventional NOR cell. To erase means to set the V_T of the cells to have a high value, while to program means to set the V_T of the selected cell to have a low value. Erasing is achieved by applying a high voltage of 10 V to the control gates and a negative voltage of −8 V to the p-well. Thus, electrons are injected into the FGs of all of the cells via channel FN tunneling. When programming, a high drain voltage of

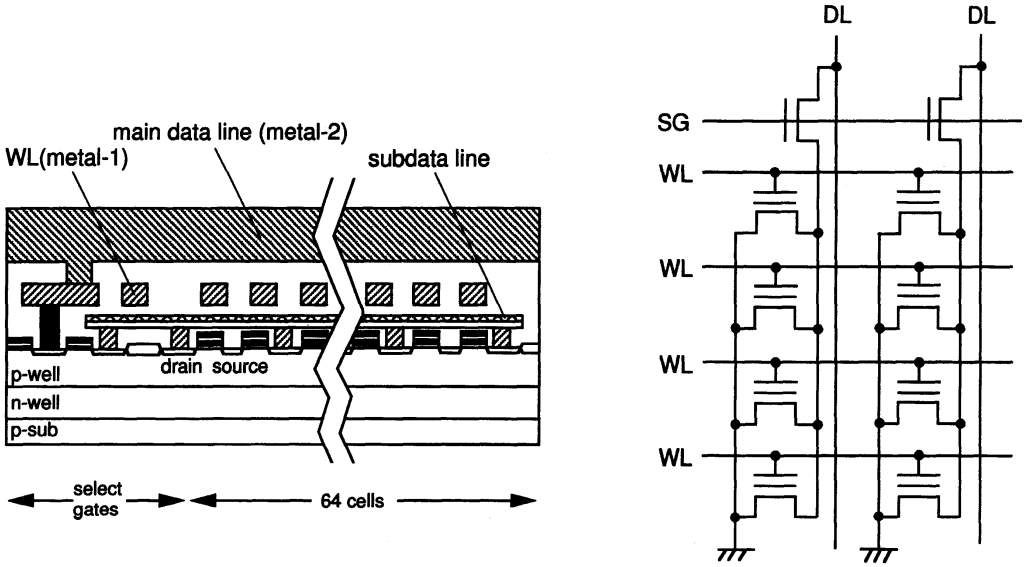


Fig. 1.38. The cross-section and equivalent circuit of DINOR cells [1.40]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 44

6 V is applied to the subdata line of the selected blocks, and a negative voltage of −8 V is applied to the selected control gate. Thus, electrons are selectively ejected from the FG though the gate/drain overlapping area. Disturbances are reduced by the high-V_T erase. Overprogramming is avoided by the V_T-verify algorithm, which ensures a normally off condition even after programming.

The AND cell. This cell operates like the DINOR cell. Figure 1.39 shows the cell structure and equivalent circuit [1.41]. It increases the gate coupling in the word-line direction and realizes a small cell area due to diffused drain and source lines. Both the drain and source lines have hierarchical structures. Each subdata line is connected to a metal main data line through a selection transistor. Each subsource line is also connected to a main source line through another selection transistor. This structure greatly reduces programming disturbances from other blocks.

The Multilevel Cell. To improve Flash memory density, the multilevel cell concept [1.42], which increases the number of possible states in a cell, has been proposed. If k bits are stored in a memory cell, the density increases by a factor of k. Instead, for a fixed maximum cell operating voltage the signal-to-noise (S/N) ratio degrades by $1/(2^k - 1)$ for the conventional 1 bit per cell scheme. Thus, a high S/N design is the key. The inherent features of the Flash cell, such as a gain cell and a V_T-adjustment capability for each cell, might overcome the S/N issue, although the concept has been difficult to implement on a commercial DRAM chip. The number of states is determined by the total available charge range, the ability to program a state accurately, ability to read a state accurately, and the disturbance of a state over time [1.43], as understood below.

If programming can be done accurately enough, the cell can store one of four discrete charge bands to achieve 2 bits per cell storage, as shown in

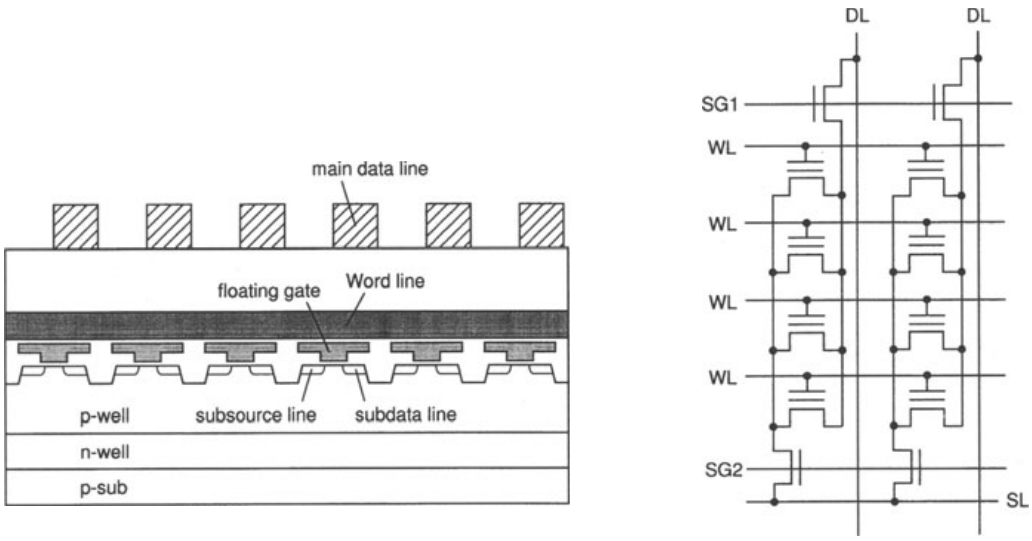


Fig. 1.39. The cell structure and equivalent circuit of an AND cell [1.41]

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 45

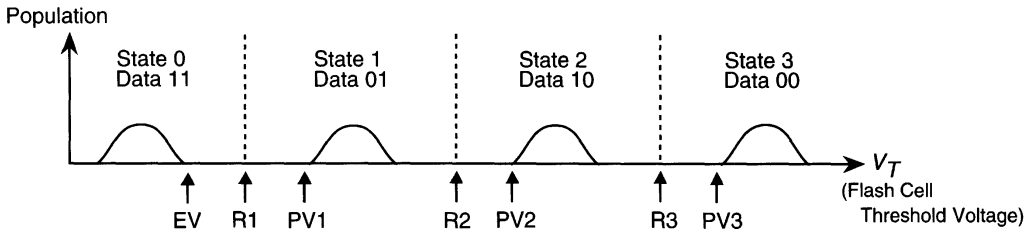


Fig. 1.40. The multilevel-cell threshold voltage distribution for 2 bits/cell storage [1.42]. The states (ranges of threshold voltages) are determined by Program Verify Reference V_T (PV) and Erase Verify Reference V_T (EV). Flash cells are used to generate the reference V_T values to compare the V_T of array memory cell with that of the cells. Logic 11 if V_T (array) $< V_T$(R2) and V_T (array) $< V_T$(R1); logic 01 if V_T (array) $< V_T$(R2) and V_T (array) $> V_T$(R1); logic 10 if V_T (array) $> V_T$(R2) and V_T (array) $< V_T$(R3); logic 00 if V_T (array) $> V_T$(R2) and V_T (array) $< V_T$(R3)

Fig. 1.40 [1.42]. State 0 (1,1) corresponds to the erased state. In the figure EV is the V_T-verify level for erasing. The remaining states are controlled by the degree of programming. PV 1, PV 2, and PV 3 are the V_T-verify levels for programming, and each state can be discriminated by a sensing circuit, shown in Fig. 1.41 [1.42]. The reference current (R2) of the first stage, that corresponds to V_T(R2), discriminates between states 0 and 1 and states 2 and 3 of the selected cell. The reference currents (R1, R3) of the next stage discriminate between state 0 and state 1, and state 2 and state 3, respectively.

1.6.3 Advances in Flash-Memory Design and Technology

Circuit innovations in the development of non-volatile memory are summarized in Table 1.6. In addition to advances in the high-density memory cell discussed so far, innovative circuits [1.18] focusing on ease of use and high speed have been developed. They are a byte-erasable scheme, the CMOS peripheral circuit, redundancy, and page (32–128 bytes) programming. Page programming effectively performs high-speed programming through parallel programming of many data, each of which has been latched into a latch circuit connected to each data line. Moreover, data polling that acknowledges completion of erase and programming, a silicon signature that enables automatic programming based on read data from ROM (in which a procedure regarding erasing and programming is stored), command-port architecture, an internal erase/erase-verify function, and a single power supply scheme have all improved ease of use.

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 46

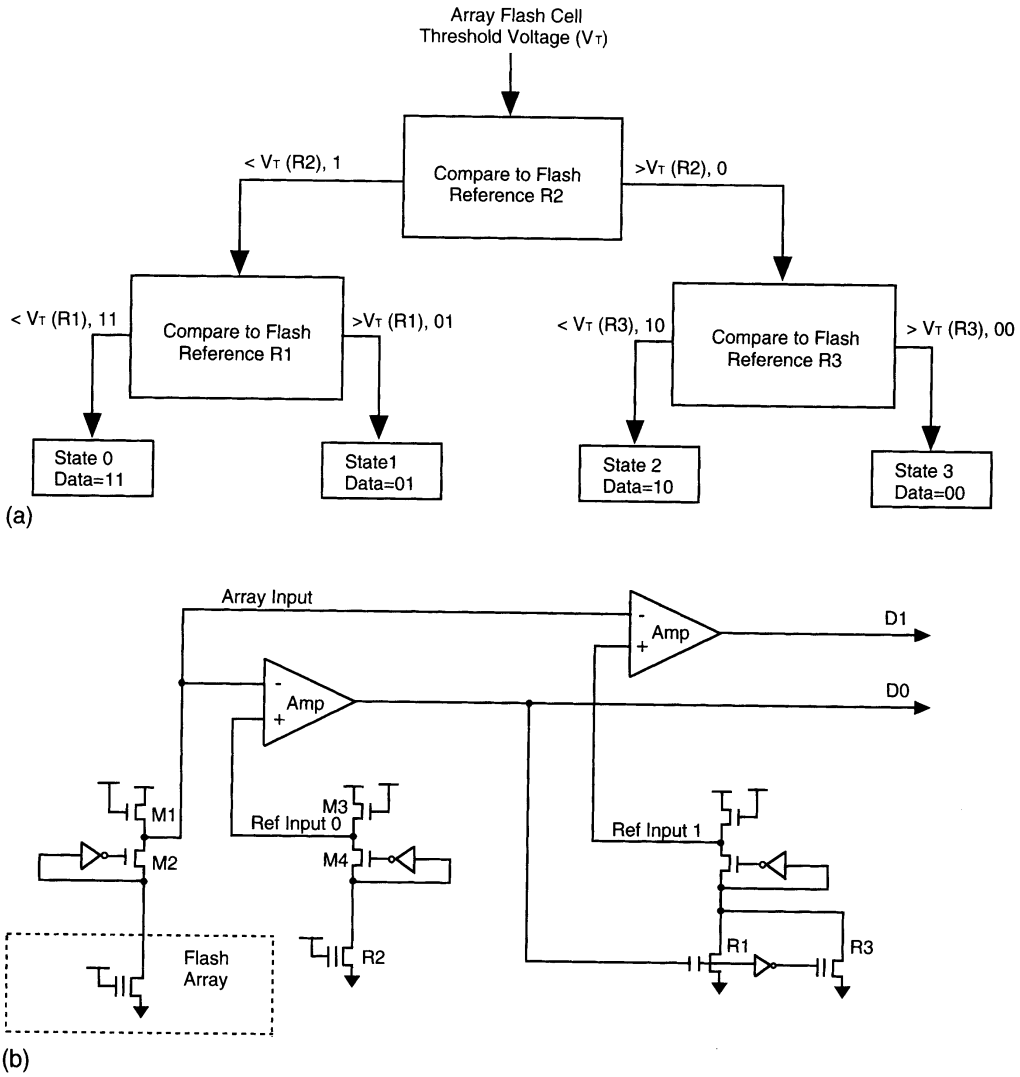


Fig. 1.41. The binary search sensing scheme (**a**) and its circuit (**b**) for 2 bits/cell storage [1.42]

Table 1.6. Circuit innovations for non-volatile memory [1.9][a]

Capacity	Power Supply, V_{PP}/V_{DD} (V)	Memory Cell	Circuits
2 Kb	~ 50 (EP)	PMOS 2-T FAMOS (EP)	
8 Kb	25 (EP)	NMOS 1-T staked gate (EP)	
16 Kb	25/5 (EP) 5/5 (EEP)	MNOS, FLOTOX, triple poly-Si (EEP)	–Byte-erasable (EEP)
64 Kb	21/5 (EP) 5/5 (EEP)		–CMOS peripheral –Redundancy –Page programming (EEP) –Data Polling (EEP)
128 Kb	21/5 (EP) 5/5 (EEP) 12/5 (FL)	Split gate (FL)	
256 Kb	12.5/5 (EP)	Stacked gate, FLOTOX	–Silicon signature (EP)
	5/5 (EEP) 12/5 (FL)	Triple poly-Si (FL)	–Command-port architecture (FL)
1 Mb	12.5/5 (EP) 5/5 (EEP) 12/5 (FL)		–Word-wide (×16) (EP) –Page programming (EP) –Internal erase/erase-verify (FL)
4 Mb	12.5/5 (EP) 12/5 (FL)	NAND (FL)	
16 Mb	12.5/5 (EP) 12/3.3, 5 (FL) 5/5 (FL)		–Negative word voltage (FL)
32 MB			–Multilevel Cell (FL)

[a]EP, EPROM; EEP, EPROM; FL, Flash memory; T, transistor; $V_{PP} = V_{DD}$ denotes single power supply.

BESANG EX2036
Micron v. BeSang, IPR2023-00900
Page 48