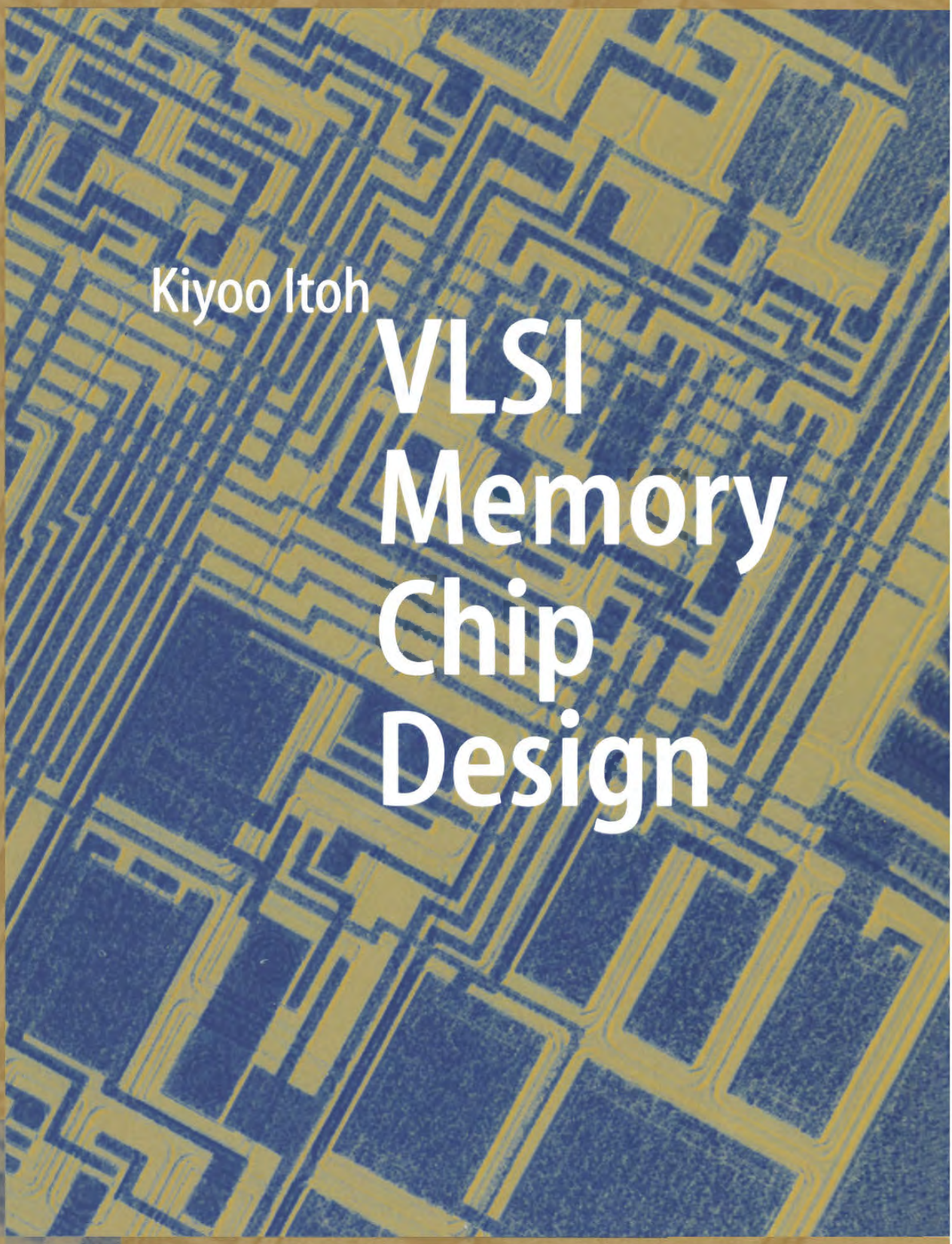
ADVANCED
MICROELECTRONICS



Kiyoo Itoh

VLSI Memory Chip Design

Springer

Itoh

VLSI Memory Chip Design

APPLE-1020

Springer Series in

ADVANCED MICROELECTRONICS 5

Physics and Astronomy ONLINE LIBRARY

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

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/

Kiyoo Itoh

VLSI Memory Chip Design

With 416 Figures and 26 Tables



Springer

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

Preface

The VLSI memory era truly began when the first production of semiconductor memory was announced by IBM and Intel in 1970. The announcement had a profound impact on my research at Hitachi Ltd., and I was forced to change fields: from magnetic thin film to semiconductor memory. This change was so exceptionally sudden and difficult, I felt like a victim of fate. Looking back, however, I realize how fortunate I was. I have witnessed an unprecedented increase in memory capacity (DRAM, for example, has had a 6-order increase in the last three decades – from the 1-Kb level in 1970 to the 1-Gb level today). I have contributed to this progress with full involvement in memory-chip development over my career. Such rapid progress would have been impossible without many of the inventions and innovative technologies, and without the effort of many talented people. Unfortunately, few systematic books on memory-chip design have been written by experts. This is a result of two factors: the difficulty of involving university professors because of rapidly changing technology requiring huge investments and development resources, and a shortage of time on the part of chip designers in industry due to severe competition in the memory-chip business. Therefore, LSI memory-chip design has been isolated from the outside, preventing a deeper understanding of the technology.

This book is based on my 30-year memory-chip (particularly DRAM) design career. In addition to memory circuits and subsystem design issues, I describe boundary issues between processes, devices, and circuits. I also attempt to systematically describe concepts that remain unclear, and discuss state-of-the-art memory-chip design. This book will be beneficial to students and engineers interested in memory-chip design, and also to process and device engineers involved in memory-chip development.

Chapter 1 describes the basics of various VLSI memory chips including DRAM, SRAM, and nonvolatile memory. Particular emphasis is paid to internal organization, operation principles and general trends in chip performance. Chapter 2 deals with the basics of RAM design and technology. The elements constituting a memory chip (MOSFETs, capacitors, and resistors), MOS memory circuits, the scaling law, and other relevant technologies are discussed. The first two chapters lay the groundwork for understanding the rest of the book. Chapter 3 focuses on DRAM chip design. After the catalog

specifications, the determinants of chip performance, the basic technologies for memory-array configuration and each of the peripheral circuits are described. Then, refreshing schemes and redundancy are explained. Chapter 4 discusses the signal-to-noise (S/N) issue in DRAM which strongly influences stable operation in the memory cell and, thus, in the chip. The relationship between memory-cell structure and its driving/sensing is explained in relation to the S/N issue. Chapter 5 describes on-chip voltage generators used for power-supply conversion. These generators are essential for power-supply standardization and stable operation. Chapter 6 discusses subsystem-memory architectures. These are increasingly important in providing wide bandwidth (i.e. throughput) for modern DRAMS. Chapters 7 and 8 describe low-power/low-voltage memory circuits, emphasizing the importance of the partial activaiton of multi-divided arrays, and of lowering power-supply voltage. Low voltage inevitably needs the subthreshold-current reduction which is the key to future LSI design.

I am indebted to many people including colleagues and the office administration staff members, Ms. Hosoda and Ms. Ohta, at Hitachi Ltd. They offered support, advice, and the material needed to finalize my work. Special thanks go to my wife, Kyoko. Without her continuing support and patience this book would not have been possible.

Stanford, January 2001 *Kiyoo Itoh*

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

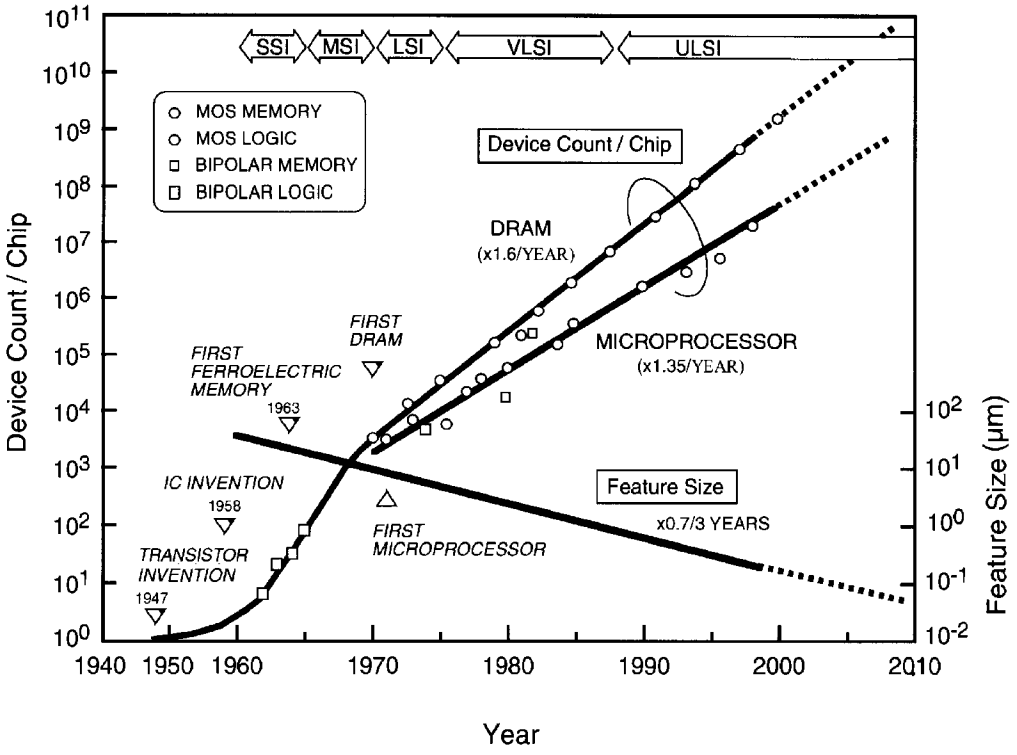


Fig. 1.1. The trend in the device count per chip [1.2]

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].

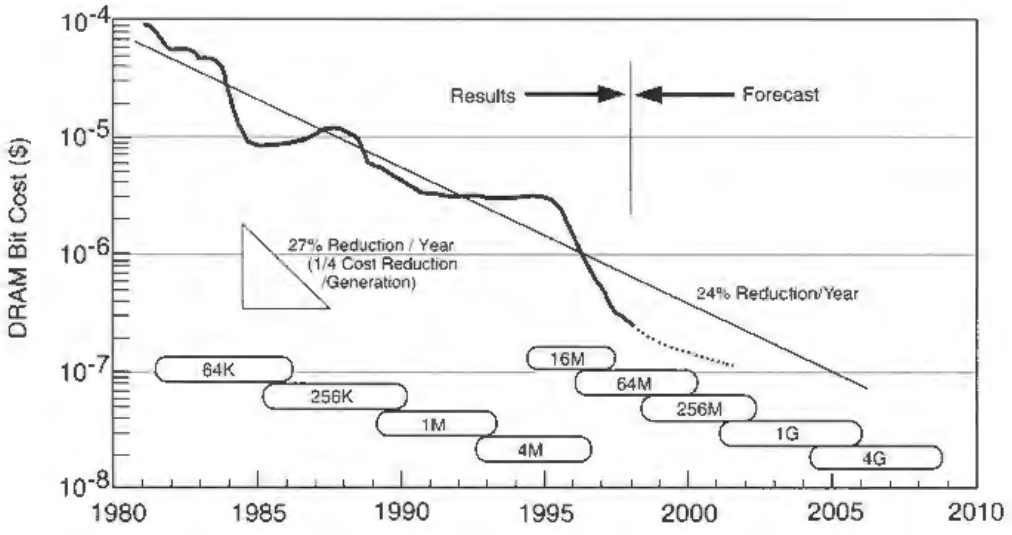


Fig. 1.2. The reduction in the bit cost of DRAM [1.7]

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

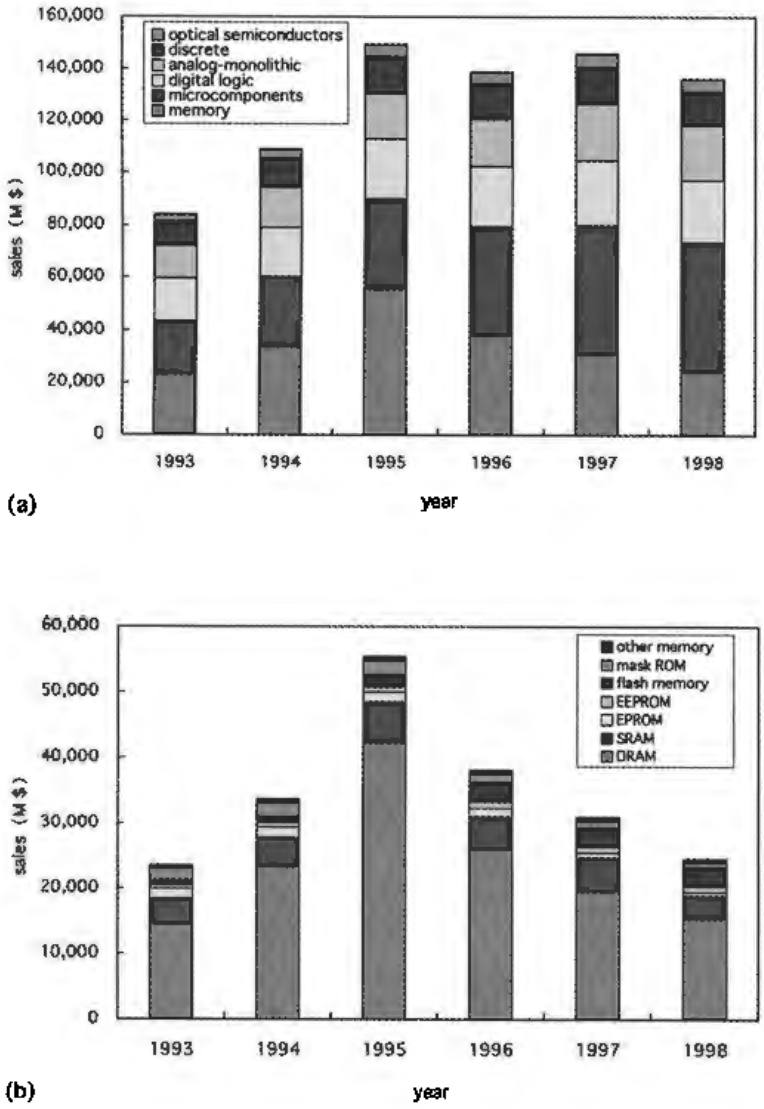


Fig. 1.3. The amount of sales of LSIs (a) and various memory chips (b) [1.8]

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.

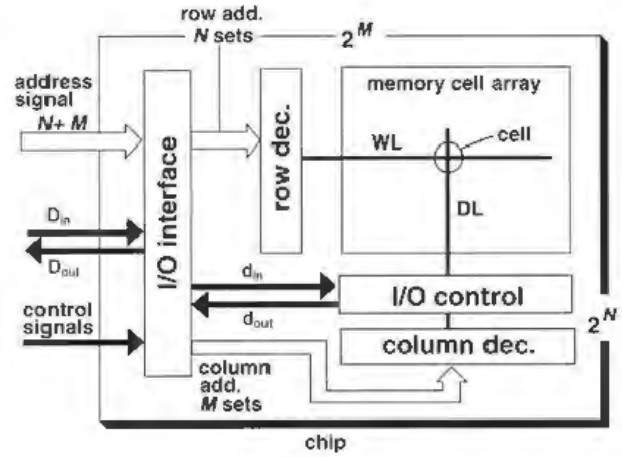


Fig. 1.4. The memory chip configuration

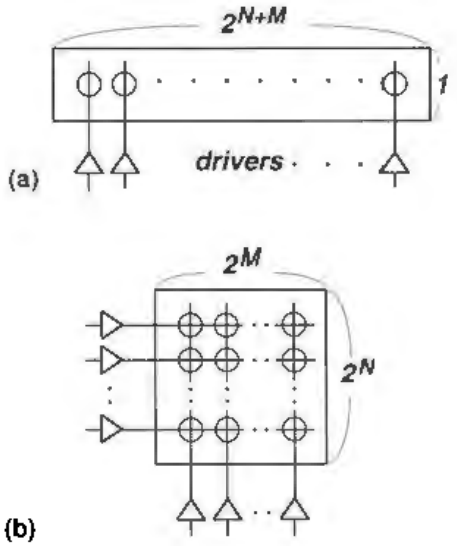


Fig. 1.5. Memory cell arrangements [1.9]. (**a**) One-dimensional; (**b**) two-dimensional

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

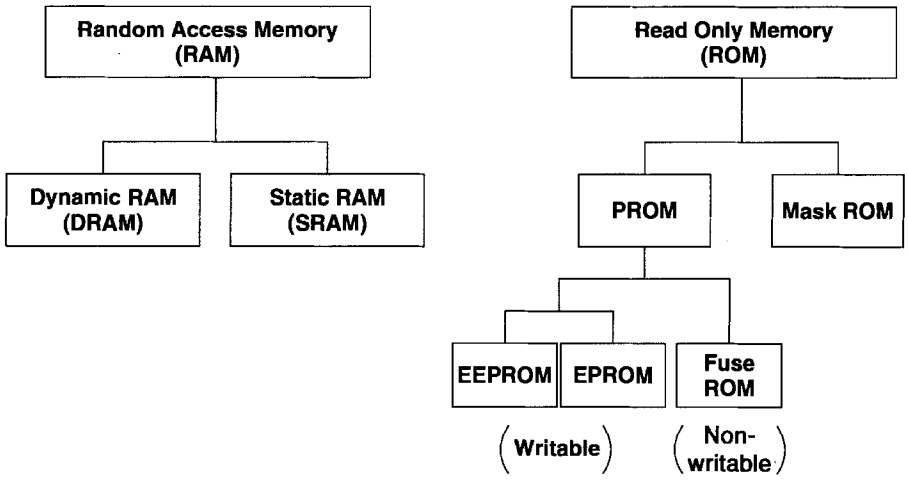


Fig. 1.6. Categories of semiconductor memories widely used [1.9]

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

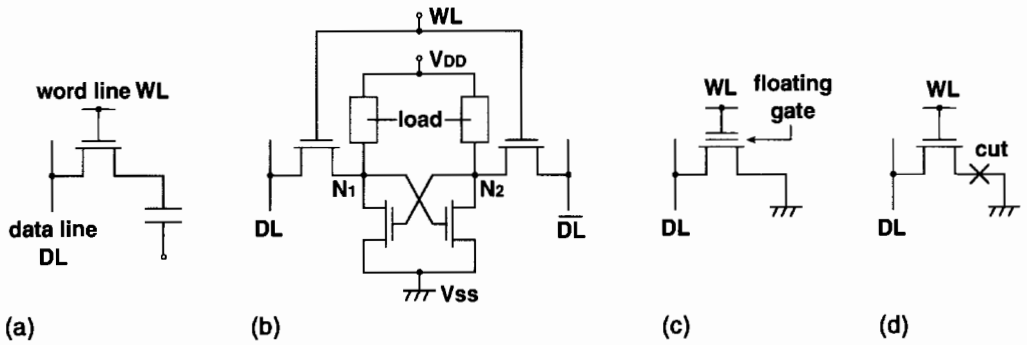


Fig. 1.7. Memory cell circuits [1.9]. (**a**) DRAM; (**b**) SRAM; (**c**) EPROM and EEPROM; (**d**) Mask ROM and Fuse ROM

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

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

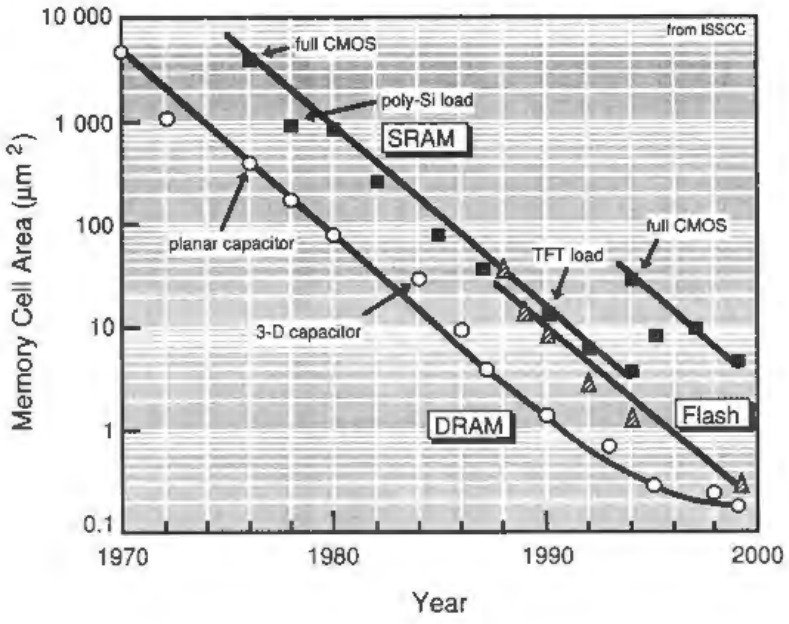


Fig. 1.8. Trends in the memory-cell area of VLSI memories [1.9, 1.44]

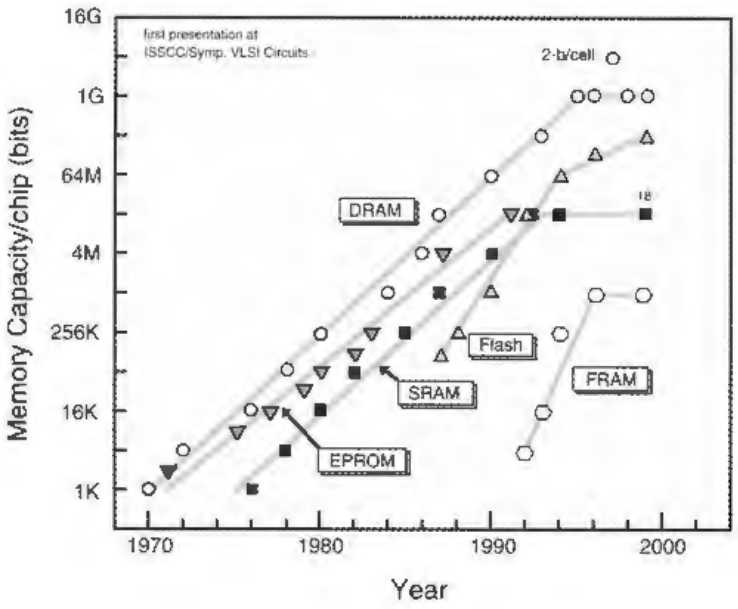


Fig. 1.9. Trends in the memory capacity of VLSI memories [1.9, 1.44]

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)

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.

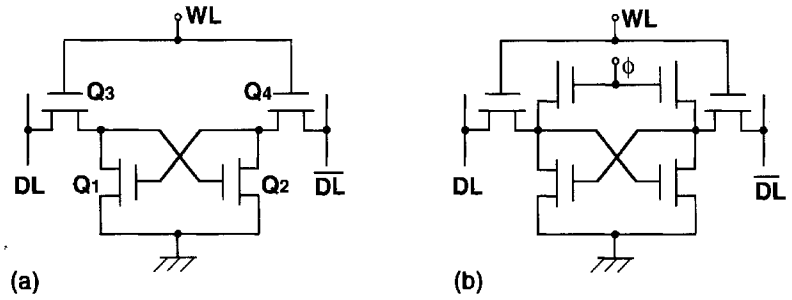


Fig. 1.10. The four-transistor cell [1.9]. (**a**) Standard; (**b**) charge pumping

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 $\overline{\text{DL}}$, 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

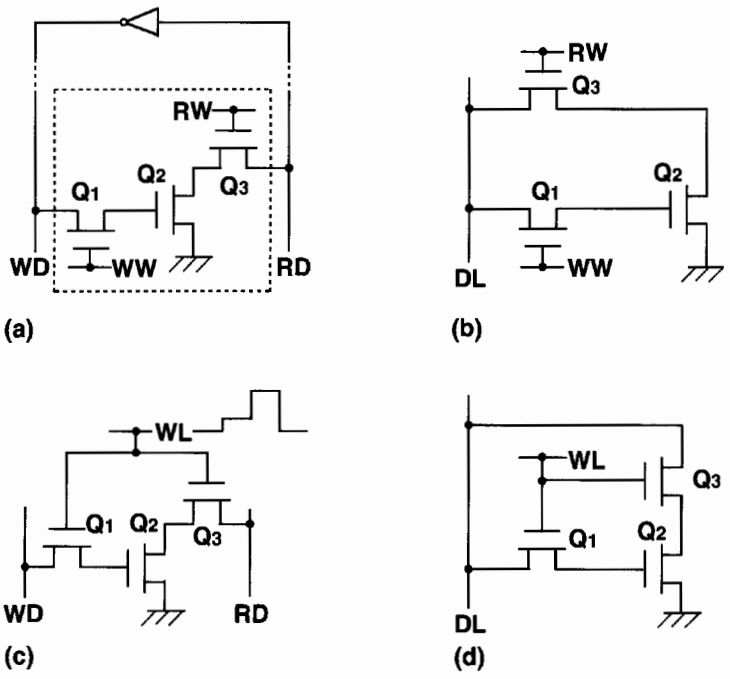


Fig. 1.11. The three-transistor cell [1.9]. (**a**) 4.5 lines; (**b, c**) 3.5 lines; (**d**) 2.5 lines

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 $2V_T$.

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.

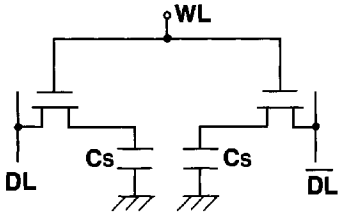


Fig. 1.12. The two-transistor cell (twin cell)

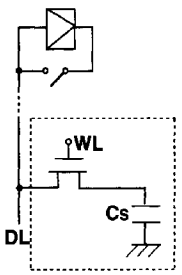


Fig. 1.13. The one-transistor cell

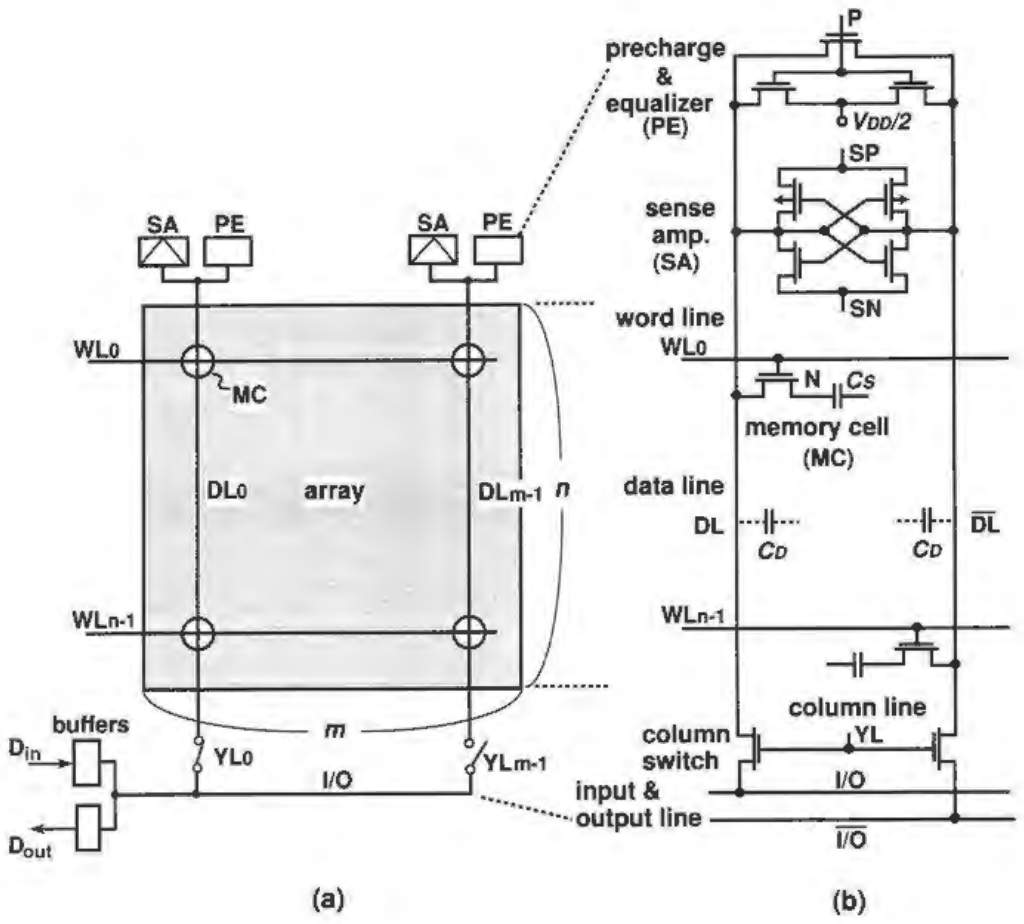


Fig. 1.14. A conceptual DRAM array (a), and an actual data-line configuration (b)

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 ($\pm\nu_S$) developed on the floating data line (for example, DL) is expressed by

$$\nu_S = \frac{V_{DD}}{2} \cdot \frac{C_S}{C_D + C_S} .$$

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 ($V_{DD}/2$, 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 ($\overline{DL}$) as a reference. Then, one of the amplified

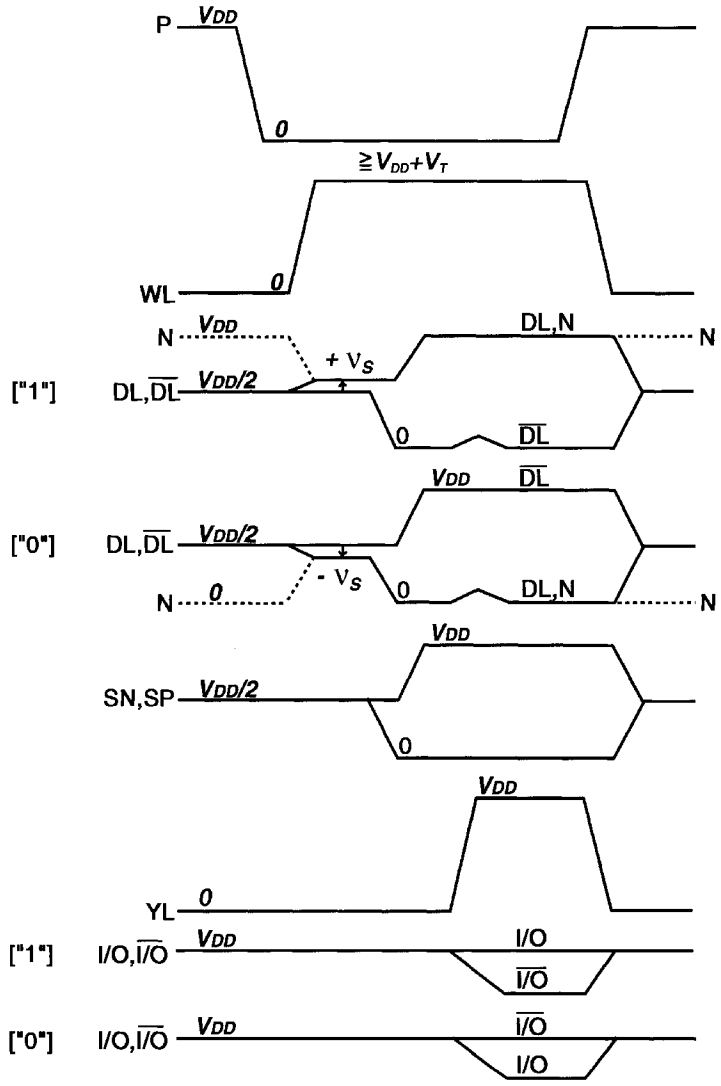


Fig. 1.15. The read operation

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

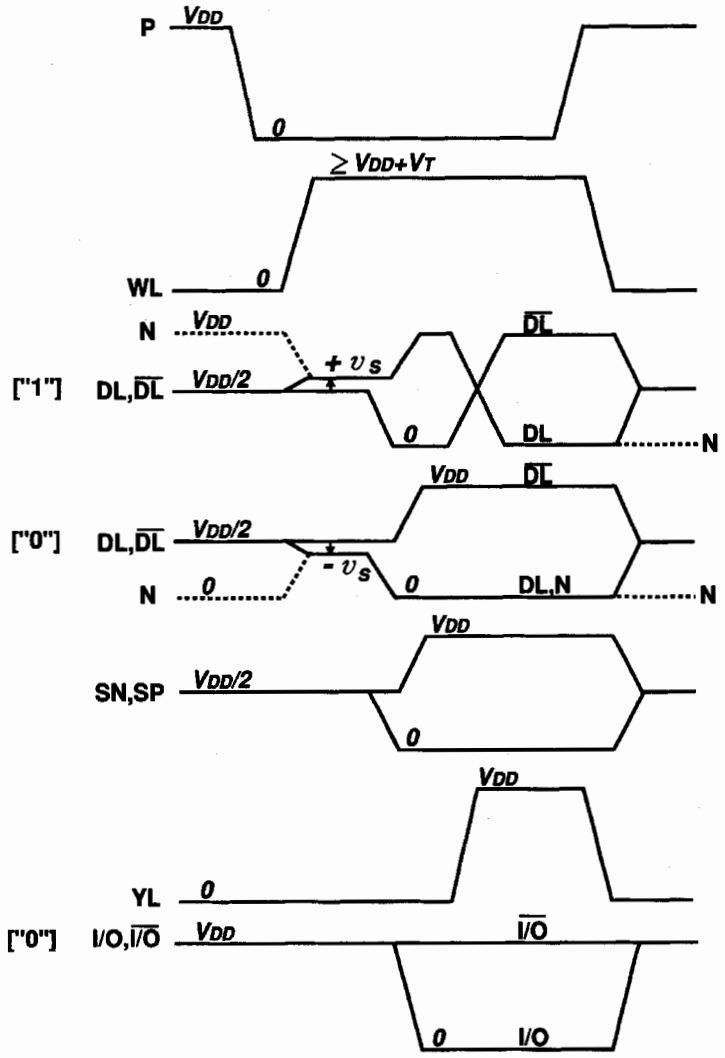


Fig. 1.16. The write operation for low-voltage ("0") binary information

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

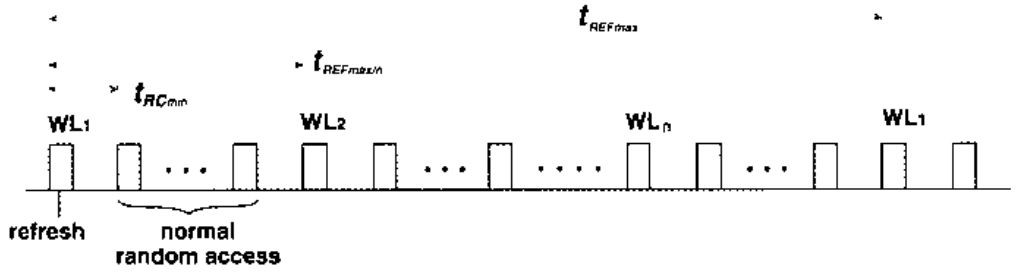


Fig. 1.17. The refresh operation. t_{RCmin}, minimum cycle time

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

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

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

Table 1.3. Advances in DRAM chip technology [1.9]

		64 Kbit	64 Mbit
Process		4in wafer	8in wafer
		3 μm NMOS	0.3 μm triple-well CMOS
		Double poly-Si,	Double poly-Si,
		single metal	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	1.3 μm^2, stacked C_S
		V_{DD} electrode	$V_{DD}/2$ electrode
	C_S/C_D	38 fF/515 fF	44 fF/250 fF, shielded DL
Circuit		–	Redundancy
		–	Multibit test (parallel test)
		V_{DD} precharge	$V_{DD}/2$ precharge
		Two-divided DL	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

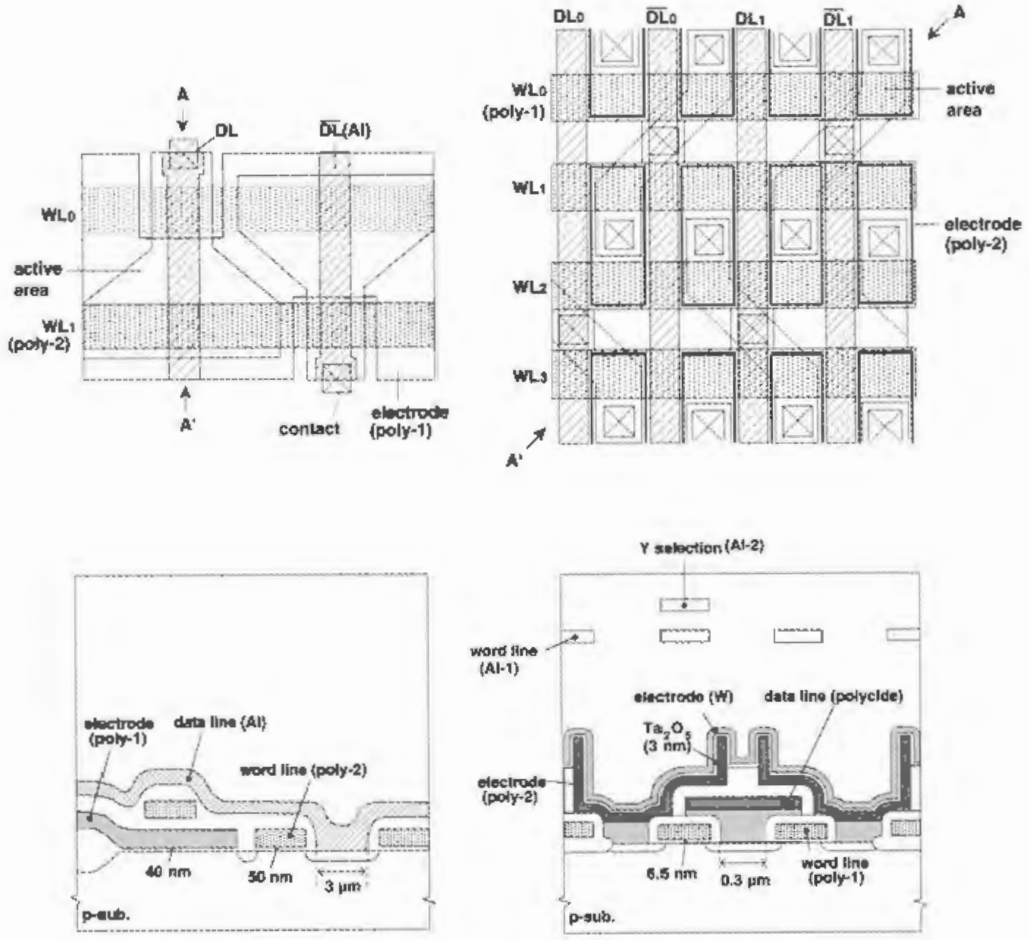


Fig. 1.18. A comparison between a 64 Kb memory cell (*left*) and a 64 Mb memory cell (*right*) [1.9]. *Top*, layout; *bottom*, cross-section A-A'. The word line Al-1, the Y-selection line Al-2, and the electrode W for the 64 Mb cell are omitted for ease of understanding

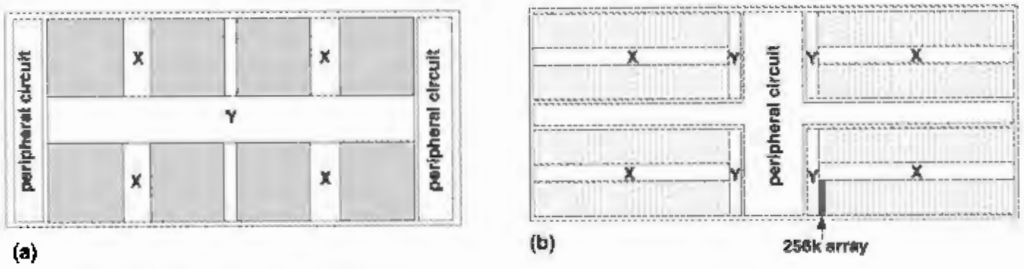


Fig. 1.19. A comparison between the configurations of a 64 Kb chip (**a**) and a 64 Mb chip (**b**) [1.9]

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.

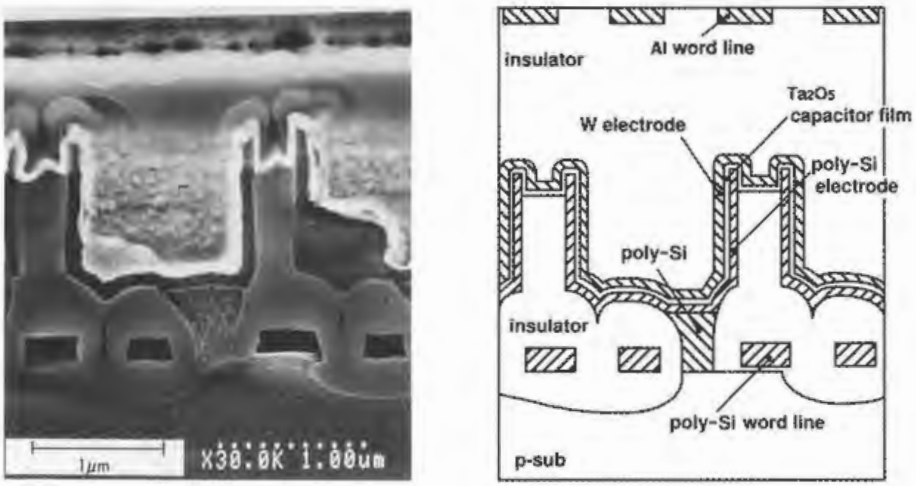


Fig. 1.20. An actual cross-section of a 64 Mb memory cell [1.9]

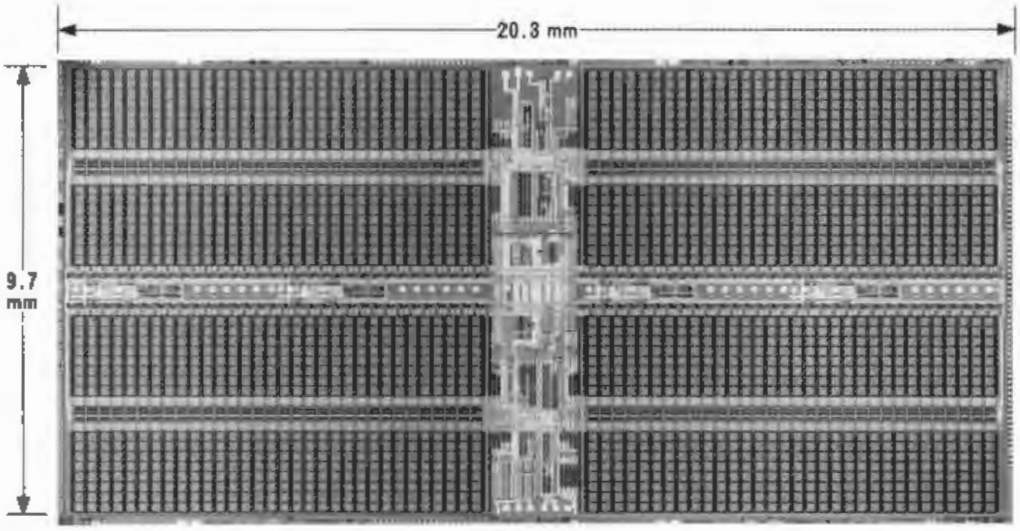


Fig. 1.21. A microphotograph of the 64 Mb DRAM chip in Fig. 1.19 [1.28]

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)

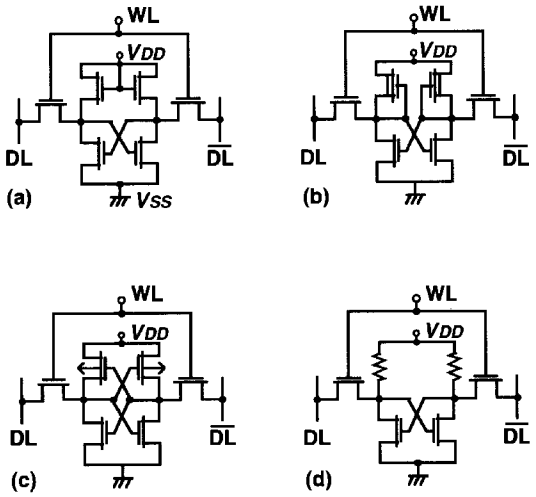


Fig. 1.22. Flip-flop SRAM memory cells [1.9]. (**a**) Enhancement MOS load; (**b**) depletion MOS load; (**c**) PMOS load (full CMOS); (**d**) polysilicon load

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

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, v_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

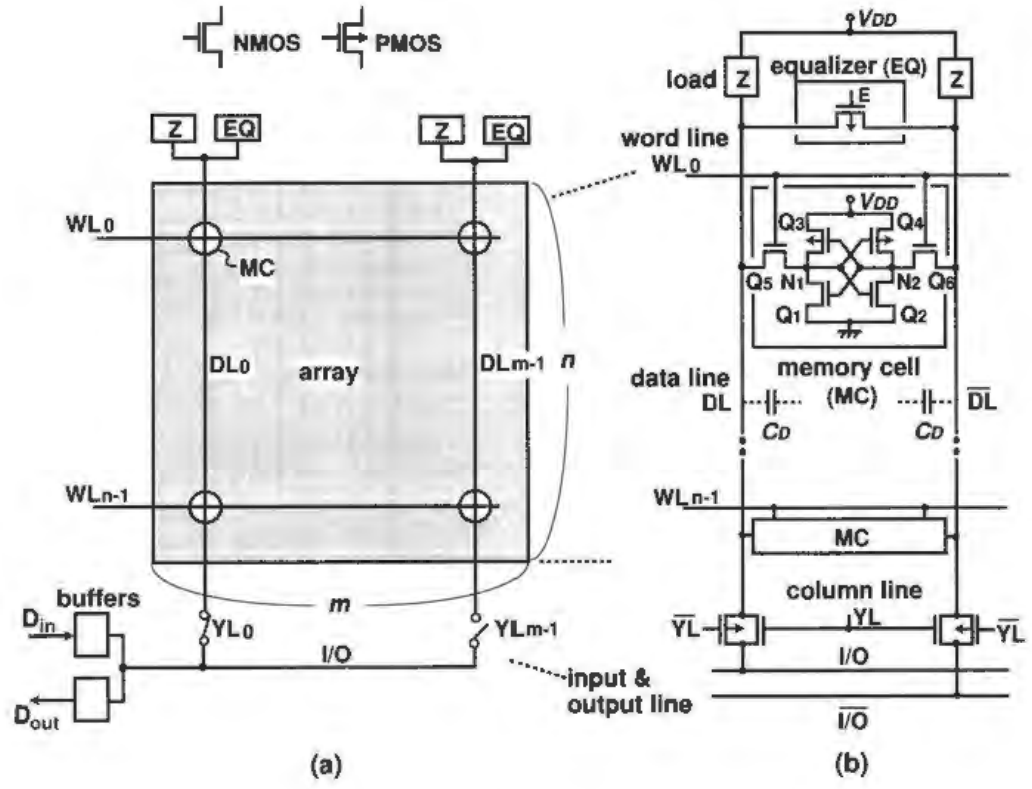


Fig. 1.23. A conceptual SRAM array (**a**) and an actual data-line configuration (**b**)

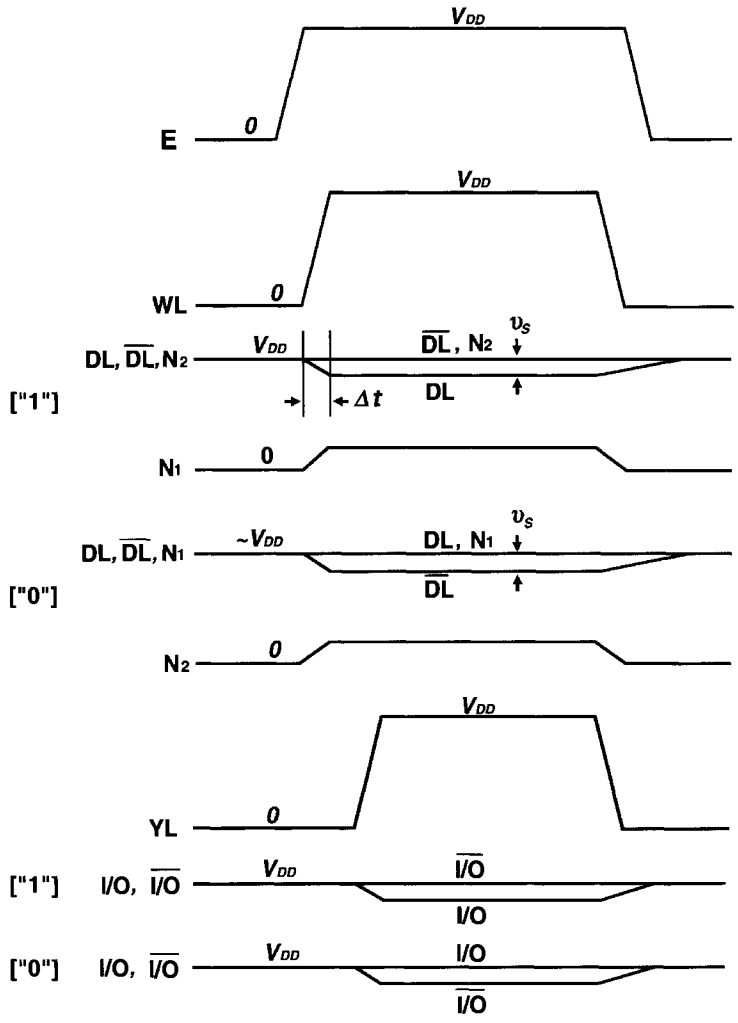


Fig. 1.24. The read operation

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

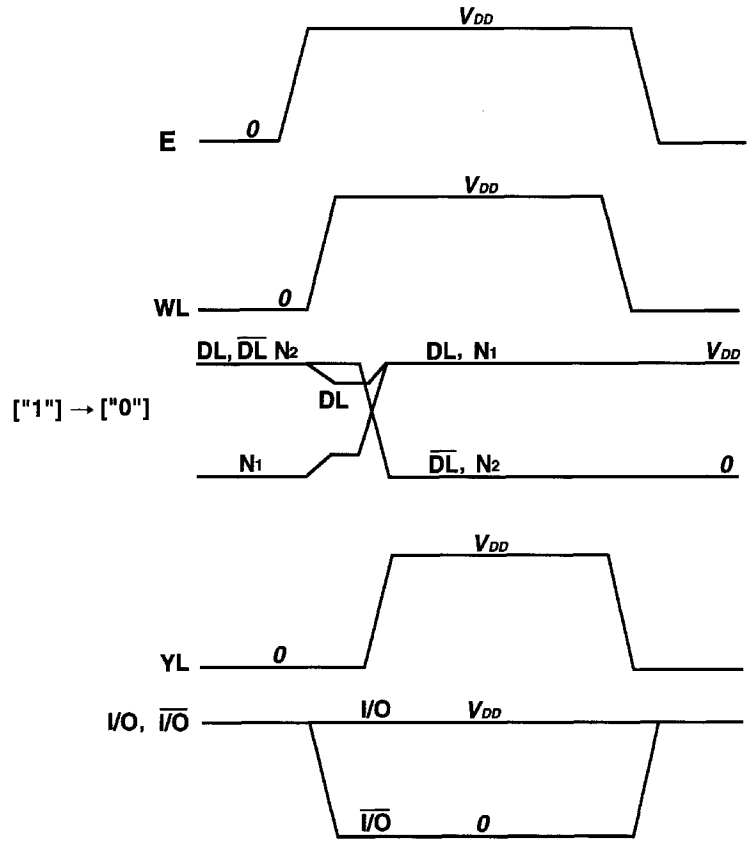


Fig. 1.25. The write operation

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

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

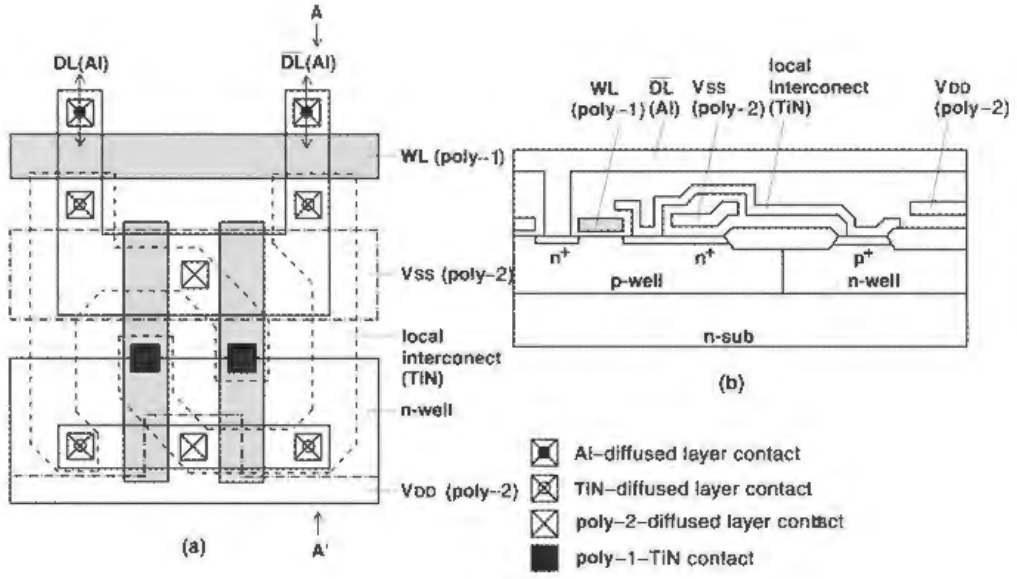


Fig. 1.26. A schematic of a full-CMOS SRAM-cell layout (**a**) and a cross-section along line A-A′ (**b**) [1.31]. The Al wirings for data lines in (**a**) are omitted for ease of understanding

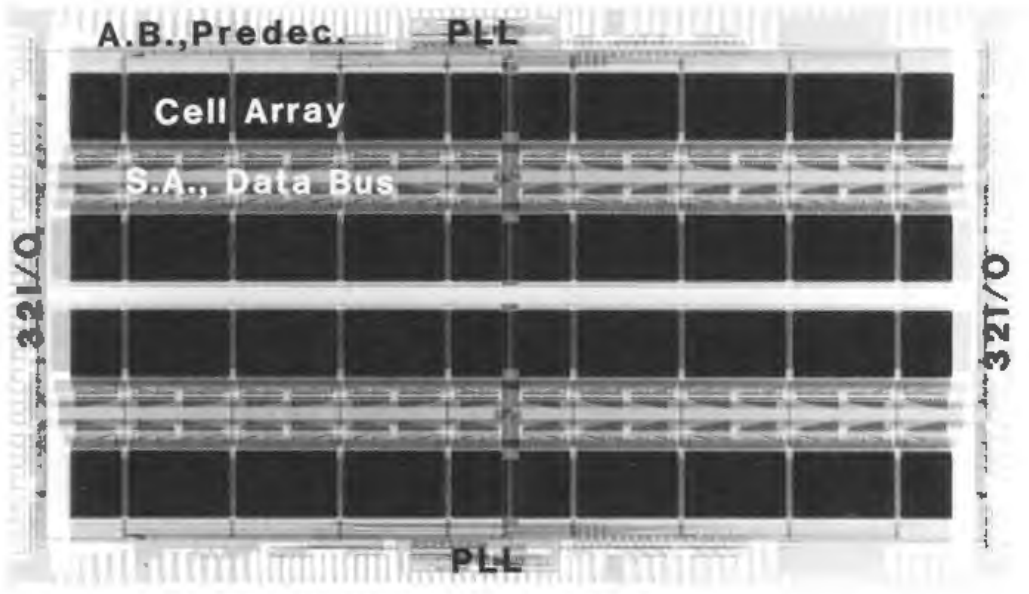


Fig. 1.27. A microphotograph of a 4 Mb CMOS SRAM [1.32]

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 $8 \times 14.1\,\text{mm}^2$, 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

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.

(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.

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,

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

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

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

$$C_{\mathrm{T}} = C_{\mathrm{FC}} + C_{\mathrm{S}} + C_{\mathrm{D}} + C_{\mathrm{B}} \,.$$

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

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

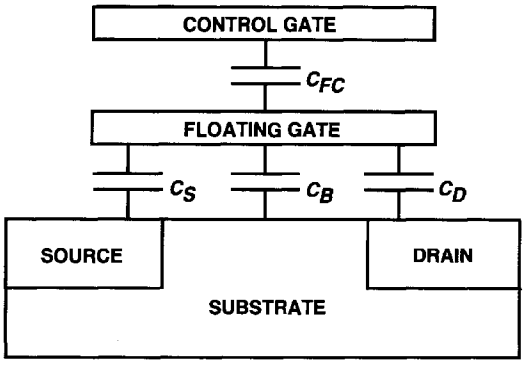


Fig. 1.28. A schematic of the cross-section of a Flash memory cell [1.35]

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

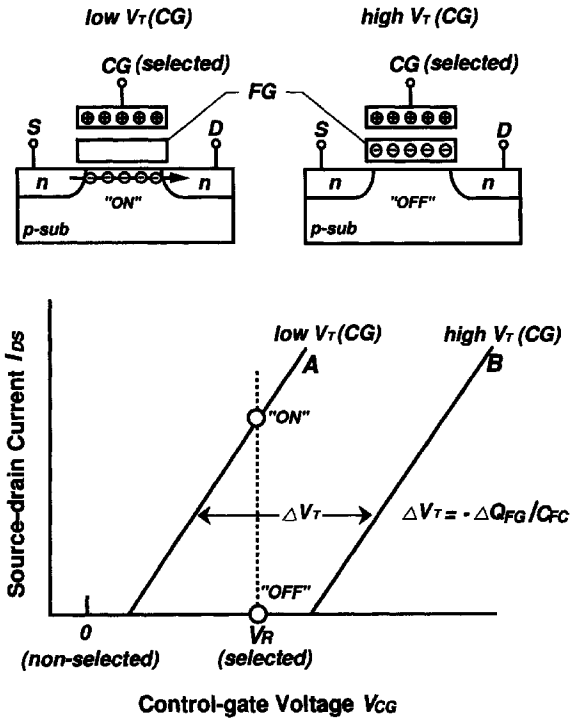


Fig. 1.29. The principle of the read operation of a Flash memory cell [1.36]

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

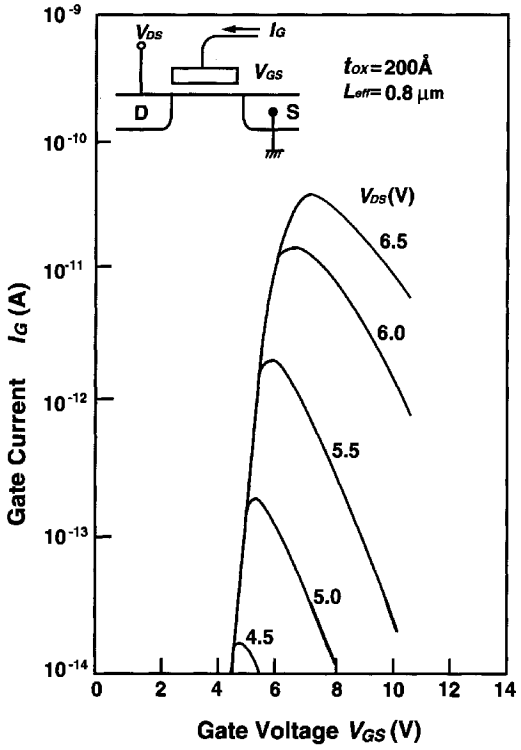


Fig. 1.30. The gate current caused by channel hot-electron injection [1.35]

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

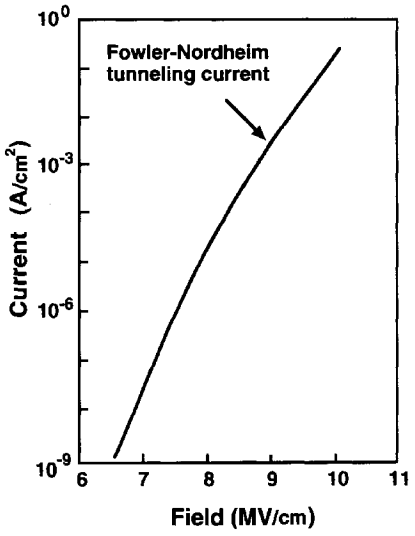


Fig. 1.31. The Fowler-Nordheim tunneling current, as a function of the electric field [1.35]

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

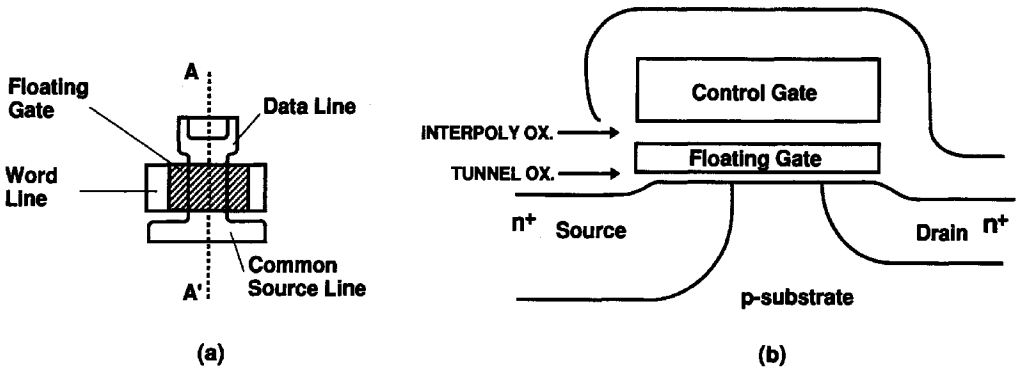


Fig. 1.32. The NOR cell [1.37]. (**a**) The layout of a typical double-polysilicon stacked gate cell; (**b**) a schematic cross-section along line A-A′

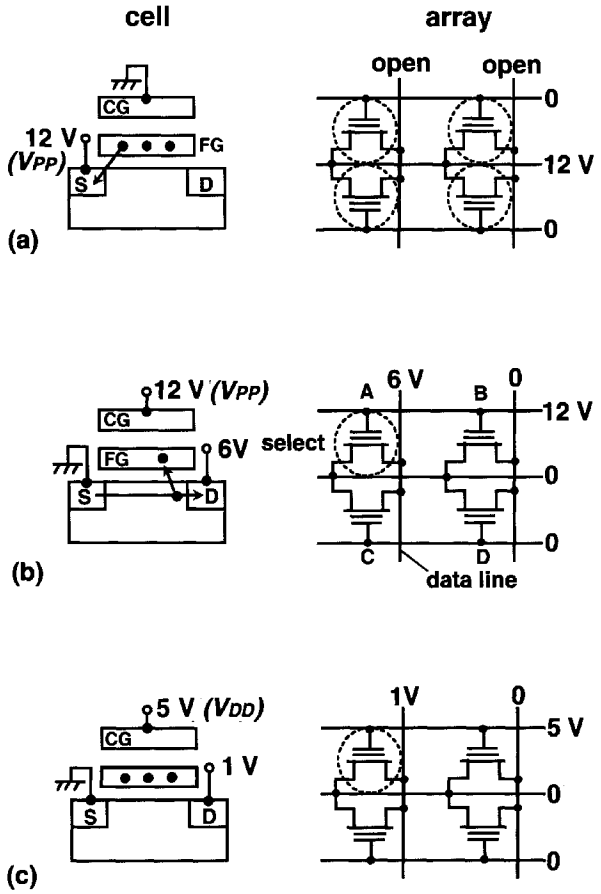


Fig. 1.33. The basic operation of the NOR cell [1.9]. (**a**) Erase; (**b**) write; (**c**) read

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"),

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

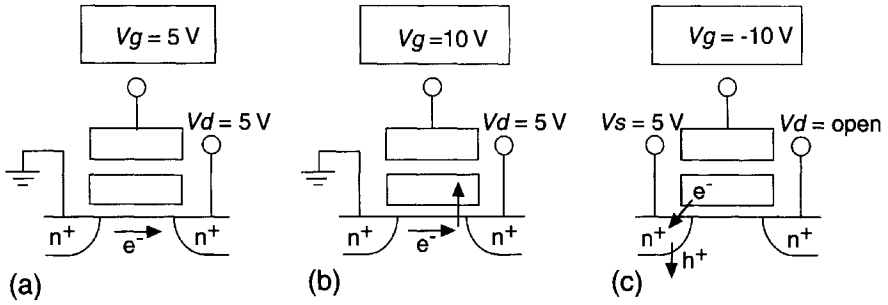


Fig. 1.34. The negative word-line voltage scheme [1.38]. (**a**) Read, 60 μA/cell; (**b**) program, 0.5 mA/cell (8 mA for 16 I/Os); (**c**) erase, 10 nA/cell (5 mA for 64 Kb block)

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

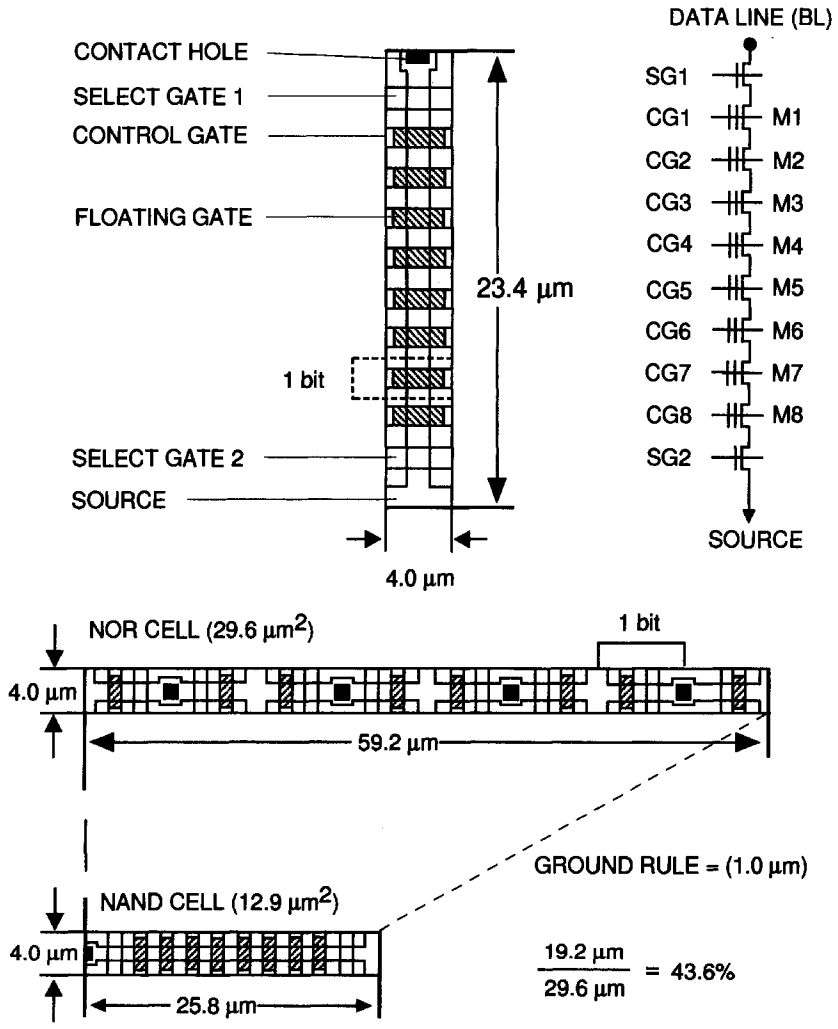


Fig. 1.35. NAND architecture. The dimensions of a NAND array are compared to those of a NOR array [1.39]

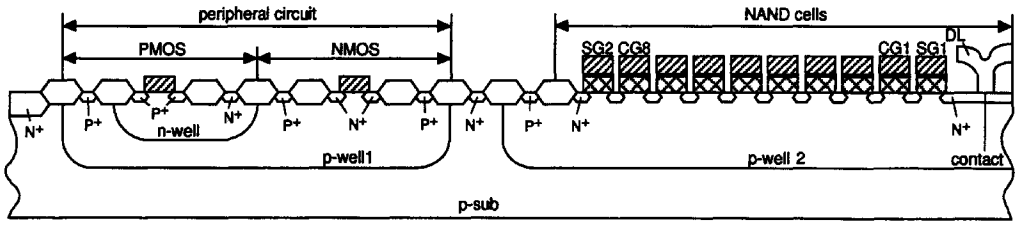


Fig. 1.36. A cross-section of NAND cells and the peripheral circuit [1.39]

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

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:

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

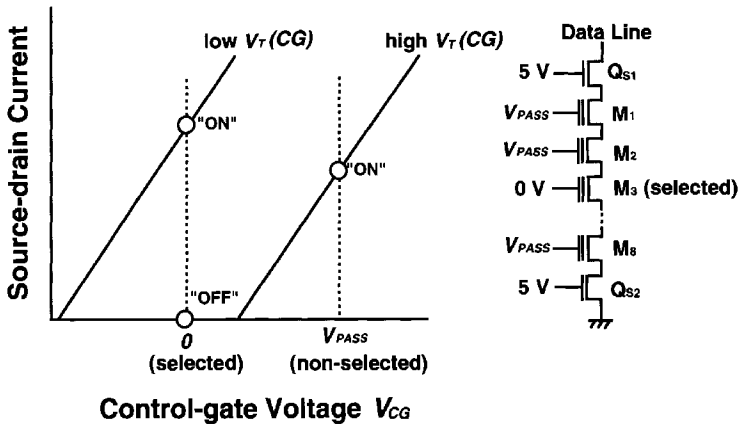


Fig. 1.37. The principle of the read operation of the NAND cell [1.39]

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

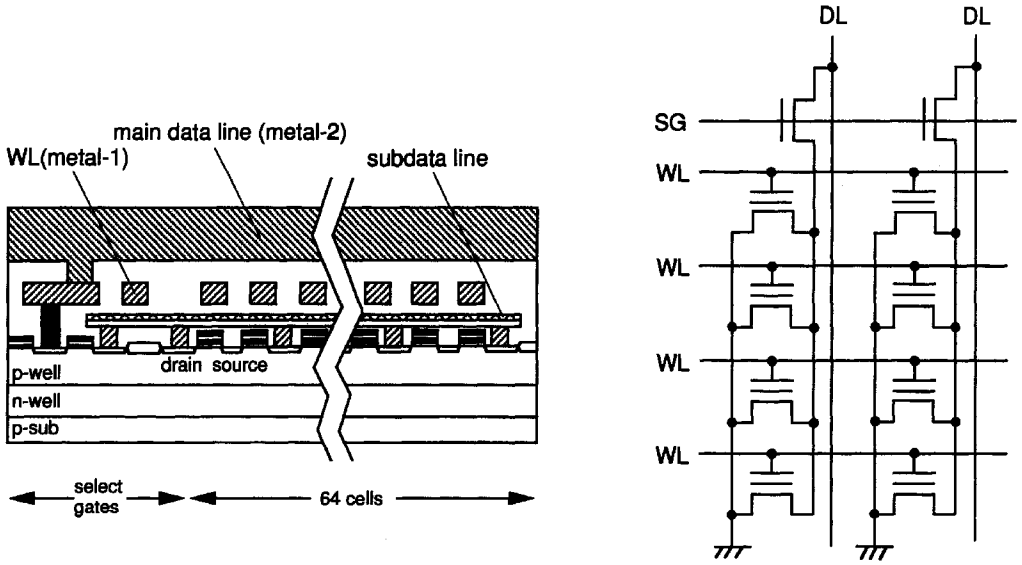


Fig. 1.38. The cross-section and equivalent circuit of DINOR cells [1.40]

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

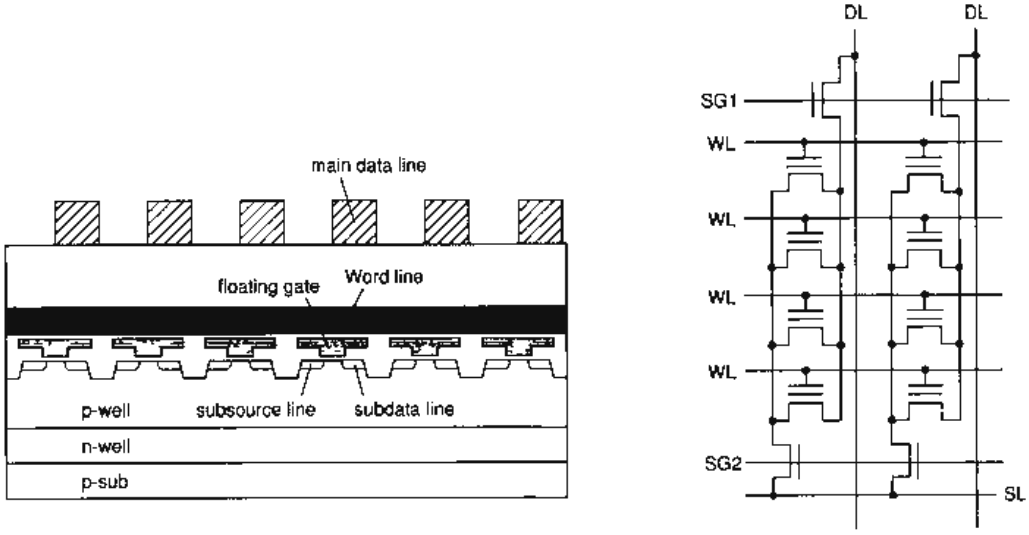


Fig. 1.39. The cell structure and equivalent circuit of an AND cell [1.41]

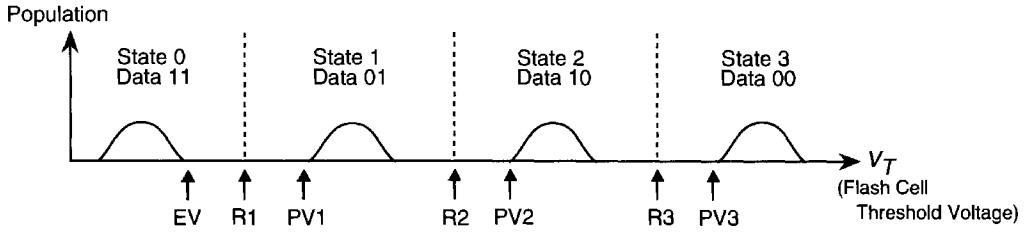


Fig. 1.40. The multilevel-cell threshold voltage distribution for 2 bits/cell storage [1.42]. The states (ranges of threshold voltages) are determined by Program Verify Reference V_T (PV) and Erase Verify Reference V_T (EV). Flash cells are used to generate the reference V_T values to compare the V_T 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.

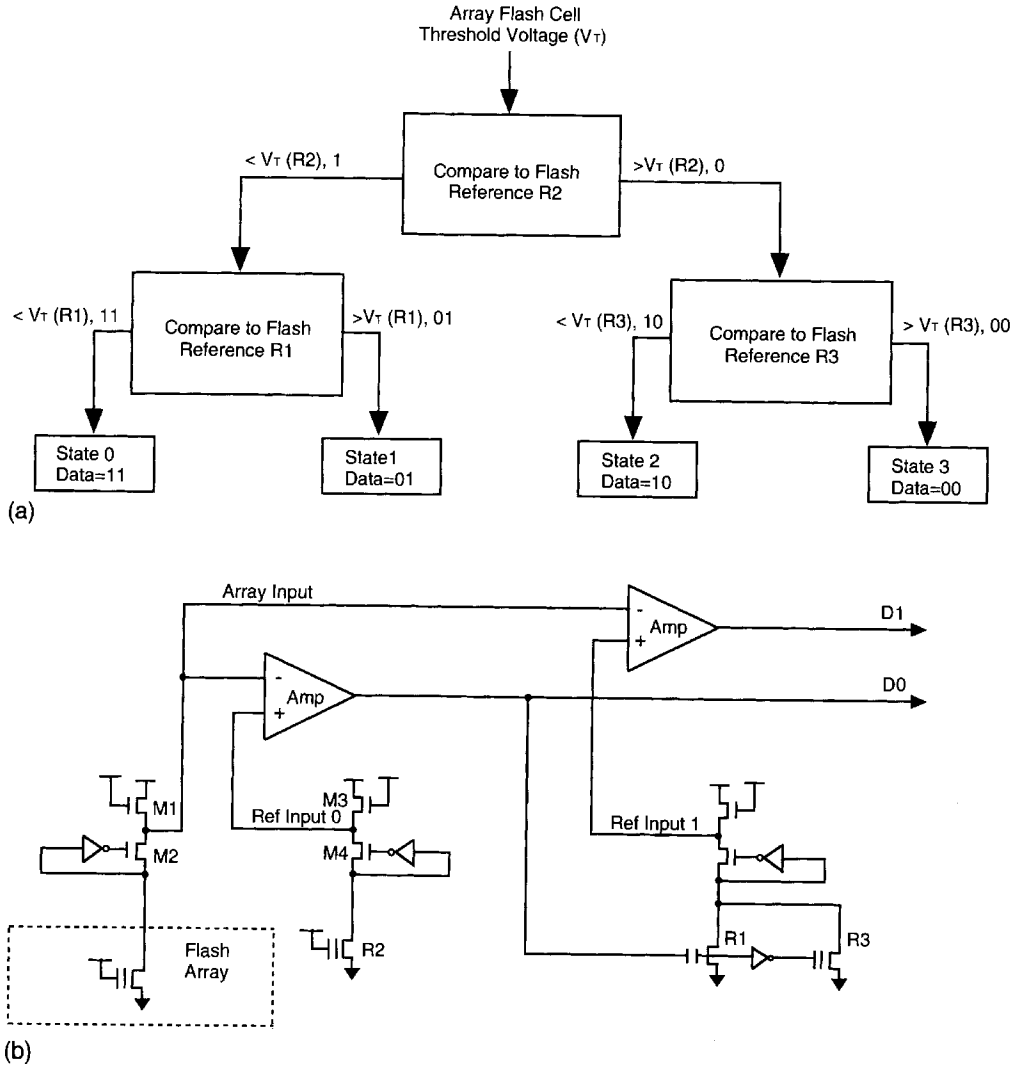


Fig. 1.41. The binary search sensing scheme (**a**) and its circuit (**b**) for 2 bits/cell storage [1.42]

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.

2. The Basics of RAM Design and Technology

2.1 Introduction

Advances in memory chip technology have been supported by extensive technologies, such as high-density, high-speed device and circuit technology, lithography and fabrication process technology, and high-density packaging technology. In this chapter, the basics of some of the above supporting technologies are discussed.

First, various devices for RAM chips, such as MOSFETs, capacitors, resistors, wiring and wiring materials, and silicon substrates, are discussed. Next, basic MOS circuits, such as the NMOS static circuit, the NMOS dynamic circuit, and the CMOS circuit, are described, followed by the basic memory circuits that are widely used in modern RAMs. Third, a device scaling law that gives a guiding principle for the performance of scaled-down devices is explained. Then, lithography technology is briefly discussed, followed by packaging technology.

2.2 Devices

2.2.1 MOSFETs

Basic Structure and Operation. All of the elements in a RAM chip are integrated on to a silicon substrate, which ensures mechanical strength and works as an ac ground because of a large self-capacitance. Figure 2.1 shows a schematic cross-section [2.1] of an N-channel MOSFET (NMOS). Two symmetric n^+ regions, whose doping concentration and depth are around $10^{20}\,cm^{-3}$ and $0.1-0.5\,\mu m$, are formed with ion-implantation on a p-type silicon substrate. The drain (D) and source (S) are interchangeable, depending on their voltage conditions. In an NMOS, one n^+ region at a higher voltage is called the drain, while the other is called the source. Therefore, the drain (or source) can change to a source (or drain) during operation. The surface of the substrate between the drain and the source is called the channel. Conduction from the source to the drain is controlled by the gate (G). The gate is usually made of poly-silicon (poly-Si) and is isolated by a thin layer of silicon dioxide (SiO_2), the so-called gate oxide, from the channel. To isolate the MOSFET

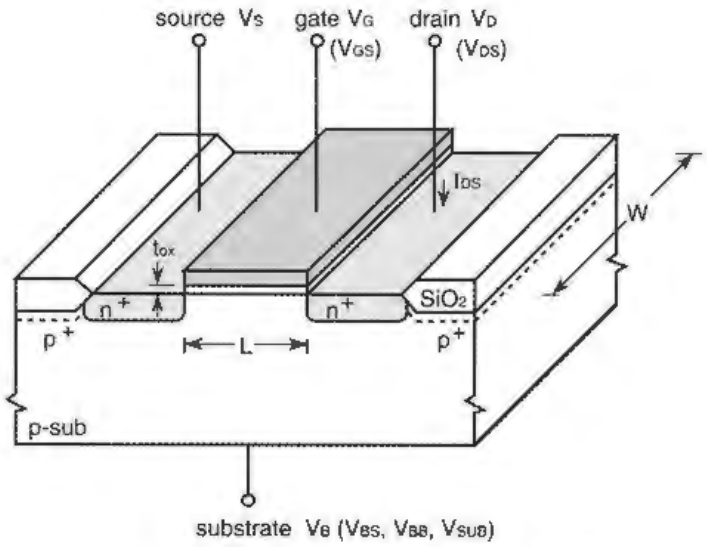


Fig. 2.1. A cross-section of an NMOSFET [2.1]

from adjacent devices, thick isolation layers composed of SiO_2 are formed. The p^+ boron (B) layers of about $10^{18}\,cm^{-3}$ implanted beneath the isolation layers improve the isolation capability.

The drain, gate, and substrate voltages measured from the source are called V_{DS}, V_{GS}, and V_{BS} (i.e. backgate bias or substrate bias V_{BB}). V_{BB} is usually negative or 0 V, and is supplied directly from an external source, or is internally generated from V_{DD} by means of an on-chip voltage converter (or V_{BB} generator). When V_{GS} is increased the drain–source current I_{DS} starts to flow at a certain V_{GS}. This value of V_{GS} is called the MOS threshold voltage V_T. I_{DS} increases with V_{GS} above V_T.

There are basically four different types of MOSFET, depending on the type of channel carrier [2.2]. The MOSFET with an n^+ source and drain, shown in Fig. 2.1, is called an N-channel MOSFET (NMOS, NMOST, or NMOSFET). In the NMOSFET, electrons flow from source to drain through an n-channel; that is, a current I_{DS} flows from drain to source. If $I_{DS} = 0$ at $V_{GS} = 0$, we must apply a positive V_{GS} to form the n-channel. This type is known as the enhancement (normally off) NMOSFET (the E-NMOSFET). If an n-channel exists at $V_{GS} = 0$, we must apply a negative V_{GS} to deplete carries in the channel. This type is called the depletion (normally on) NMOSFET (the D-NMOSFET). Similarly, we have the p-channel enhancement (normally off) MOSFET (the E-PMOSFET) and the depletion (normally on) MOSFET (the D-PMOSFET). The transfer characteristics of the four types are shown in Fig. 2.2 [2.1]. Note that for the enhancement MOSFET, the NMOSFET takes positive values for both V_{DS} and V_T, while the PMOSFET takes negative values for these voltages. Figure 2.3 shows circuit expressions and voltage relationships for the E-NMOSFET [2.1], in which V_{DD} and V_{SS}

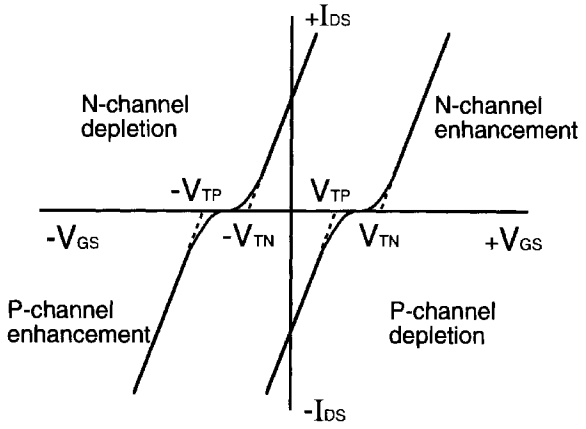


Fig. 2.2. The transfer characteristics of four types of MOSFET [2.1]

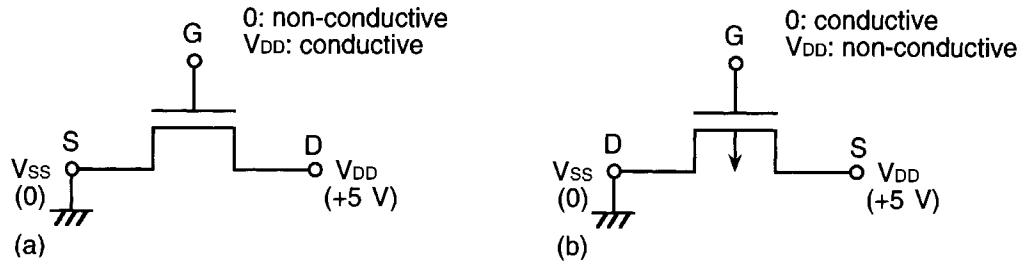


Fig. 2.3. Circuit expressions and voltage relationships for an E-MOSFET [2.1]. (a) NMOS; (b) PMOS

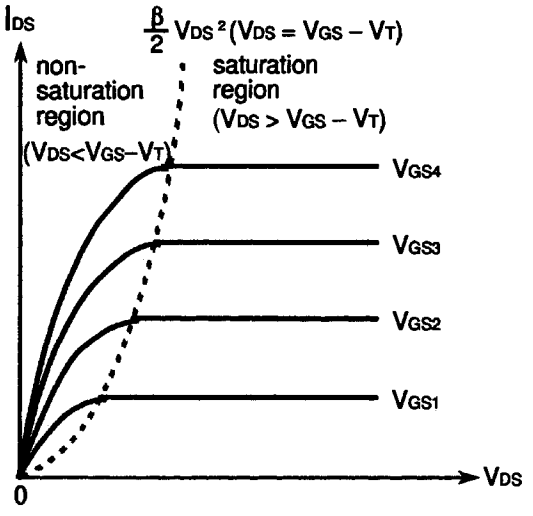


Fig. 2.4. I_{DS} versus V_{DS} for a given V_{BB} [2.1]

are external supply voltages. In the following there are dc characteristics of E-NMOSFET [2.3–2.5].

Figure 2.4 shows I_{DS} versus V_{DS} for different values of V_{GS} at a given V_{BB}. Three regions of operation can be distinguished, as follows [2.3].

The Cutoff Region. This region corresponds to $V_{GS} < V_T$ and $I_{DS} \simeq 0$. It is also referred to as the subthreshold region, where I_{DS} in this region is much

smaller than its value when $V_{GS} > V_T$. Thus, in many NMOS circuits for $V_{GS} < V_T$, the transistor is considered off and $I_{DS} = 0$. However, the small value of I_{DS} (i.e. the subthreshold current) in the region could affect the performance of the circuit, as in MOS dynamic circuits. In ultra-low-voltage operations with a lower V_T, the ever-increasing subthreshold current is an emerging issue, as discussed in detail in Chap. 8.

The Non-Saturated Region. This region is also referred to as the "triode region", in which I_{DS} increases with V_{DS} for a given V_{GS} larger than V_T. The current is given by

$$\begin{aligned} V_{GS} &\geq V_T \ , \\ V_{DS} &= V_{GS} - V_T \ , \\ I_{DS} &= \beta \left[(V_{GS} - V_T) V_{DS} - \tfrac{1}{2} V_{DS}^2 \right] \end{aligned} \tag{2.1}$$

where β is the channel conductance, which is explained later. When $V_{DS} \ll V_{GS} - V_T$, the I–V charasteristics can be approximated by a straight-line, $I_{DS} = \beta(V_{GS} - V_T)V_{DS}$, and the transistor is modeled by a resistor of

$$R_{ON} = \frac{1}{\beta(V_{GS} - V_T)} \ . \tag{2.2}$$

The Saturated Region. In this region I_{DS} becomes constant, independently of V_{DS}. It is determined only by V_{GS}, which is given by

$$\begin{aligned} V_{GS} &\geq V_T \ , \\ V_{DS} &= V_{GS} - V_T \ , \\ I_{DS} &= \tfrac{\beta}{2}(V_{GS} - V_T)^2 \ . \end{aligned} \tag{2.3}$$

The conductance β is given by

$$\beta = \frac{W}{L} \frac{\varepsilon_{OX} \mu}{t_{OX}} = \frac{W}{L} \beta_0 = \frac{W}{L} \mu C_{OX} \ , \tag{2.4}$$

where β_0 is the conduction factor, ε_{OX} is the permittivity of the gate oxide, t_{OX} is the thickness of the gate oxide, μ is the average surface mobility of carriers (μ_n in the case of electrons in NMOS, and μ_p in the case of holes in PMOS), C_{OX} is the gate capacitance per unit area, L is the channel length between the n^+ source and drain edges, which has a typical value of 0.1–0.5 μm (the metallurgical length of the gate is the gate length L_g, which is usually larger than L), and W is the channel width perpendicular to L.

The conduction factor β_0 is technology-dependent and is specified for a given MOS process. Thus it is not a circuit design variable. An NMOSFET has a larger value of β_0 compared with a PMOSFET because $\mu_n = (2.5 - 3.0)\mu_p$. The factor β_0 is a function of temperature, and is given by

$$\beta_0 / \beta_0' = (T/T_0)^{-3/2} \tag{2.5}$$

where β_0' is the value of β_0 at room temperature ($T = 298\,\mathrm{K}$) and T is the absolute temperature (K). The ratio W/L is a circuit design parameter. The minimum value of L ($= L_{\min}$) is determined by the MOS fabrication process, mainly by the mask channel length, the tolerances on that length, and the lateral diffusions of both the source and drain regions. The minimum value of W is usually of the order of $L_{\min}$. Increasing W/L increases I_{DS} for a given set of operating voltages. However, increasing W increases the gate, source, and drain diffusion areas, and thus increases the gate capacitance, and the source–substrate and drain–substrate junction capacitances.

The threshold voltage V_{T} is given by

$$V_{\mathrm{T}} = V_{\mathrm{T0}} + \Delta V_{\mathrm{T}}(V_{\mathrm{BB}}) \,, \tag{2.6a}$$

$$V_{\mathrm{T0}} = V_{\mathrm{FB}} + K\sqrt{2\psi} + 2\psi \,, \tag{2.6b}$$

$$\Delta V_{\mathrm{T}}(V_{\mathrm{BB}}) = K\sqrt{|V_{\mathrm{BB}}| + 2\psi} - \sqrt{2\psi} \,, \tag{2.6c}$$

$$K = \sqrt{2\varepsilon_{\mathrm{S}} q N}/C_{\mathrm{OX}} \,, \quad \psi = (kT/q)\ln(N/n_i) \tag{2.6d}$$

where V_{T0} is V_{T} for $V_{\mathrm{BB}} = 0$, V_{FB} is the flat-band voltage, ψ is the Fermi potential, N is the substrate doping concentration, n_i is the intrinsic carrier density of silicon, ε_{S} is the permittivity of silicon, q is the magnitude of electric charge, and k is the Boltzmann constant. The body-effect coefficient or substrate-bias effect constant K, which is about $0.1\text{–}1.0\,\mathrm{V}^{1/2}$, represents the sensitivity of V_{T} for V_{BB}. V_{T} decreases linearly, as given by

$$V_{\mathrm{T}}(T) = V_{\mathrm{T}}(0) - a(T - T_0) \,, \tag{2.7}$$

where $V_{\mathrm{T}}(0)$ is for room temperature ($T_0 = 298\,\mathrm{K}$), and a is $0.5\text{–}5.0\,\mathrm{mV/K}$. Note that the temperature dependence of I_{DS} tends to be canceled with reductions in β_0 and V_{T} as the temperature increases.

Table 2.1 shows an example of $0.5\,\mu\mathrm{m}$ MOSFET characteristics [2.1].

Table 2.1. The characteristics of the $0.5\,\mu\mathrm{m}$ MOSFET [2.1]

	NMOS	PMOS
t_{OX}	$15\,\mathrm{nm}$	$15\,\mathrm{nm}$
L	$0.6\,\mu\mathrm{m}$	$0.6\,\mu\mathrm{m}$
β_0^a	$115\,\mu\mathrm{S/V}$	$34\,\mu\mathrm{S/V}$
V_{T}^b	$0.5\,\mathrm{V}$	$0.5\,\mathrm{V}$

[a] $\beta_0 = \varepsilon_0\varepsilon_{\mathrm{OX}}\mu/t_{\mathrm{OX}}$, $\mu_{\mathrm{n}} = 500\,\mathrm{cm}^2/\mathrm{Vs}$, $\mu_{\mathrm{p}} = 150\,\mathrm{cm}^2/\mathrm{Vs}$, $\varepsilon_0 = 8.854 \times 10^{-12}\,\mathrm{F/m}$, $\varepsilon_{\mathrm{OX}} = 3.9$.

[b] V_{GS} at $I_{\mathrm{DS}} = 10\,\mathrm{nA}$, $W = 15\,\mu\mathrm{m}$.

Increases in V_{T} in Source–Follower Mode. When the source voltage, and thus source–substrate voltage, change, V_{T} changes even at a fixed V_{BB}, as exemplified in Fig. 2.5a. V_{T} increases as node A – which is the source in

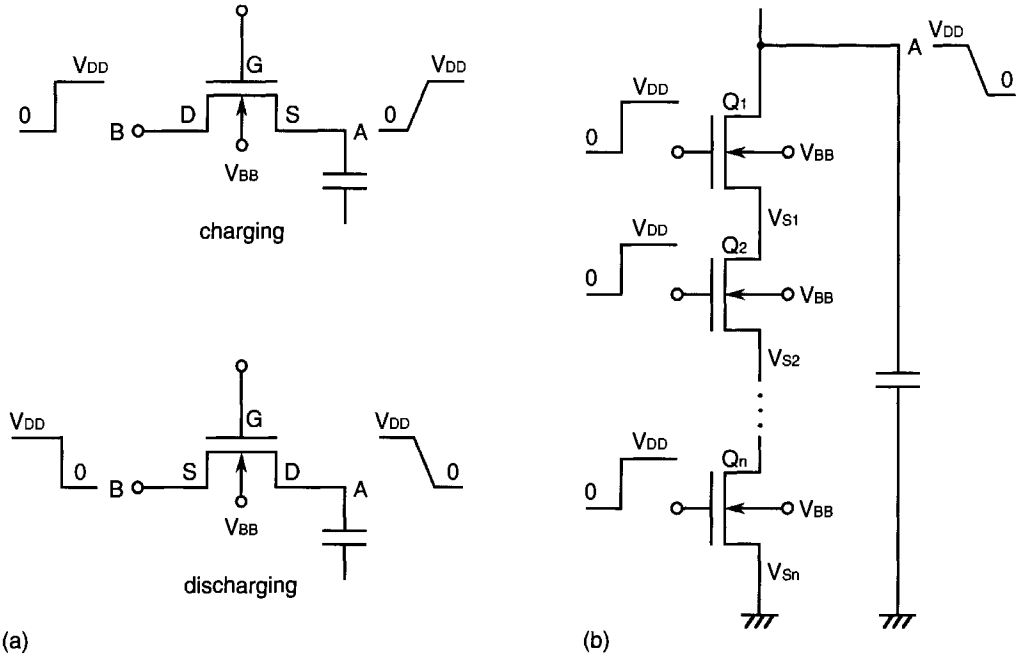


Fig. 2.5. Examples of source-follower mode. (**a**) Charging and discharging of node A; (**b**) discharging of node A by stacked NMOSFETs

this case – is charged up by applying a sufficiently high voltage to the gate and V_{DD} to node B of an NMOSFET. Finally, V_T becomes a maximum at a source–substrate voltage of $V_{BB} + V_{DD}$, as follows:

$$V_{Tmax} = V_{T0} + K\left(\sqrt{|V_{BB}| + V_{DD} + 2\psi} - \sqrt{2\psi}\right) . \tag{2.8}$$

On the contrary, when node A is discharged from V_{DD} to 0 V by fixing node B (which is in turn the source), V_T becomes a minimum, as follows:

$$V_{Tmin} = V_{T0} + K\left(\sqrt{|V_{BB}| + 2\psi} - \sqrt{2\psi}\right) . \tag{2.9}$$

The difference is thus

$$V_{Tmax} - V_{Tmin} = K\left(\sqrt{|V_{BB}| + V_{DD} + 2\psi} - \sqrt{|V_{BB}| + 2\psi}\right) . \tag{2.10}$$

For example, the difference is as large as 0.67 V at $V_{BB} = 0$ V for $V_{DD} = 5$ V, $K = 0.3\,V^{1/2}$ and $2\psi = 0.6$ V. If V_{Tmin} is set to be 0.5 V, as for the usual 5 V V_{DD} designs, V_{Tmax} is 1.17 V. Hence, to charge node A to V_{DD}, with elimination of the reduction in V_T, the gate voltage must be higher than 6.17 V ($= V_{DD} + V_{Tmax}$), giving the gate oxide a high stress voltage. A higher $|V_{BB}|$ reduces the difference (e.g. 0.31 V at $|V_{BB}| = 3$ V), enabling a lower gate voltage (e.g. 5.81 V).

The discharging of node A by stacked MOSFETs, shown in Fig. 2.5b, is another example. During discharging, the V_T values of all NMOSFETs except Q_n rise due to their instantaneous source voltages. In particular, the V_T of the upper MOSFET Q_1 is maximized at the beginning of discharging,

because $V_{S1} > V_{S2} > \cdots > V_{Sn}$. The raised V_T eventually limits the number of stacked NMOSFETs for high speed.

Channel Length Modulation. The I_{DS} equation at saturation is derived by simply assuming that I_{DS} at saturation does not change when V_{DS} increases, resulting in a zero slope of the I_{DS} versus V_{DS} characteristics. In the actual MOSFET characteristics [2.1] shown in Fig. 2.6, however, I_{DS} increases slightly with V_{DS} even in the saturation region, which can be accounted for by empirically multiplying the voltage-dependent term in (2.3) by the factor $1 + \lambda V_{DS}$ [2.3]; i.e.

$$I_{DS} = \frac{\beta}{2}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \tag{2.11}$$

where λ is the channel length modulation parameter ($= 0.1$–$0.01\,V^{-1}$). The resultant source–drain resistance r is given by the reciprocal of the drain conductance g_{DS}, as [2.4]

$$g_{DS} = \frac{\Delta I_{DS}}{\Delta V_{DS}} = \frac{\beta}{2}(V_{GS} - V_T)^2\lambda = \frac{I_{DS}\lambda}{1 + \lambda V_{DS}} \simeq I_{DS}\lambda\,;$$
$$r = \frac{1}{g_{DS}} \simeq \frac{1}{\lambda I_{DS}}\,. \tag{2.12}$$

Small-Size Effects. The dc equations described above are for a sufficiently large MOSFET, since they are based on one-dimensional analysis. However, if the channel is short and/or narrow, behavior of actual MOSFETs differs from that predicted by the equations. A typical example is the short-channel effects that arise from a two-dimensional potential distribution and high electric fields in the channel region. An excessively short channel or excessively high drain voltage even causes a punch-through, in which the depletion region of the drain junction is punched through to the depletion region of the source junction.

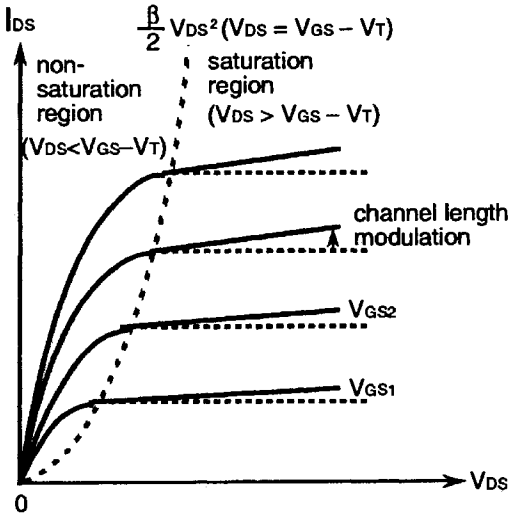


Fig. 2.6. Actual I_{DS}–V_{DS} characteristics [2.1]

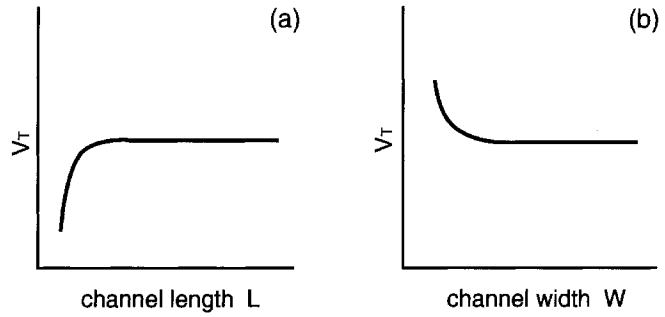


Fig. 2.7. Small size effects in a MOSFET [2.1]. (**a**) The short-channel effect; (**b**) the narrow-channel effect

The lowering of V_T in a shorter channel, shown in Fig. 2.7a [2.1, 2.2], is a result of short-channel effects. On the contrary, the increase in V_T in a narrower channel, shown in Fig. 2.7b, is a small-size effect, that is related to the depletion region spreading laterally in the substrate along the channel width [2.2]. For a given size tolerance, such shifts in V_T make the V_T variation more prominent as MOSFETs are miniaturized. Therefore, MOSFET sizes appropriate to circuit operations must be chosen: the use of an excessively short channel must be avoided for differential circuits such as sense amplifiers and comparators, in order to reduce the V_T imbalance (i.e. offset) between paired MOSFETs. If the setting of the nominal L and W is changed despite the same W/L ratio, the resultant I_{DS} characteristics can be changed, since the short-channel effects are changed.

Advanced Structures and Power-Supply Voltages. Figure 2.8 shows advances in MOSFET structures [2.6]. The miniaturization of the MOSFET calls for a low operating voltage, because of reduction of the hot-carrier breakdown voltage of the MOSFET (BV_{HC}). In DRAMs, the external standard power-supply voltage (V_{DD}) has changed from 12 V to 5 V at the 64 Kb generation. After that, a 5 V V_{DD} has been maintained up to the 4 Mb generation, using devices as small as 0.8 μm. Accordingly, the maintenance of BV_{HC} despite the ever-smaller FET has imposed improvements in the FET drain structure, such as a double drain structure (DDD) in which n^+ layer is covered by a lightly dosed n (n^-) layer, and a lightly dosed drain (LDD) which features an offset structure with an n^- layer. At the 16 Mb generation (0.5–0.6 μm), even the structures suffered from a low BV_{HC}. Then an on-chip voltage down-converter approach was adopted. This lowers a 5 V V_{DD} to 3.3 V internally, and then an internal voltage of 3.3 V is supplied to the internal major circuits. In the subsequent 0.35 μm 64 Mb generation, V_{DD} has been reduced to 3.3 V. In the 64 Mb generation, two approaches coexist. One is that a whole chip operates directly at 3.3 V. The other is that only the array operates at an internal voltage of 2.0–2.5 V, while the remaining circuits operate at 3.3 V.

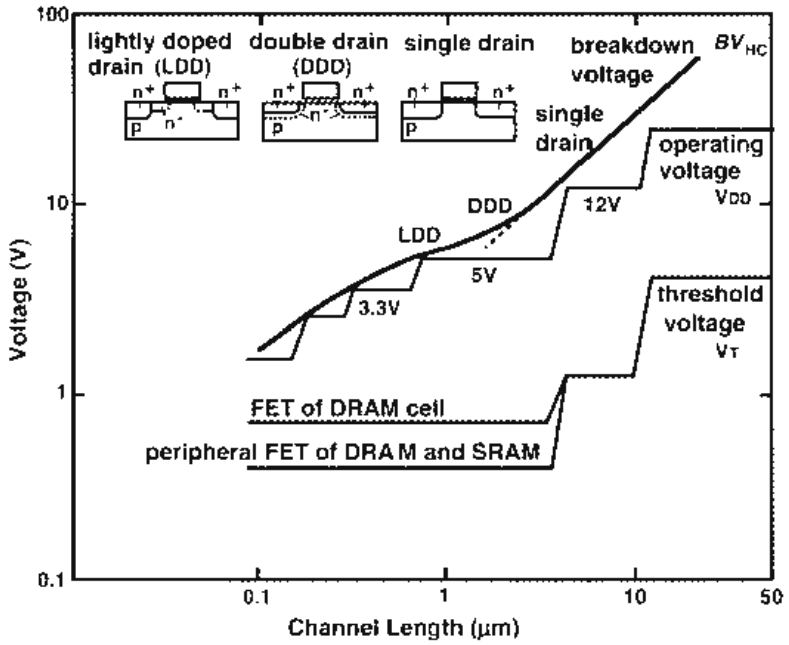


Fig. 2.8. Breakdown voltage, threshold voltage, and operating voltage versus the channel length of a MOSFET [2.6]

2.2.2 Capacitors

The capacitors in a RAM chip are MOS-gate capacitors, p n-junction capacitors, and capacitors of wiring layers (i.e. interconnection layers). Reductions of their capacitances are essential for high speed, because they form loads on circuits. In DRAM, however, higher capacitances, utilizing the MOS capacitor and high-permittivity insulators, have been used intensively for memory-cell and bootstrap capacitors. In general, all of the capacitances per unit area are reduced with device miniaturization.

The MOS Capacitor. The capacitor components of an E-MOSFET are the overlap capacitance C_{OV} $(= C_{OX} l_{OV} W)$ between the gate and the n^+ regions at the source and drain edges, the gate source capacitance C_{GS}, the gate–drain capacitance C_{GD}, and the gate substrate vapacitance C_{GB} [2.3], as shown in Fig. 2.9a. Figure 2.9b shows the gate capacitance C_G versus V_{GS}, excluding the less contributing C_{OV}. Here, C_G is the sum of C_{GS}, C_{GD}, and C_{GB}. At $V_{SG} = 0$, C_G is equal to C_{GB}, which is the gate-oxide capacitance $C_{OX}LW$, because no channel is formed. When V_{GS} is increased, a series connection of the gate capacitance and the depletion-layer capacitance, formed at the silicon surface beneath the gate, is established. Thus C_G is decreased as the depletion capacitance is decreased with V_{GS}. When V_{GS} exceeds V_T, however, C_G increases. In this region C_{GB} becomes zero, because a channel (i.e. an inversion layer or an n-type thin conduction layer) is formed and works as a shield. Instead, C_{GS} appears as the resnlt of a channel formed at the

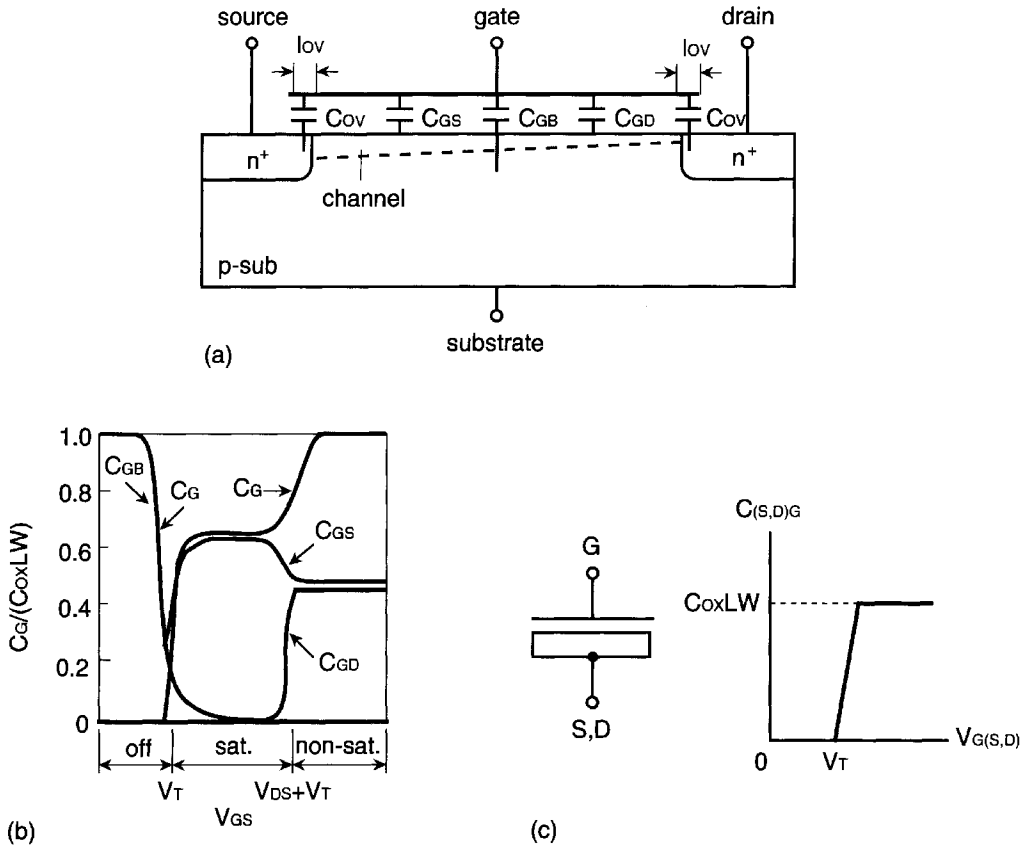


Fig. 2.9. The capacitance characteristics of an E-NMOSFET [2.1]. (**a**) Capacitor components; (**b**) C_G versus V_{GS}; (**c**) the MOS capacitor

source. In the non-saturation region at $V_{GS} > V_{DS} + V_T$, the channel spreads fully between the source and the drain. Hence, the gate-oxide capacitance, which equals C_G, is equally distributed between C_{GS} and C_{GD}.

When the source (S) and drain (D) are connected to each other, a large capacitance is formed between the S–D terminal and the gate when V_{GS} exceeds V_T, as shown in Fig. 2.9c.

Figure 2.10 shows various MOS capacitors [2.1]. Structure (a) is the basic one described above. Structure (b) eliminates the drain. It was widely used for DRAM-cell capacitors in the 16 Kb and 64-Kb generations, in which V_{DD} was applied to the gate. However, the drawback is a charge loss due to a reduction in V_T at the gate capacitor: when V_{DD} is inputted to the source, a full V_{DD} is stored at the small p–n junction capacitance of the source, while a partial voltage of $V_{DD} - V_T$ due to the drop in V_T is stored at the large source–gate capacitance. The resultant small amount of charge at the junction is quickly lost by a given p–n junction current, allowing the source voltage to decay quickly to $V_{DD} - V_T$. After that, a large amount of charge stored at the large capacitance dominated by the gate capacitance starts to be lost, allowing the stored voltage to decay slowly. Eventually, the capacitor cannot

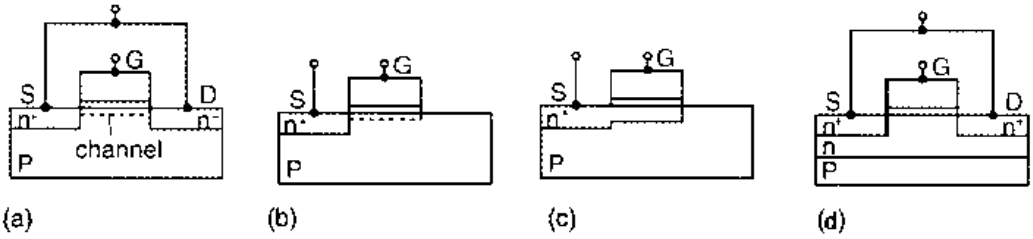


Fig. 2.10. Various capacitors that utilize a MOSFET structure [2.1]. (**a, b**) E-type, variable; (**c, d**) D-type (fixed)

store a full V_{DD} signal charge. To solve the problem, structure (c) uses an n^+ layer beneath the gate, which realizes a pure gate capacitance. This structure was also widely used with an application of 0 V or $V_{DD}/2$ to the gate in the 256 Kb and 1 Mb generations. The application of $V_{DD}/2$ is particularly preferable, since the stress voltage across the gate insulator is halved, or the capacitance is doubled at a fixed stress voltage. Structure (d), in which an NMOSFET is embedded in an n-well, is useful when a large capacitance is needed in peripheral circuits.

Figure 2.11 shows vertical capacitors for DRAM cells [2.1]; stacked and trench capacitors. They have been used widely since the 4 Mb generation. The plate voltage is $V_{DD}/2$, to increase the capacitance. To further increase the capacitance, high-permittivity insulators [2.7] such as multi layered SiO_2, Si_3N_4, and Ta_2O_5 films have been developed. The details are explained in Chap. 3.

The p–n Junction Capacitor. The p–n junction capacitance C_j [2.1] is given by

$$C_j/C_{j0} = (1 + V_j/\phi)^n \,, \tag{2.13}$$

where V_j is the reverse bias voltage at the junction, C_{j0} is the capacitance at $V_j = 0$ V, ϕ is the built-in potential (about 0.6 V), and n varies between -0.5 and -0.3. The capacitance is small, and thus it is not useful as a capacitor element. However, the resultant parasitic capacitance degrades the circuit speed. Table 2.2 compares the various capacitors [2.1, 2.4, 2.8] formed in a chip, including the wiring capacitors, that are discussed later.

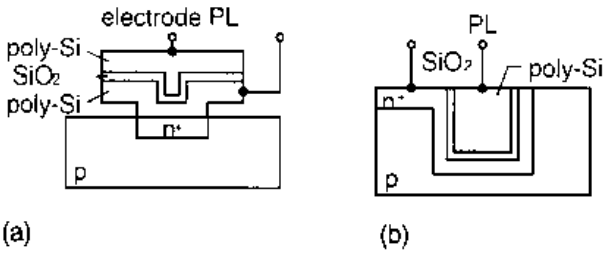


Fig. 2.11. Vertical capacitors [2.1]. (**a**) Stacked; (**b**) trench

Table 2.2. Capacitor elements [2.1, 2.4, 2.8]

		Capacitance	Conditions
MOS capacitor, C_{OX}		2.3 fF/μm^2	$t_{OX} = 15$ nm
Junction caption	n^+	0.18 fF/μm^2	Bias 2 V, area[a]
	p^+	0.51 fF/μm^2	Bias 0 v, area[a]
	n-well to p-sub	0.10 fF/μm^2	Bias 3.3 V
memory cell capacitor, C_s		6.9 fF/μm^2	$t_{OX} = 5$ nm
Single line capitor	Al	0.22 fF/μm	$W_{int}/t_{int} = 2.9, H_{int}/t_{int} = 1.4$
	Poly-Si	0.26 fF/μm	$W_{int}/t_{int} = 5, H_{int}/t_{int} = 0.5$

[a] Area Component

2.2.3 Resistors

Resistors have been widely used for voltage conversions and for eliminating floating nodes [2.1]. Resistances as high as 100 kΩ–1 MΩ are needed to suppress the chip standby current to 10 μA–100 μA. They are usually made of poly-silicon because it has no voltage or temperature dependences, and because of ease of use, despite large variations in resistance during volume production. Here, let us evaluate the length of poly-silicon line with width of 1 μm and a sheet resistance of 100 Ω/square needed to make a 1 MΩ resistor. The sheet resistance ρ_S, which is defined by the resistance (Ω/square, or Ω/□) of a square with thickness t_{int} (Fig. 2.12), is related to the resistivity ρ (Ω · cm) [2.1] as follows:

$$\rho_S = \rho / t_{int} \ . \tag{2.14}$$

The resistance R of a poly-silicon line with a width of W_{int} and a length of l_{int} is

$$R = \rho_S (l_{int}/W_{int}) \ . \tag{2.15}$$

The length can thus be as long as 10 mm, but it is practical, since the necessary area becomes small relative to the ever-increasing chip area.

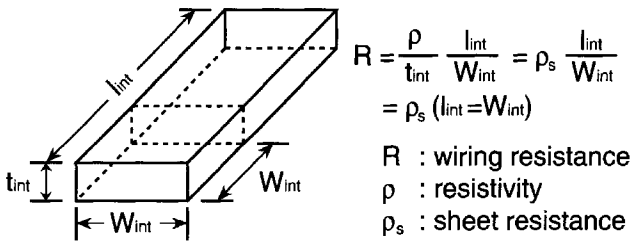


Fig. 2.12. Wiring resistance [2.4]

Diffused n and p layers could be used for resistors, although they have voltage and temperature dependences and need a large area due to small sheet resistances. A combination of an n-well resistor and a poly-silicon resistor has been proposed for a refresh timer [2.9]. Different temperature dependences of resistance between the two – large (+0.6%/°C) for the n-well and almost zero for the poly-silicon – enables detection of the junction temperature and accordingly variation of the refresh interval. Table 2.3 compares various resistors [2.10], including wiring resistors that are discussed later.

Table 2.3. Resistor elements [2.10]

Materials		Resistivity	Sheet resistance W/□ [a]	
		μΩcm	**Width** 2 μm	**Width** 0.5 μm
Poly-Si		1000	15	60
Diffused Layer (n^+)		1500	23	90
Silicide (WSi_2)		130	2	8
Metal Wiring	Al	3	0.05	0.18
	W	10	0.15	0.60

[a] Width= 3× thickness.

2.2.4 Wiring and Wiring Materials

A small parasitic capacitance and resistance, and high reliability against open failures and corrosion, are required for wiring (i.e. interconnections), although these are sometimes contradictory requirements.

Figure 2.13 shows average load-capacitance components for 33 critical circuits on the access path of a 1.3 μm 1 Mb CMOS DRAM [2.1, 2.11]. Obviously, the gate capacitance dominates with 65% of the total, and the wiring capacitance occupies as much as about 33%, although the occupancy is different for row and column circuits, and array-driving circuits. Even for a 3 μm NMOS 64 Kb DRAM [2.12], the wiring capacitance occupies about 30% of the total. Therefore, approximately one-third of the total load capacitance of an average critical circuit is for wiring capacitance. This implies that reduction of wiring capacitance is important for high speed. Wiring resistance should be also reduced as much as possible, since even a small wiring resistance, if coupled to a large gate capacitance at the load, could cause a large *RC* delay. For a 7.8 ns access 4 Mb DRAM [2.13], the delay caused by wiring resistance occupies about 23% (1.8 ns) of the access time. The parasitic capacitance and resistance are closely related to the reliability of the wiring.

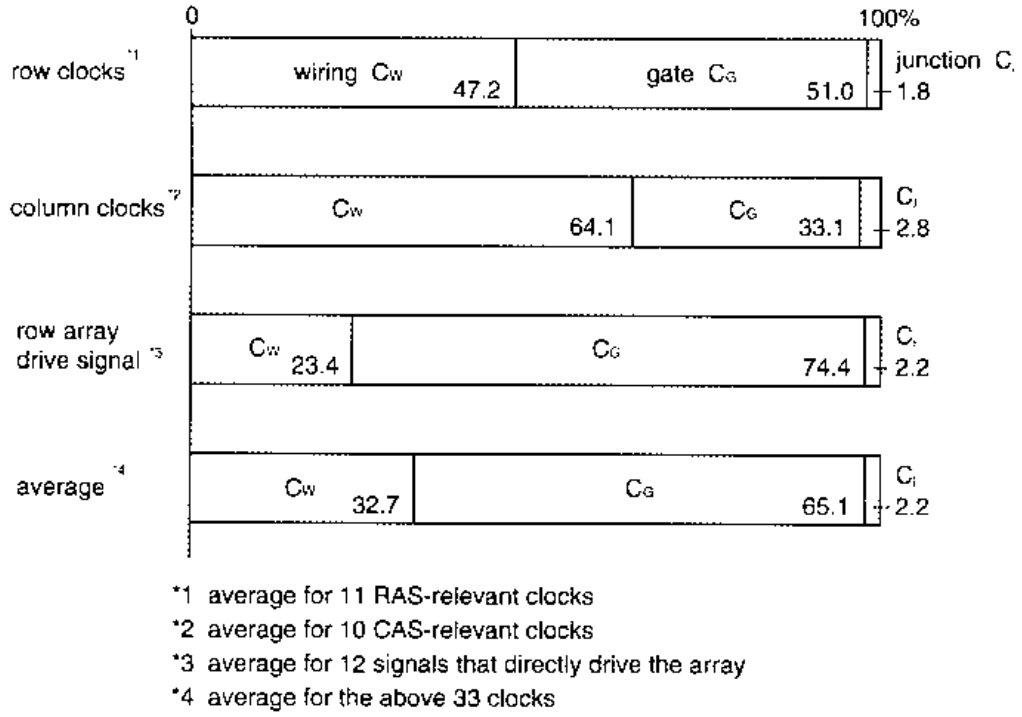


Fig. 2.13. Average load-capacitance components for critical circuits on the access path of a 1.3 μm CMOS 1 Mb chip ($t_{OX} = 25\,\text{nm}$, $L_{eff} = 1.5\,\mu\text{m}$)

Wiring Parasitic Capacitance. The capacitance influences not only speed, but also the spike current, the power consumption and the circuit noise [2.10]. In particular, for the 0.5 μm 16 Mb DRAM generation and beyond, crosstalk is an emerging issue, because the coupling capacitances between adjacent lines are relatively increased. The increase stems from the fact that the thickness cannot be so scaled down as the areal size, to reduce wiring resistance and to ensure the breakdown voltage of the interlayer dielectric film, and immunity from electromigration.

Figure 2.14 shows the capacitance C_{int} of a single strip line [2.8]. For a fixed thickness t_{int}, C_{int} decreases with a decreasing width W_{int}. However, it is saturated to 1 pF/cm when $W_{int} \simeq t_{int} \simeq H_{int}$, due to a fringe effect. Therefore, reducing only W_{int} is not a very effective way of reducing C_{int}. Figure 2.15 shows the capacitance C_1 of one of a number of strip lines [2.8]. The capacitance C_1 becomes lowest at $W_{int}/t_{int} \simeq 1$, when W_{int}/t_{int} is decreased. At $W_{int}/t_{int} < 1$, however, it increases again due to increased coupling capacitances $2C_{12}$. When another set of parallel-running strip lines crosses over the total capacitance, C_1 increases by 25%. Note that the cross-capacitance of two intersecting strip lines, whose widths are 1.4 μm for the lower one and 2.0 μm for the upper one with the same thickness (0.4 μm), and with a 1 μm thick SiO_2 interlayer dielectric film, is as large as six times that calculated by a parallel-plate capacitor approximation [2.14].

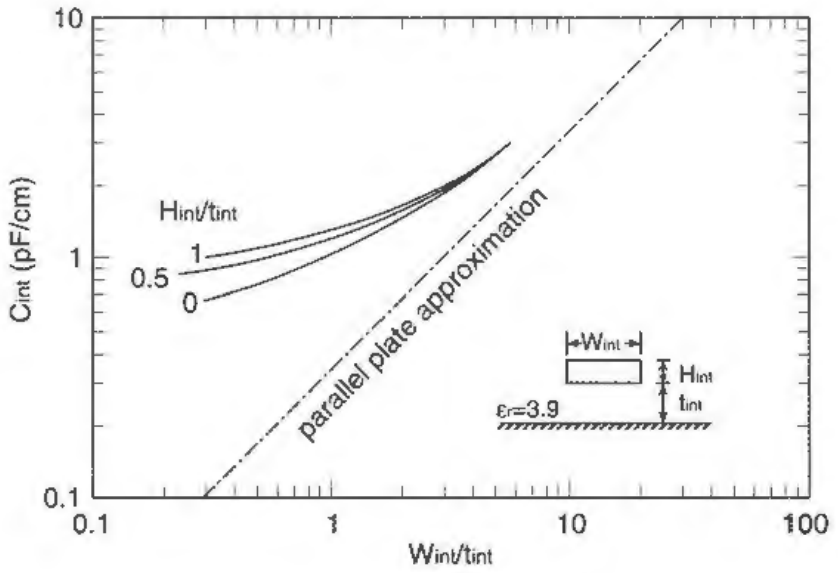


Fig. 2.14. The capacitance of a single strip line [2.8]

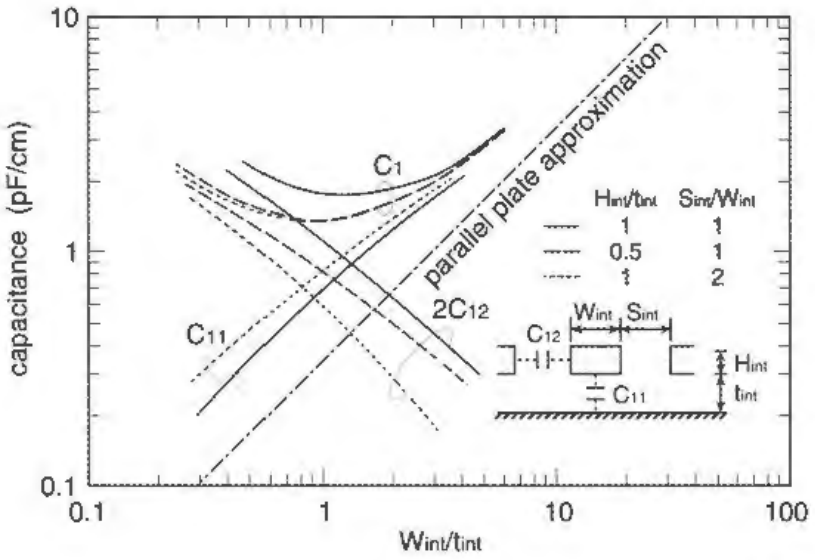


Fig. 2.15. The capacitance of one of a number of strip lines [2.8]

Wiring Parasitic Resistance [2.1, 2.10]. Various wiring materials, are used as shown in Table 2.3. Poly-silicon and diffused p and n layers are suitable for high-resistance elements, but they are not useful for wiring materials because their resistances are rapidly increased as devices are scaled down. Thus, they are only used for short wiring lines. For relatively long and fine lines, such as data lines and word lines in a memory array, a polycide structure or silicide materials are useful. The polycide structure, which is composed of a silicide layer stacked over a poly-silicon layer, concurrently achieves a low-

resistance line due to a low-resistivity silicide, and a self-aligned process and highly reliable MOSFET, realized by the poly-silicon gate process. For long signal lines, low-resistivity metals such as aluminum (Al) and copper (Cu) are absolutely vital. However, even an Al line creates quite a large RC delay with miniaturization. For example, the resistance of a 0.5 μm wide, 2.8 mm long Al line is about 1 kΩ, and the RC delay is thus about 1 ns for a load capacitance of 1 pF.

Reliability of Wiring [2.15]. Despite a low resistivity, Al does not fully meet the ever-stronger needs for fine patterning capability and high reliability, which is closely related to current density. Thus, since the 256-Kb generation, an AlSi alloy has been used to prevent Al from penetrating the silicon substrate at contact holes. In the 1 Mb generation, an AlCuSi alloy was used to avoid electromigration, followed by stacked metals with Al and refractory metals, to avoid stress migration at the 4 Mb generation. Recently, tungsten and copper have started to be used in commercial chips.

Electromigration phenomena cause open failures of metal wiring while a current is flowing. When the current density of Al wiring exceeds a certain magnitude, exemplified by $10^5\,\mathrm{A\,cm^{-2}}$, which corresponds to 10 mA for a 1 μm thick, 10 μm wide Al line, electromigration occurs. The phenomena are characterized by diffusion of Al atoms caused by electron kinetic energy. Consequently, voids are created inside the film, and the resultant increased current density causes a melting of the Al. The lifetime of the metal to failure is given by

$$\tau = AW_{\mathrm{int}}t_{\mathrm{int}}J^{-2}\exp(\phi/kT)\,, \tag{2.16}$$

where A is a constant, W_{int} is the width, t_{int} is the thickness, J is the current density, ϕ is the activation energy (0.5–0.6 eV for pure Al and AlSi), k is the Boltzmann constant, and T is the absolute temperature. Thus, the lifetime is shortened as the current density and temperature are increased. To reduce the phenomena, the cross-section $W_{\mathrm{int}}t_{\mathrm{int}}$ must be kept as large as possible, causing difficulty in reducing t_{int} when reducing W_{int}. The addition of 1% copper extends the lifetime, with a raised activation energy ϕ of 0.6–0.8 eV.

Stress migration phenomena also cause open failures of metal wiring, even without any current flow. This failure occurs even in a high-temperature acceleration test, depending on the width and thickness and thus preventing from fine patterning. It is caused by mechanical stresses being transmitted from the surrounding insulator and silicon substrate to the Al wiring. Stacked structures with Al and stress-migration-immune refractory metals, such as Ti, W, TiW, WSi_2, $MoSi_2$, and TiN, are useful to reduce migration. However, the refractory metals eventually cause difficulty in fine patterning and corrosion after etching. In addition, the resistivities of stacked structures increase as the migration is reduced, due to the increasing content of refractory metal in the pure Al.

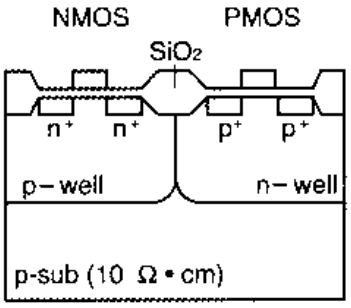


Fig. 2.16. A double-well structure [2.1]

2.2.5 Silicon Substrates and CMOS Latch-Up

Silicon Substrates. In the NMOS DRAM era, in the 1970s and early 1980s, a p-type substrate with a resistivity of about $10\,\Omega \cdot$ cm was used. A substrate bias voltage V_{BB} of -2 to $-3\,V$ was supplied to the substrate to ensure stable operations and isolation between NMOS memory cells. In the 1980s this was replaced by a double-well (or tub) CMOS structure, shown in Fig. 2.16 [2.1]. In principle, the p-well is unnecessary, because a p-type substrate is used. The double well, however, allows NMOS parameters such as V_T to be controlled by adjusting only the p-well dose concentration [2.1]. In order to avoid a forward-biasing of the p–n junction, the p-well and p-substrate are back-biased at a voltage lower than any NMOSFET source voltage, while the n-well is back-biased at another voltage higher than any PMOSFET source voltage. Since the lowest source voltage is V_{SS} (0 V), and the highest PMOSFET source voltage is V_{DD}, the p-well voltage is 0 V or a negative voltage, and the n-well voltage is V_{DD} or a higher voltage. Recently, a triple-well structure has been used to protect a memory-cell array from minority carrier injections or to enable ultra-low-voltage operations, as explained in Chaps. 4 and 7.

CMOS Latch-Up. Latch-up occurs in CMOS structures [2.1], in which a parasitic thyristor consisting of parasitic npn and pnp bipolar transistors is easily formed. As soon as the thyristor is triggered by a spike noise, it is turned on by positive feedback. As a result, a large current flows from V_{DD} to ground, and melts junctions and metal wiring throughout the chip. For example, in the structure forming an n-well on to a p-substrate shown in Fig. 2.17 [2.1], an npnp thyristor composed of an NMOS source, a p-substrate, an n-well and a PMOS source is developed. Here, R_W and R_S are the n-well resistance and substrate resistance, respectively. These resistors are represented by lumped devices in the figure, although they are actually distributed. The resultant circuit features positive feedback, and thus it causes the latch-up when either of two transistors is turned on. When a positive noise is inputted to the Q_1-base, Q_1 is turned on. The resultant collector current creates a voltage drop across R_W, so that the base-emitter of Q_2 is forward-biased. When the bias exceeds a certain voltage, Q_2 is turned on, and the collector current makes the Q_1-base voltage increases further. Thus, the current grows if the

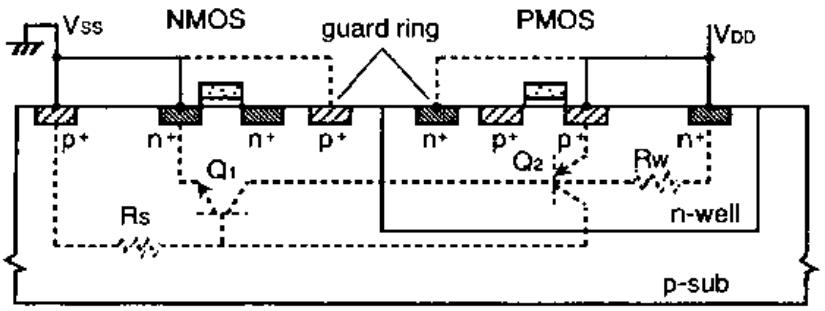


Fig. 2.17. Parasitic bipolar transistors in a CMOS structure [2.1]

product of the current gains of both transistors is larger than unity. Latch-up can be triggered by transient noises, which forward-bias p–n junctions, such as local voltage variations in the substrate and wells, caused by capacitive couplings, ringing waveforms on the signal lines, overshoots exceeding V_{DD} and undershoots exceeding V_{SS} to gate protection diodes at the input and output pins, and noises coupled to power supply lines. In order to suppress latch-up, the current gains of parasitic bipolar transistors must be reduced, which is realized by deepening the wells for the vertical transistor (Q_2), and isolating MOSFETs from the well edges as much as possible for the lateral transistor (Q_1). Reductions of the parasitic resistances R_W and R_S are also effective for suppression. They are achieved by the use of guard rings (Fig. 2.17), a V_{SS}-supplied p^+ guard ring surrounding an NMOSFET, and a V_{DD}-supplied n^+ guard ring surrounding a PMOSFET. The guard rings also lower the current gains, because they capture minority carriers injected before reaching the bases. Latch-up immunity and the usefulness of guard rings have been reported as follows [2.16]. Latch-up occurs more easily with a higher temperature and a wider pulse width of noise. For example, for an npn parasitic transistor, the latch-up starting base-voltage was lowered from 1.2 V at 25 °C and a 10 ns pulse width to 0.4 V at 125 °C and a 50 ns pulse width. An n^+ guard ring improved the latch-up immunity from 0.4 V to 1.2 V. An additional p^+ guard ring brought about a further improvement, to 2 V.

Guard rings are not so effective for the parasitic transistors formed in the deeper regions of substrate and wells, since they lower only the surface resistances. Therefore, increases in the doping concentrations of the substrate and the wells further improve latch-up immunity with lowered R_S and R_W. Excessive doping, however, increases junction capacitances, degrades the mobilities of MOSFETs, and raises V_T. An epitaxial wafer that drastically improves immunity solves the above problems, since it comprises a thin silicon layer with an appropriate doping concentration, grown on to quite a high doping concentration wafer.

2.2.6 Other Devices

Gate-protection resistors, diodes that protect against surges at input pins of chips [2.17], and poly-silicon laser-fuses are also indispensable for higher-density, large-memory-capacity chips.

2.3 NMOS Static Circuits

An inverter is a basic element of MOS digital circuits, such as logic and SRAM-cell circuits. It can be categorized as either a static inverter or a dynamic inverter. A static inverter outputs a voltage determined by the ratio of the effective resistances of two MOSFETs, while a dynamic inverter outputs a voltage determined by the charging and discharging of the output capicitor. NMOS static inverters were widely used in the 1970s due to the high-speed advantage of NMOSFETs over PMOSFETs, which stems from a higher electron mobility, and the stable operation of the static inverter. The high-power dissipation of the static inverter was still acceptable in an era of relatively small-scale integration. Nowadays, static inverters are rarely used, since CMOS circuits dominate. However, they represent basic circuit structures for understanding MOS circuit analysis and design. Figure 2.18 depicts NMOS static inverters, each consisting of an E-MOS driver (Q_D) and a load. The load is one of the following: (a) a saturated E-NMOS, with its gate connected to its drain; (b) a non-saturated E-NMOS; (c) a D-NMOS; or (d) a poly-silicon resistor. Type (a) in particular was widely used in the early digital integrated circuits.

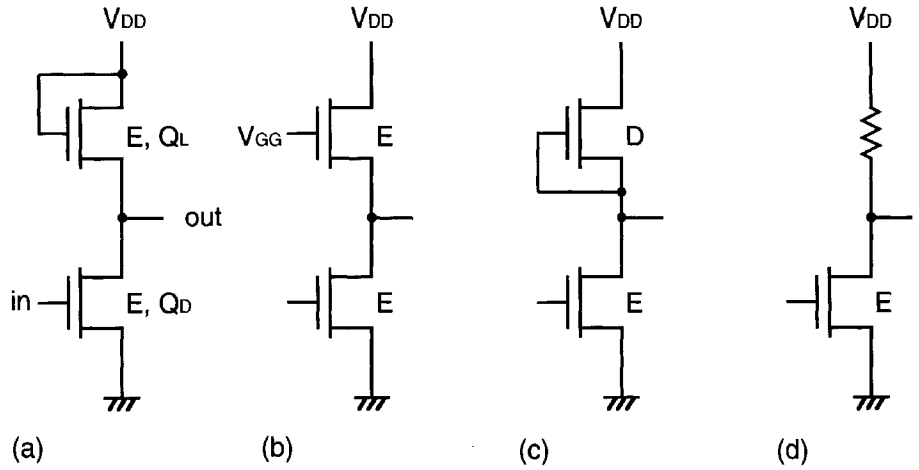


Fig. 2.18. Various NMOS inverters [2.1, 2.3]. (**a**) Saturated E-MOS load; (**b**) non-saturated E-MOS load; (**c**) D-MOS load; (**d**) resistor load

2.3.1 The dc Characteristics of an Inverter

Saturated E-NMOS Load Inverter [2.1, 2.3]. The inverter is analyzed here by using the notation D and L for the parameters of the driver MOS Q_D and the load MOS Q_L, respectively (Fig. 2.19).

When the input voltage $V_{IN} < V_{TD}$, Q_D is off ($I_D \simeq 0$) and $V_{OUT} = V_{DD} - V_{TL}$. When V_{IN} is still low with $V_{IN} > V_{TD}$, Q_D conducts and Q_L is in saturation, since $V_{DS} > V_{GS} - V_{TL}$. Thus, the load current I_L is given by

$$I_L = \frac{\beta_L}{2}\{(V_{DD} - V_{OUT}) - V_{TL}\}^2, \tag{2.17}$$

$$\begin{aligned} V_{TL} &= V_T(|V_{BB}| + V_{OUT}) = V_{T0} + \Delta V_T(|V_{BB}| + V_{OUT}) \\ &= V_{T0} + K\left(\sqrt{|V_{BB}| + V_{OUT} + 2\Psi} - \sqrt{2\Psi}\right). \end{aligned} \tag{2.18}$$

Note that the source–substrate bias of Q_L is increased to the sum of V_{BB} and V_{OUT}, since the source voltage of Q_L becomes V_{OUT}. The driver Q_D is also in saturation, which allows the driver current I_D to be expressed as

$$I_D = \frac{\beta_D(V_{IN} - V_{TD})^2}{2} \qquad (V_{IN} - V_T \le V_{OUT}), \tag{2.19}$$

$$V_{TD} = V_{T0} + K\left(\sqrt{|V_{BB}| + 2\Psi} - \sqrt{2\psi}\right) \tag{2.20}$$

Here, we assume $V_{BB} = 0$ for simplicity and replace V_{TD} with V_T. Hence, the equilibrium conduction of $I_L = I_D$ yields

$$V_{IN} - V_T = \frac{1}{\sqrt{\beta_R}}\left\{(V_{DD} - V_{OUT} - V_T) - K\left(\sqrt{2\psi + V_{OUT}} - \sqrt{2\Psi}\right)\right\}$$
$$(V_{IN} - V_T \le V_{OUT}), \tag{2.21}$$

where β_R (the β-ratio) $= \beta_D/\beta_L$.

When V_{IN} is raised sufficiently, Q_D is in non-saturation. Thus,

$$I_D = \beta_D\left\{(V_{IN} - V_T)V_{OUT} - \tfrac{1}{2}V_{OUT}^2\right\} \qquad (V_{IN} - V_T \ge V_{OUT}). \tag{2.22}$$

By setting $I_D = I_L$,

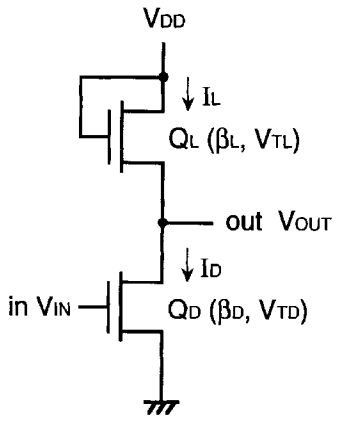


Fig. 2.19. A typical MOS inverter

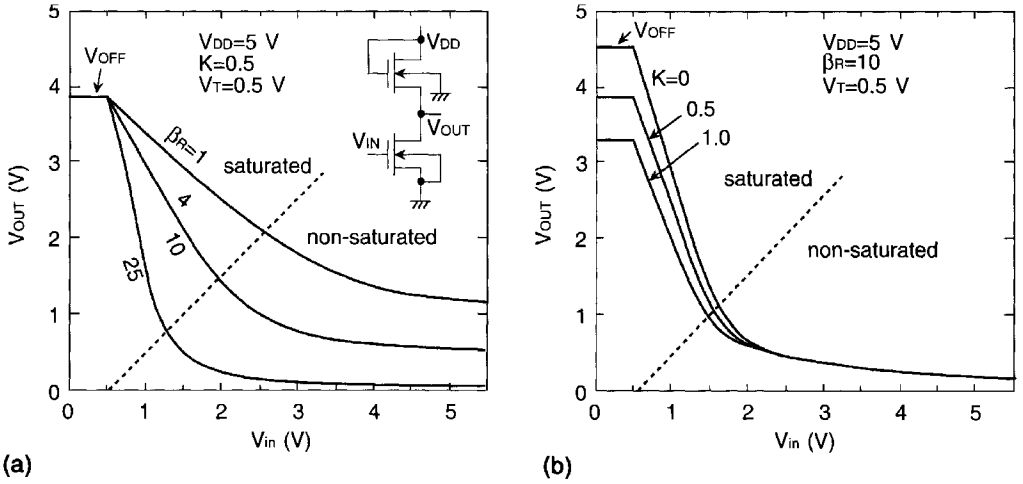


Fig. 2.20. The transfer characteristics of a saturated E-MOS load inverter [2.1]. (**a**) β_R dependence; (**b**) K dependence

$$V_{IN} - V_T = \tfrac{1}{2}V_{OUT} + \frac{\left\{(V_{DD} - V_{OUT} - V_T) - K\left(\sqrt{2\Psi + V_{OUT}} - \sqrt{2\Psi}\right)\right\}^2}{2\beta_R V_{OUT}}.$$

$$(V_{IN} - V_T \geq V_{OUT}) \tag{2.23}$$

Figure 2.20 illustrates relationships between V_{IN} and V_{OUT} as parameters of β_R and K [2.1], which were calculated using (2.21) and (2.23). At $V_{IN} \leq V_T$ V_{OUT} is $V_{OFF}(= V_{DD} - V_{TL})$, which is higher with a small K. When V_{IN} is increased Q_D is turned on, and V_{OUT} is decreased with a further increasing V_{IN}. The decrease is more prominent with a larger β_R. When Q_D is in non-saturation, V_{OUT} gradually decreases with V_{IN}. The magnitude of V_{OUT} (i.e. on-level V_{ON}) becomes smaller with a larger β_R. The inverter is called a ratio circuit because V_{OUT} is determined by β_R; that is, the ratio of the W/Ls of the two MOSFETs.

Other Inverters [2.1, 2.3]. A drawback of the saturated E-NMOS load inverter is a reduced off-level $V_{OFF}(= V_{DD} - V_{TL})$ that raises the on-level V_{ON} of the succeeding inverter. This problem is solved by other inverters in Fig. 2.18. Figure 2.18b shows a non-saturated load with its gate voltage V_{GG} raised to over $V_{DD}+V_{TL}$. A higher V_{GG} enables a larger load current and thus a higher speed or a small W/L of Q_L. However, further drawbacks are the need for an additional power supply V_{GG} and the application of a higher stress voltage to the Q_L-gate insulator. A D-MOS load in Fig. 2.18c solves these problems. The negative V_{TL} that characterizes the D-MOSFET ensures an off-level of a full V_{DD}. Moreover, the gate–source connection of Q_L provides a constant current I_0 given by $\beta_L V_{TL}^2/2$, which is independent of V_{DS} and can be adjusted with V_{TL}. The constant-current characteristics enable a high speed, as discussed later. Assuming that V_{IN} is high enough (i.e. $V_{IN} = V_{DD}$) and V_{OUT} is thus low enough (i.e. $V_{ON} \ll V_{DD} - V_{TD}$), β_R is given by setting the Q_L-current equal to the Q_D-non-saturation current, as follows:

$$I_0 = \frac{\beta_L}{2} V_{TL}^2 = \beta_D \left\{ (V_{DD} - V_{TD}) V_{ON} - \tfrac{1}{2} V_{ON}^2 \right\}$$
$$\simeq \beta_D (V_{DD} - V_{TD}) V_{ON} ; \tag{2.24}$$
$$\therefore \beta_R = \beta_D / \beta_L = \frac{V_{TL}^2}{2 V_{ON} (V_{DD} - T_D)} .$$

In the resistor load (d), $V_{OUT} \simeq V_{DD}$ for a sufficiently low value of V_{IN}. When V_{IN} is high enough and Q_D is thus in non-saturation (i.e. $V_{ON} \ll V_{DD}$),

$$I_0 = (V_{DD} - V_{ON})/R = \beta_D \left\{ (V_{DD} - V_{TD}) V_{ON} - \tfrac{1}{2} V_{ON}^2 \right\} ;$$
$$\therefore I_0 \cong V_{DD}/R \cong \beta_D (V_{DD} - V_{TD}) V_{ON} ;$$
$$\therefore R \cdot \beta_D = \frac{V_{DD}}{V_{ON} (V_{DD} - V_{TD})} .$$

The resistor load has been used for SRAM memory cells. The resistor is made of an extremely highly-resistive poly-silicon to reduce the SRAM-standby current with a high resistance and a small load.

2.3.2 The ac Characteristics of an Inverter

Saturated E-NMOS Load Inverter [2.1, 2.3]. *Discharging Times.* The β-ratio β_R is usually large enough to ensure a large on–off ratio, V_{OFF}/V_{ON}. Therefore, the Q_L current is negligible compared with the Q_D current while discharging the load capacitance C. The discharging time of the inverter shown in Fig. 2.21 is thus calculated as follows [2.1, 2.3].

When a positive step pulse is inputted, the output is discharged from the initial off-voltage V_{OFF} by the Q_D conducting current. At a certain V'_{OUT}, Q_D enters into the non-saturation region from the saturation region. The time $t_{DIS}(1)$ to V'_{OUT} is as follows:

$$t_{DIS}(1) = C(V_{OFF} - V'_{OUT})/I_{DS} ,$$
$$I_{DS} = \beta_D (V_{IN} - V_{TD})^2 / 2 ,$$
$$V'_{OUT} = V_{IN} - V_{TD} ;$$
$$\therefore t_{DIS}(1) = \frac{C\{V_{OFF} - (V_{IN} - V_{TD})\}}{\beta_D / 2 (V_{IN} - V_{TD})^2} = 2\tau_D \frac{V_{OFF} - (V_{IN} - V_{TD})}{V_{IN} - V_{TD}} ,$$
$$\tau_D = \frac{C}{\beta_D (V_{IN} - V_{TD})} . \tag{2.25}$$

The discharging current after $t_{DIS}(1)$ is not constant because of non-saturated Q_D. The time $t_{DIS}(2)$ from V'_{OUT} to V_{DN} is as follows:

$$-C \frac{dV_{OUT}}{dt} = \beta_D \left\{ (V_{IN} - V_{TD}) V_{OUT} - \tfrac{1}{2} V_{OUT} - \tfrac{1}{2} V_{OUT}^2 \right\} ; \tag{2.26}$$
$$\therefore V_{OUT} = (V_{IN} - V_{TD}) \frac{2 e^{-t/\tau_D}}{1 + e^{-t/\tau_D}} .$$

Thus, $t_{DIS}(2)$ and the total discharging time t_{DIS} are given by

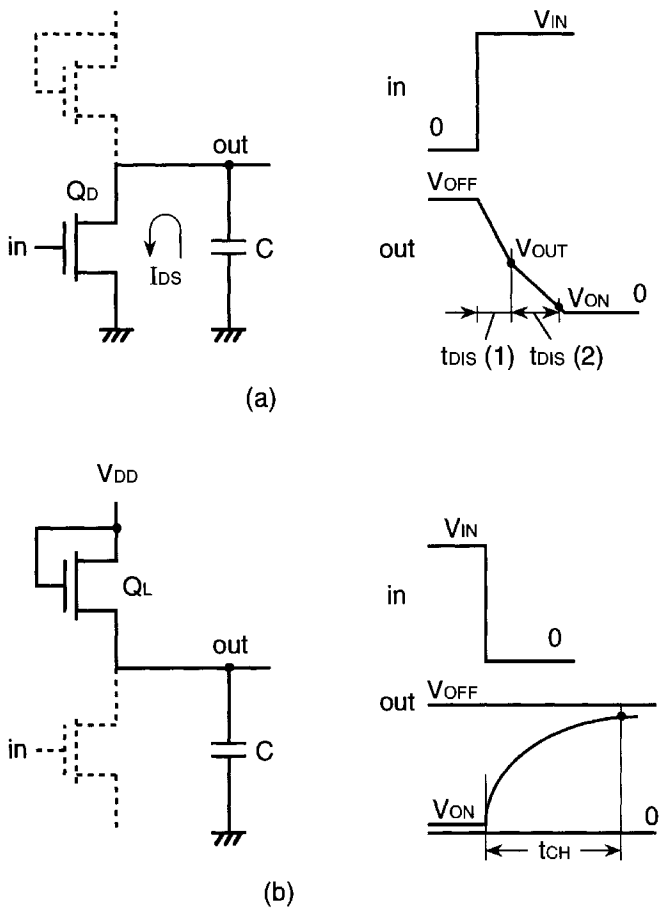


Fig. 2.21. The discharging (**a**) and charging (**b**) of a capacitive load [2.1, 2.3]

$$t_{\mathrm{DIS}}(2) = \tau_{\mathrm{D}} \ln \left\{ \frac{2(V_{\mathrm{IN}} - V_{\mathrm{TD}})}{V_{\mathrm{ON}}} - 1 \right\} , \tag{2.27}$$

$$t_{\mathrm{DIS}} = t_{\mathrm{DIS}}(1) + t_{\mathrm{DIS}}(2) = \tau_{\mathrm{D}} \left[\frac{2\{V_{\mathrm{OFF}} - (V_{\mathrm{IN}} - V_{\mathrm{TD}})\}}{V_{\mathrm{IN}} - V_{\mathrm{TD}}} + \ln \left\{ \frac{2(V_{\mathrm{IN}} - V_{\mathrm{TD}})}{V_{\mathrm{ON}}} - 1 \right\} \right] . \tag{2.28}$$

Since, usually, $t_{\mathrm{DIS}}(2) \gg t_{\mathrm{DIS}}(1)$, $V_{\mathrm{IN}} = V_{\mathrm{OFF}} \gg V_{\mathrm{TD}}$ and $V_{\mathrm{OFF}} \gg V_{\mathrm{ON}}$:

$$t_{\mathrm{DIS}} \simeq \tau_{\mathrm{D}} \ln \frac{2(V_{\mathrm{OFF}} - V_{\mathrm{TD}})}{V_{\mathrm{ON}}} . \tag{2.29}$$

If $V_{\mathrm{ON}} = (V_{\mathrm{OFF}} - V_{\mathrm{TD}})/10$,

$$t_{\mathrm{DIS}} \simeq \tau_{\mathrm{D}} \ln 20 . \tag{2.30}$$

Thus, a larger value of β_{D} and/or a larger value of $(V_{\mathrm{IN}} - V_{\mathrm{TD}})$ are desired to discharge quickly.

Figure 2.22 shows the output voltage versus time [2.1], derived from (2.26). The discharging time from 90% to 10% of the amplitude is about

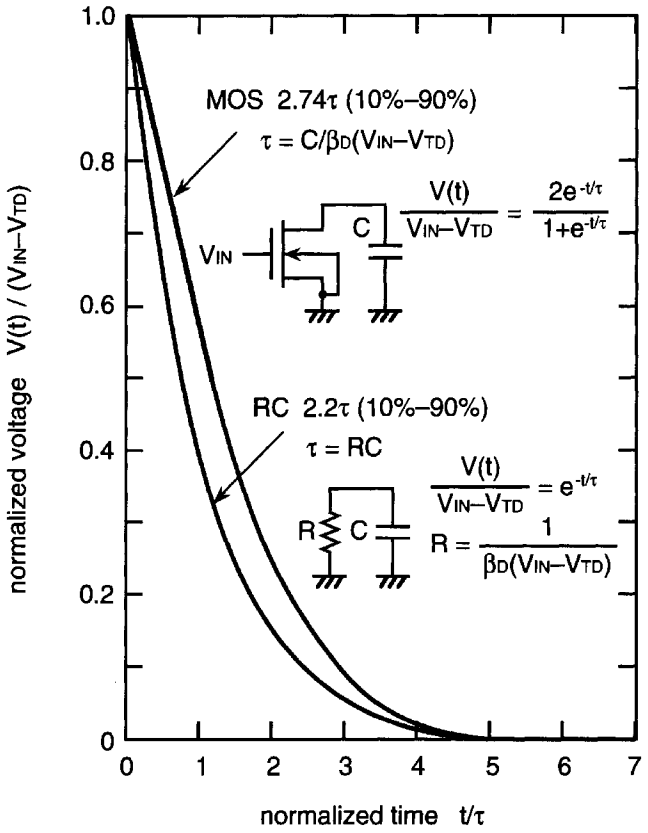


Fig. 2.22. Normalized discharging waveforms [2.1]

$2.7\tau_{\mathrm{D}}$, which is slightly slower than the RC delay ($= 2.2\tau_{\mathrm{D}}$) for $R = [\beta_{\mathrm{D}}(V_{\mathrm{IN}} - V_{\mathrm{TD}})^{-1}]$.

Charging Time. When $V_{\mathrm{IN}} = 0\,\mathrm{V}$, $\mathrm{Q_D}$ is turned off and thus the output voltage increases from V_{ON} to V_{OFF}. The charging time is calculated by assuming that there is no substrate bias effect and thus a fixed V_{TL}, as follows.

Setting the current flowing into C equal to the $\mathrm{Q_L}$ conducting current yields

$$C\frac{\mathrm{d}V_{\mathrm{OUT}}}{\mathrm{d}t} = \frac{\beta_{\mathrm{L}}}{2}(V_{\mathrm{DD}} - V_{\mathrm{OUT}} - V_{\mathrm{TL}})^2 .$$

Using the initial conduction of $V_{\mathrm{OUT}} = V_{\mathrm{ON}}$ at $t = 0$,

$$V_{\mathrm{OUT}} = (V_{\mathrm{OFF}} - V_{\mathrm{ON}})\frac{t/\tau_{\mathrm{L}}}{2 + (t/\tau_{\mathrm{L}})} + V_{\mathrm{ON}} ,$$

$$\tau_{\mathrm{L}} = \frac{C}{\beta_{\mathrm{L}}(V_{\mathrm{DD}} - V_{\mathrm{IN}})} , \qquad V_{\mathrm{OFF}} = V_{\mathrm{DD}} - V_{\mathrm{TL}} . \tag{2.31}$$

Thus, V_{OUT} gradually approaches the off-level $V_{\mathrm{DD}} - V_{\mathrm{TL}}$ over time. The charging time τ_{CH} from V_{ON} to 90% of the pulse amplitude ($= V_{\mathrm{OFF}} - V_{\mathrm{ON}}$) is calculated using (2.31), as

$$t_{\mathrm{CH}} = 18\tau_{\mathrm{L}} . \tag{2.32}$$

Therefore, a smaller τ_L is essential for quick charging-up. Note that for the ratio circuit, usually with $\beta_D/\beta_L \geq 10$, charging is rather slower than discharging, as given by

$$\frac{t_{CH}}{t_{DIS}} = \frac{18}{\ln 20}\frac{\tau_L}{\tau_D} = \frac{18}{\ln 20}\frac{\beta_D}{\beta_L}\frac{V_{OFF} - V_{TD}}{V_{DD} - V_{TL}} \gg 1 . \tag{2.33}$$

The slow charging comes from the ever-decreasing V_{GS} of the load MOS: the load MOS current decreases as the output voltage increases, so that the slow speed is more enhanced with time. This is clearly seen in the normalized charging-up waveform shown in Fig. 2.23 [2.1]. As a result, the speed is quite slow compared with the RC delay for $R = [\beta_L(V_{DD} - V_{TL})]^{-1}$ that is shown for inverter (c) in Fig. 2.18. In any case, the inverter speed is eventually determined by the charging time.

The above analysis is based on the assumption of no substrate bias effect. In practice, however, the ever-increasing output voltage causes an ever-increasing threshold voltage of the load MOS, with its increased source voltage, further degrading the speed.

Other Inverters. The charging speed described above is improved by the other inverters shown in Fig. 2.18. Analyses for inverters (b) and (c) are

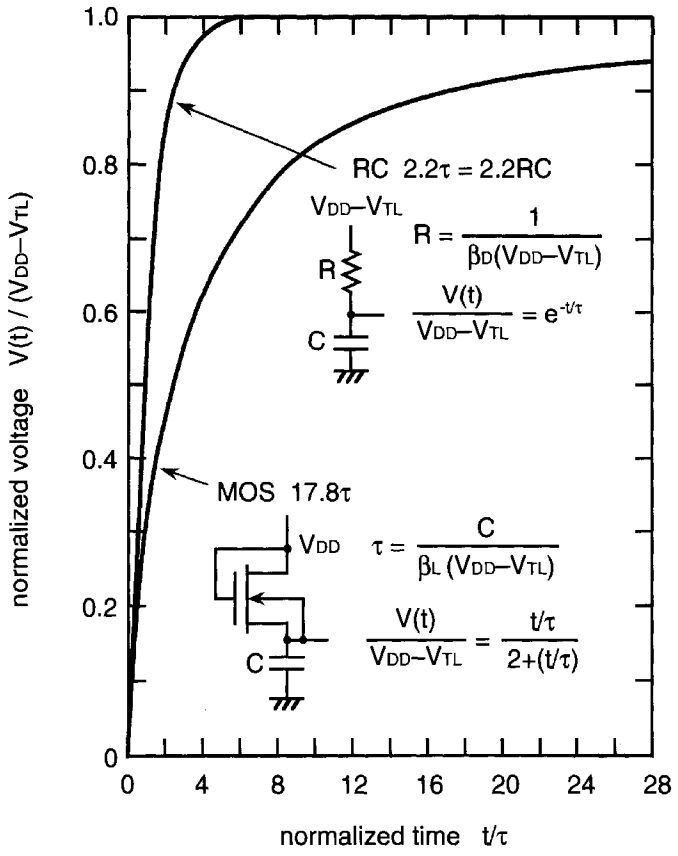


Fig. 2.23. Normalized charging waveforms [2.1]

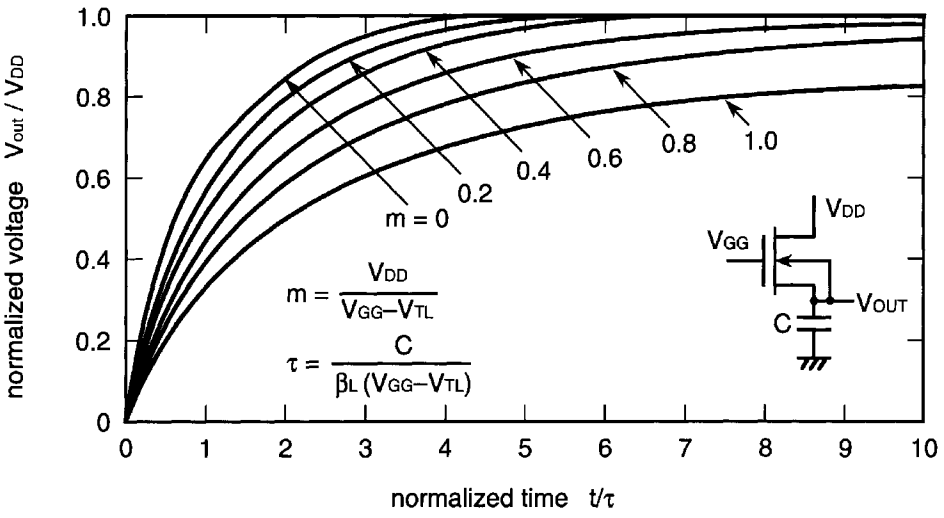


Fig. 2.24. Normalized charging waveforms with a non-saturated MOS load inverter [2.1]

presented here, because the waveform for inverter (d) has been shown in Fig. 2.23. For inverter (b) the following equation is established, as discussed previously:

$$C\frac{\mathrm{d}V_{\mathrm{OUT}}}{\mathrm{d}t} = \frac{\beta_{\mathrm{L}}V_{\mathrm{DD}}^2}{2m}\left(1-\frac{V_{\mathrm{OUT}}}{V_{\mathrm{DD}}}\right)\left(1-m\frac{V_{\mathrm{OUT}}}{V_{\mathrm{DD}}}\right).$$

Assuming a constant V_{TL} (i.e. no substrate bias effect), V_{OUT} is thus given by

$$V_{\mathrm{OUT}} = V_{\mathrm{DD}}\frac{(2-m)\{1-\mathrm{e}^{-(t/\tau_{\mathrm{L}})(1-m)}\}}{2-m\{1+\mathrm{e}^{-(t/\tau_{\mathrm{L}})(1-m)}\}}, \tag{2.34}$$

$$m = \frac{V_{\mathrm{DD}}}{V_{\mathrm{GG}}-V_{\mathrm{TL}}}, \qquad \tau_{\mathrm{L}} = \frac{C}{\beta_{\mathrm{L}}(V_{\mathrm{GG}}-V_{\mathrm{TL}})}.$$

Figure 2.24 shows normalized charging-waveforms as a function of the biasing parameter m [2.1], calculated using (2.34). It is obvious that the speed for $m = 1$ is slow, which is an almost saturated E-NMOS load. The speed is improved with increasing m (i.e. increasing V_{GG}). It is fastest at a value of $m = 0$, approaching that for the resister load.

For inverter (c) the load is charged up more quickly than for inverter (a), because of a constant I_0 that is independent of the load voltage. The charging time is given by $C(V_{\mathrm{DD}} - V_{\mathrm{ON}})/I_0$.

2.3.3 The Improved NMOS Static Inverter

The ratio operation of the static inverter increases the size, power dissipation, and input capacitance, which forms a significant load capacitance for the preceding inverter. The channel width of the load MOS tends to be large to

shorten the charging speed that dominates the total inverter speed. A large β-ratio is needed for a successful logic operation with a sufficiently reduced on-level. These requirements for the speed and voltage level inevitably increase the size and input capacitance of the inverter. The resultant ratio current also increases the power dissipation. For the saturated E-NMOS load, even the V_T reduction is involved. The push–pull inverter and the bootstrap inverter partly solve these problems.

Figure 2.25 shows an inverter that consists of a static inverter and an NMOS push–pull inverter [2.3]. It features the fast driving capability of a heavy load capacitance with a small power dissipation. The push–pull drives the load without consuming any dc current, because either Q_1 and Q_2 is conductive, allowing the W/L of driver Q_2 to be small. The size, power, and input capacitance of the static inverter can be small because its load capacitance is quite small, being only due to the push–pull inverter. However, there is still a V_T-reduction at the Q_1 gate.

Figure 2.26a shows a bootstrap inverter that uses a bootstrap capacitor C_B [2.3]. The biasing MOS Q_1 has charged node N_1 up to $V_{DD} - V_{T1}$ (where V_{T1} is the V_T of Q_1), while the input voltage is high, and the output voltage is thus low as a result of a ratio operation of Q_L and Q_D. As soon as the input changes to a low level and Q_D is cut off, Q_L starts to charge up the load. The voltage change at the output is fed back to the Q_L gate by a large MOS capacitor C_B, so that Q_L drives the load more quickly with the resultant increased gate voltage: the voltage change ΔV at the output allows node N_1 to be raised by $\Delta V' = \Delta V \cdot C_B/(C_B + C')$, where C' is an N_1-parasitic capacitance. Since Q_1 is cut off and isolated from N_1, $C_B \gg C'$ and thus $\Delta V' \simeq \Delta V$. Upon reaching $\Delta V = V_{DD}$, the N_1 voltage is boosted to $2V_{DD} - V_{T1}$ while the gate–source voltage of Q_L is maintained at $V_{DD} - V_{T1}$. During this boosting process, Q_L operates in the non-saturated region. In this circuit,

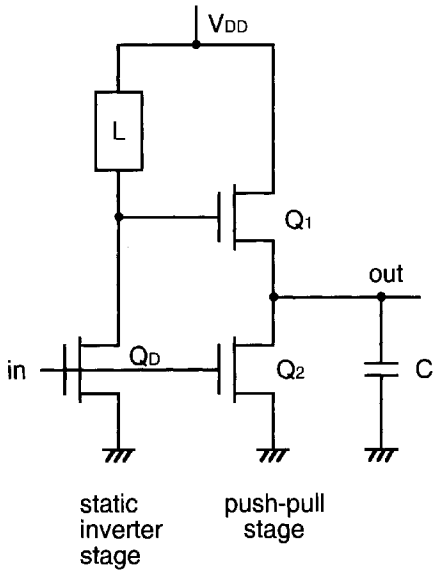


Fig. 2.25. The push-pull inverter [2.3]

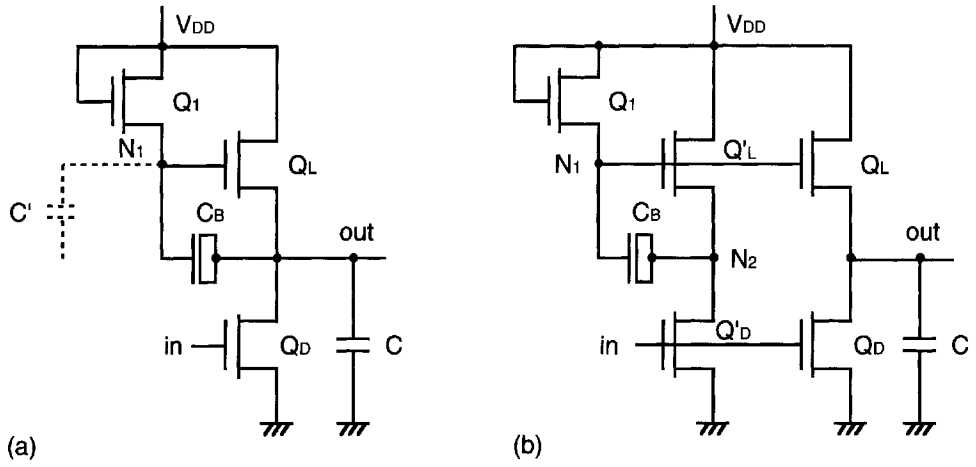


Fig. 2.26. The bootstrap inverter [2.1, 2.3]. (a) Direct drive; (b) load-isolation drive

however, a large value of C/β_L makes the output rise slowly and the resultant delayed boosting at N_1 prevents Q_L from driving quickly. The problem is solved by the circuit shown in Fig. 2.26b, in which a bootstrap stage is isolated from the output. The resultant extremely small N_2 capacitance and the small N_1 capacitance allow the Q_L gate to be boosted quickly to $2V_{DD} - V_{T1}$, independently of the load. After that, Q_L starts to drive the load. It improves the speed because the gate–source voltage of Q_L is increased to $2V_{DD} - V_{T1}$ [2.1].

2.4 NMOS Dynamic Circuits

In the dynamic circuit, charges held at the output capacitance in advance, by a precharge operation, are discharged or preserved, depending on the input signal information. The preserved charges must be maintained at the floating output capacitance until the next signal arrives. Thus, an additional precharge circuit and its activation clock are needed, and there must be no leakage path at the output.

2.4.1 The Dynamic Inverter

Figure 2.27 shows a basic dynamic inverter [2.1]. After charging the output capacitance C to a high level using the precharge transistor Q_1, Q_1 is turned off. Then, an input signal is applied. If it is low, Q_2 is kept switched off, so that the high level is held at the floating output node. If it is high, Q_2 is turned on, so that the load is discharged to a low level. Thus, this circuit works as an inverter.

Unlike static inverters, in this inverter no dc current flows. Hence the power dissipation is extremely low. Only a charging current from the power

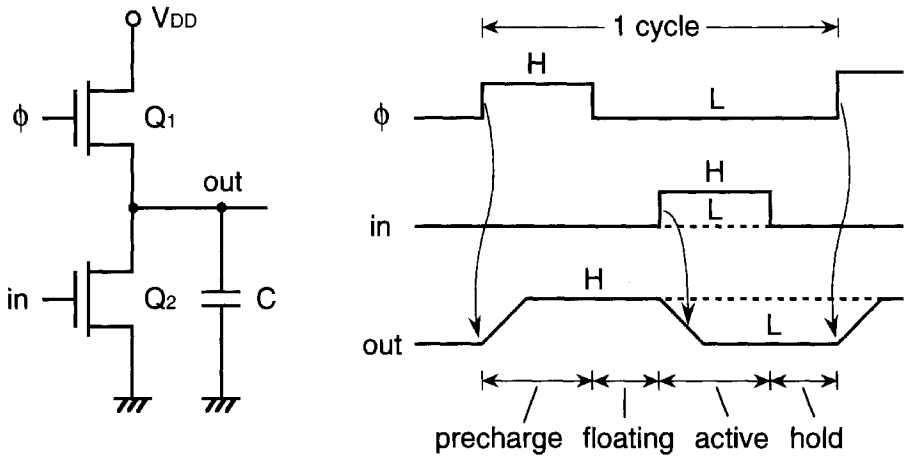


Fig. 2.27. The dynamic inverter [2.1]

supply is needed. The transistor sizes of Q_1 and Q_2 can be determined independently, being released from the restriction of the ratio operation of static inverters. Even the same and minimum feature size transistor for the both can be used, enabling a higher density. However, careful attention must be paid to achieving stable operation. Errors can occur if the node is discharged by spurious signals, since the operation depends on the charges held at the floating node. The spurious signals would consist of p–n junction leakage currents at Q_1 and Q_2, small positive noises at the Q_2 gate, negative bounces at the Q_2 source, and the subthreshold current of Q_2, discussed in Chap. 8.

2.4.2 The Bootstrap Driver

In the static inverter shown in Fig. 2.26, the ratio current is eliminated if node N_1 can be kept low until a low input pulse is applied. This is achieved by using a clock ϕ_p, as shown in Fig. 2.28 [2.1]. The input and node N_1 are kept low until a precharge period finishes. When the input increases Q_1 becomes highly conductive with its gate voltage boosted by its drain–gate MOS capacitance, so that node N_1 quickly goes to V_{DD}. Thus, low ratio voltages are developed at node N_2 and at the output. However, both N_3 and N_4 are discharged after a short delay τ, allowing N_2 to begin to be charged up by Q'$_L$ and Q_1 to be isolated from N_1. Thus, a voltage change at N_2 makes N_1 fully boosted, so that Q_L quickly drives the output. The boosted voltage at N_1 is suppressed to about $1.5V_{DD}$ by adjusting the magnitude of C_B, since excessive boosting gives a high stress voltage to Q_2. The numbers in the figure are ratios of the channel widths that are necessary to obtain a rise 5 ns, delay 5 ns output pulse with 3 μm NMOSFETs ($L_g = 3\,\mu m, L = 2.2\,\mu m, t_{OX} = 50\,nm, V_T = 0.5\,V$) for a 64 Kb DRAM. These values are obtained as a result of optimizing the power and speed. Note that C_B needs 50 times as large capacitance as the gate of Q_3. Once the channel width of Q_L necessary to obtain a desired rise

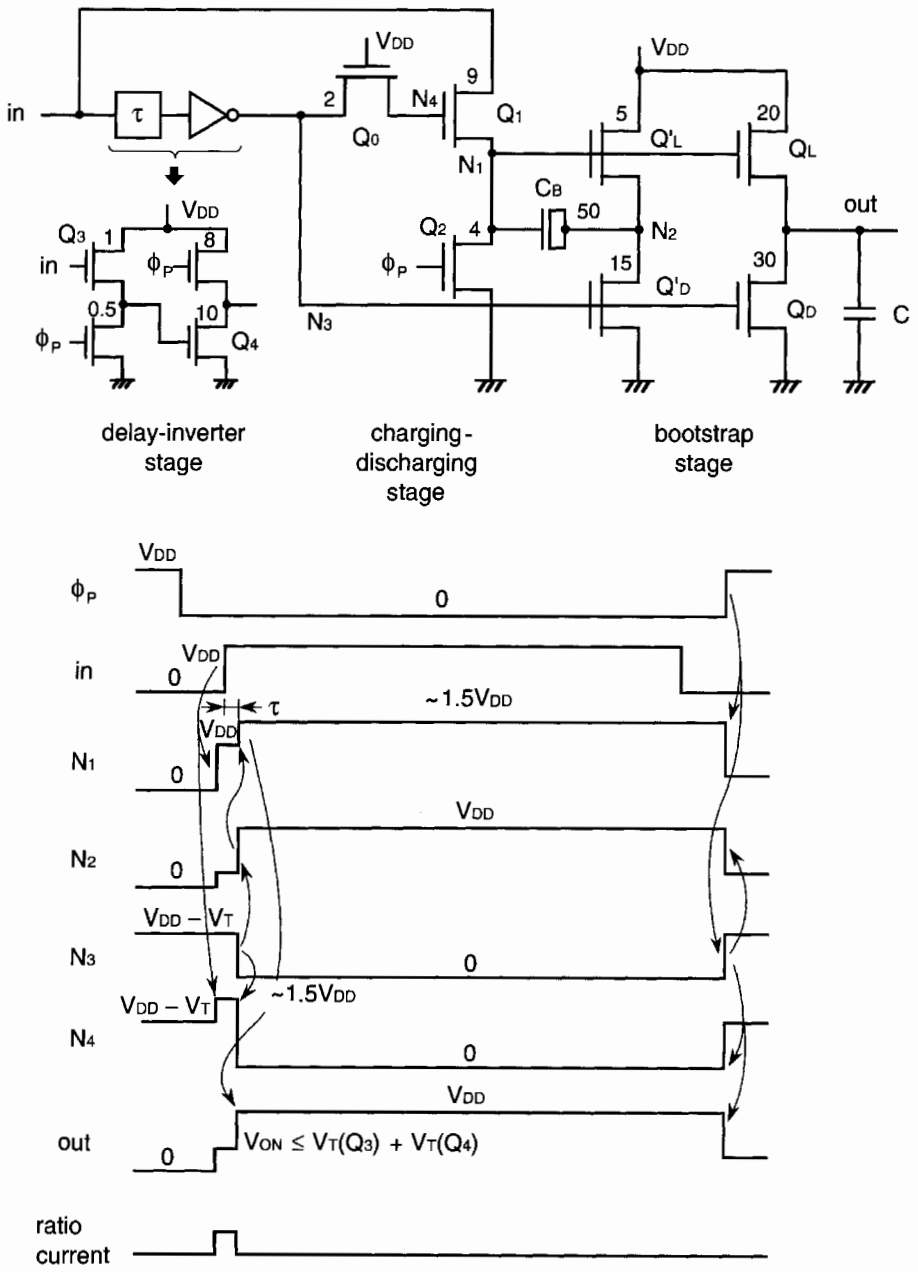


Fig. 2.28. The dynamic bootstrap driver [2.1]

time for a given load capacitance is determined, the other sizes are obtained automatically.

In this driver, the instantaneous ratio voltage must be less than the sum of the V_T's of Q_3 and Q_4, to prevent a node of the succeeding driver, which has the same circuit configuration, from discharging. The drawbacks of the driver are the existence of floating nodes, the need for node boosting and thus for many MOSFETs, a large input capacitance that is the sum of the

gate capacitances of Q_1, Q_3, Q'_L, and Q_L, and a relatively large ratio current despite a short period. However, the driver was widely used to drive heavy load capacitances in the 16 Kb and 256 Kb DRAM generations.

2.5 CMOS Circuits

The key issues of LSIs are low power, ease of design, and a wide voltage margin. Low power, especially for inactive circuits, is crucial for a larger-scale integration. This is because almost all circuits in a memory or MPU/ASIC chip are inactive, while only a limited number of circuits are active, thus calling for an extremely low power for inactive circuits. Otherwise, no LSI chip could be designed successfully, due to the ever-increasing power that accompanies integration. Thus dynamic circuits, particularly bootstrap circuits, were used in the NMOS DRAM era. Dynamic circuits, however, still involve high power, a complicated design and thus a narrow voltage margin, in addition to the reliability issue caused by node boosting, as discussed previously. CMOS circuits solve these problems, despite a process complexity caused by using two kinds of channel. This is the main reason why CMOS circuits replaced NMOS circuits in the early 1980s.

Figure 2.29 shows a basic complementary CMOS inverter and its schematic operations [2.1]. An NMOSFET discharges the load capacitance quickly and completely, while it charges up slowly, entailing a reduction in V_T, as discussed previously. On the other hand, a PMOS charges up quickly without a reduction in V_T, while it discharges slowly, which does entail a reduction in V_T. The CMOS circuit combines the advantages of the two kinds of MOSFET, enabling the fast charging up by a PMOSFET and fast discharging by an NMOSFET, with a full swing of V_{DD}. It features an extremely low power, since one of the two devices is conductive, depending on the input voltage, without any floating and boosted node. The simple circuit configuration is

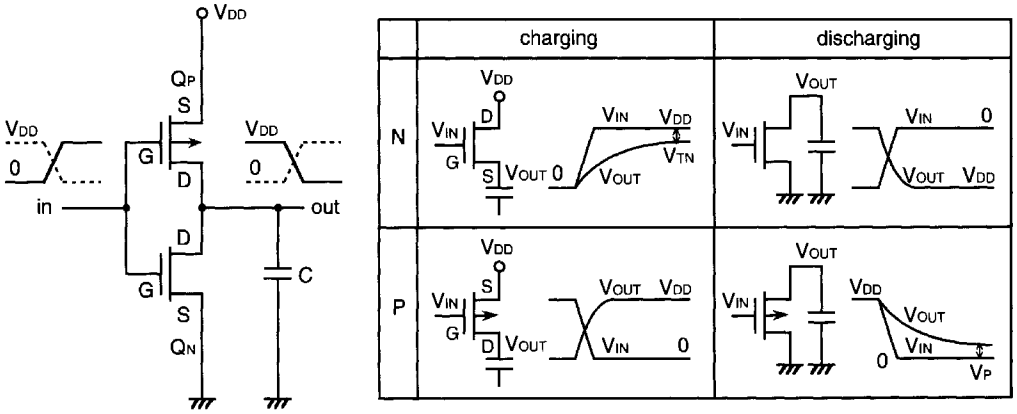


Fig. 2.29. The CMOS inverter [2.1]

another advantage. In addition, there are no substrate bias effects, because the source voltages are fixed. The dc and ac characteristics of the CMOS inverter are as follows.

2.5.1 The dc Characteristics

The operations are as follows, when increasing the input voltage from 0 V to V_{DD}, [2.1, 2.3]. Here, the voltage relationship shown in Fig. 2.30 and the transfer characteristics given in Fig. 2.31 are used, the notations N and P denote NMOSFET and PMOSFET, β_N and β_P are given by

$$\beta_N = \frac{W_D}{L_D}\mu_N C_{OX}\,, \qquad \beta_P = \frac{W_L}{L_L}\mu_P C_{OX}\,,$$

and V_{TP} is negative.

Region I: since $V_{IN} \leq V_{TN}$, Q_N is off and Q_P is on, and thus V_{OUT} is V_{DD}.

Region II: since Q_N is in saturation and Q_P is in non-saturation,

$$V_{OUT} \geq V_{IN} - V_{TN}\,, \qquad I_{DN} = \frac{\beta_N}{2}(V_{IN} - V_{TN})^2\,,$$
$$V_{OUT} \geq V_{IN} - V_{TP}\,,$$
$$I_{DP} = \beta_P\left\{(V_{IN} - V_{DD} - V_{TP})(V_{OUT} - V_{DD}) - \tfrac{1}{2}(V_{OUT} - V_{DD})^2\right\}.$$

The equilibrium condition of $I_{DN} = I_{DP}$ gives the relationship between V_{IN} and V_{OUT}.

Region III: both MOSFETs are in saturation. Hence,

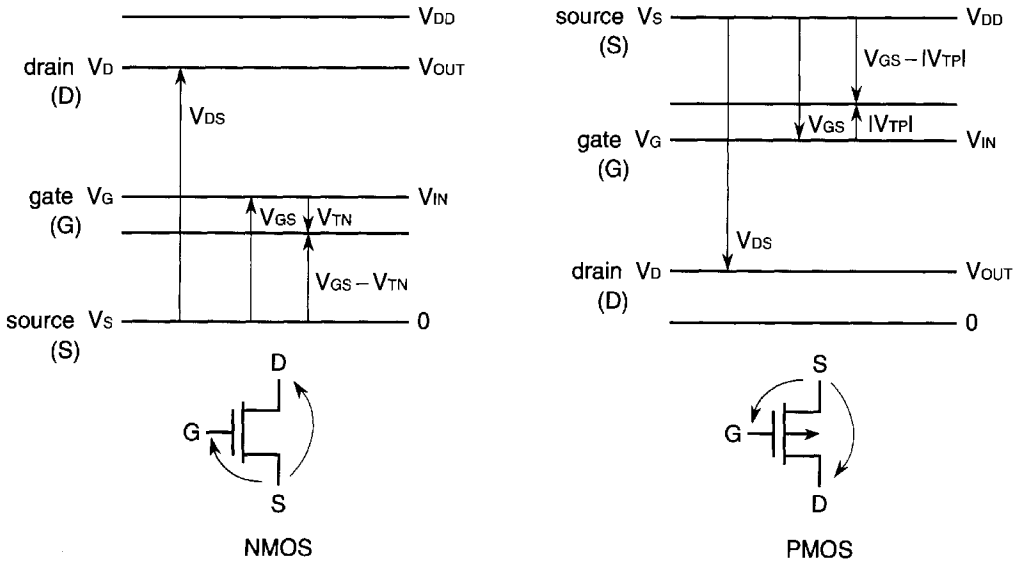


Fig. 2.30. Voltage relationships in the CMOS inverter [2.1]

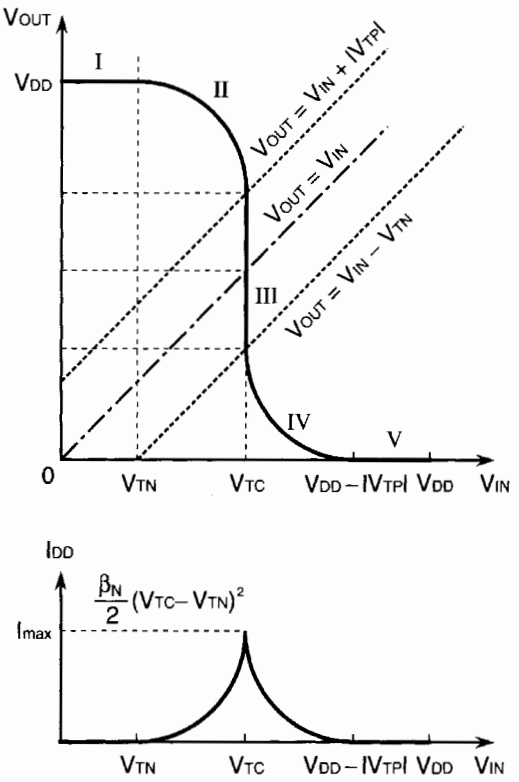


Fig. 2.31. The transfer characteristics of the CMOS inverter [2.1]

$$V_{\mathrm{OUT}} \geq V_{\mathrm{IN}} - V_{\mathrm{TN}}\,, \qquad I_{\mathrm{DN}} = \frac{\beta_{\mathrm{N}}}{2}(V_{\mathrm{IN}} - V_{\mathrm{TN}})^2\,,$$

$$V_{\mathrm{OUT}} \geq V_{\mathrm{IN}} - V_{\mathrm{TN}}\,, \qquad I_{\mathrm{DP}} = \frac{\beta_{\mathrm{P}}}{2}(V_{\mathrm{IN}} - V_{\mathrm{DD}} - V_{\mathrm{TP}})^2\,.$$

Setting $I_{\mathrm{DN}} = I_{\mathrm{DP}}$ yields the threshold voltage of the circuit (V_{TC}), at which the output is in a critical condition at a high or low level, which is expressed as

$$V_{\mathrm{IN}} = V_{\mathrm{TC}} = \frac{V_{\mathrm{DD}} + V_{\mathrm{TP}} + V_{\mathrm{TN}}\sqrt{\beta_{\mathrm{R}}}}{1 + \sqrt{\beta_{\mathrm{R}}}}\,, \qquad \beta_{\mathrm{R}} = \beta_{\mathrm{N}}/\beta_{\mathrm{P}}\,.$$

$V_{\mathrm{TC}} = V_{\mathrm{DD}}/2$ for $\beta_{\mathrm{R}} = 1$ and $V_{\mathrm{TN}} = -V_{\mathrm{TP}}$, and the dc current here is a maximum (I_{max}), and is given by $\beta_{\mathrm{N}}(V_{\mathrm{TC}} - V_{\mathrm{TN}})^2/2$.

Region IV: $\mathrm{Q_N}$ is in non-saturation while $\mathrm{Q_P}$ is in saturation. Thus,

$$V_{\mathrm{OUT}} < V_{\mathrm{IN}} - V_{\mathrm{TN}}\,, \qquad I_{\mathrm{DN}} = \beta_{\mathrm{N}}\left\{(V_{\mathrm{IN}} - V_{\mathrm{TN}})V_{\mathrm{OUT}} - \tfrac{1}{2}V_{\mathrm{OUT}}^2\right\}\,,$$

$$V_{\mathrm{OUT}} \leq V_{\mathrm{IN}} - V_{\mathrm{TP}}\,, \qquad I_{\mathrm{DP}} = \frac{\beta_{\mathrm{P}}}{2}(V_{\mathrm{IN}} - V_{\mathrm{DD}} - V_{\mathrm{TP}})^2\,.$$

By setting $I_{\mathrm{DN}} = I_{\mathrm{DP}}$ the relationship between V_{IN} and V_{OUT} is obtained.

Region V: $V_{\mathrm{IN}} \geq V_{\mathrm{DD}} + V_{\mathrm{TP}}$. Thus, $\mathrm{Q_N}$ is on while $\mathrm{Q_P}$ is off, resulting in $V_{\mathrm{OUT}} = 0\,\mathrm{V}$.

2.5.2 The ac Characteristics

When a positive step pulse is inputted, the load capacitor starts to be discharged. The discharging time t_{DIS} to $V_{DD}/10$ is given from (2.30) as

$$t_{DIS} \simeq \tau_N \ln 20\,, \qquad \tau_N = \frac{C}{\beta_N(V_{DD} - V_{TN})}\,,$$

assuming that $V_{TN}/V_{DD} \ll 1$. The charging time t_{CH} for $|V_{TP}|/V_{DD} \ll 1$ is also given by

$$t_{CH} \simeq \tau_P \ln 20\,, \qquad \tau_N = \frac{C}{\beta_N(V_{DD} - V_{TN})}\,.$$

Furthermore, the average delay time per stage, defined at the 50% point of the input and output waveforms, for a chain of serial inverters is about $2C/(\beta_N V_{DD})$ for $\beta_N = \beta_P$ [2.3].

The power consumption of the inverter consists of two components [2.3]. One is the charging and discharging power (P_1) of the load, which is given by CV_{DD}^2/t_C (where t_C is the cycle time). The other is the dc power (P_2) due to a current from V_{DD} to Q_P and Q_N. The current flows twice, at t_1 and t_2, every cycle, corresponding to an input voltage that ranges from V_{TN} to $V_{DD} - V_{TP}$ in Figs. 2.31 and 2.32. Assuming a triangular current waveform, P_2 is given by

$$P_2 = \tfrac{1}{2} I_{max}(V_{DD} - V_{TN} - |V_{TP}|)\frac{t_1 + t_2}{t_c}\,, \qquad I_{max} = \frac{\beta_N}{2}(V_{TC} - V_{TN})^2$$

if $t_1, t_2 \ll t_C$ and $P_1 \gg P_2$.

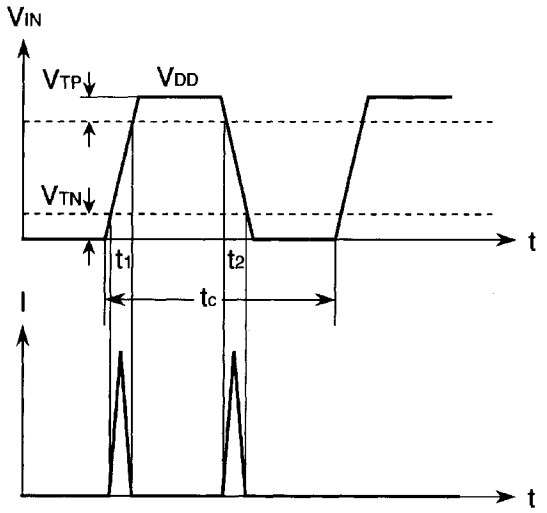


Fig. 2.32. Current flows in the CMOS inverter [2.3]

2.6 Basic Memory Circuits

2.6.1 The Inverter and the Basic Logic Gate

Figure 2.33 shows cross-sections of basic NMOS and CMOS inverters [2.1]. Figure 2.34 illustrates logic diagrams, truth tables, circuits, and layouts for a typical logic gate used in a memory chip [2.1]. A small circle is added at the output in the logic diagram if the logic input is inverted at the output. The circle is omitted for a non-inverted logic gate, as exemplified by two serially-connected inverters.

2.6.2 The Current Mirror

Figure 2.35 shows a current mirror [2.1]. If both Q_1 and Q_2 are in saturation,

$$I_1 = (\beta_1/2)(V_G - V_T)^2\,, \quad I_2 = (\beta_2/2)(V_G - V_T)^2\,;$$
$$\therefore I_2/I_1 = \beta_2/\beta_1 = W_2/W_1\,. \tag{2.35}$$

Hence, if I_1 is supplied from a constant-current generator, a constant current I_2 multiplied by W_2/W_1 is available.

2.6.3 The Differential Amplifier

The differential amplifier is indispensable for detecting a small signal in a noisy environment, as discussed in Chap. 3. It can be categorized as a normal differential amplifier or a cross-coupled differential amplifier. The former simply amplifies a small signal, with a separated input and output. A good example is the current-mirror amplifier, which has been used as a main amplifier on common I/O lines of DRAMs and SRAMs. The latter amplifies a small signal to V_{DD} and latches the amplified signal at the input. It has been widely used for DRAMs because of its simple circuit configuration, low power, and suitability for rewrite operations.

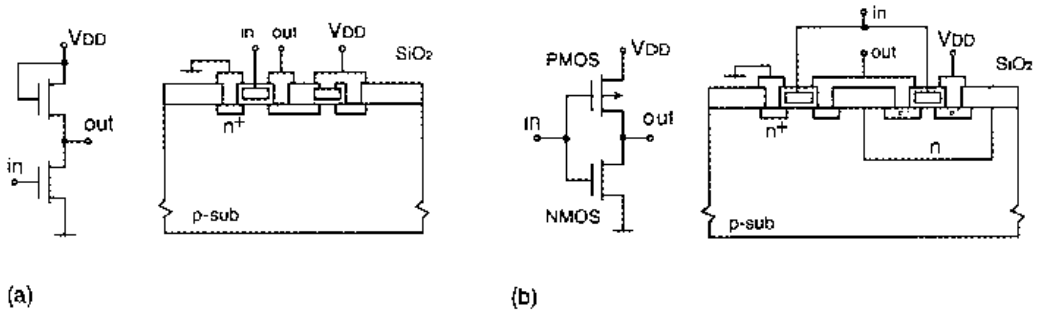


Fig. 2.33. The NMOS inverter (**a**) and the CMOS inverter (**b**) [2.1]

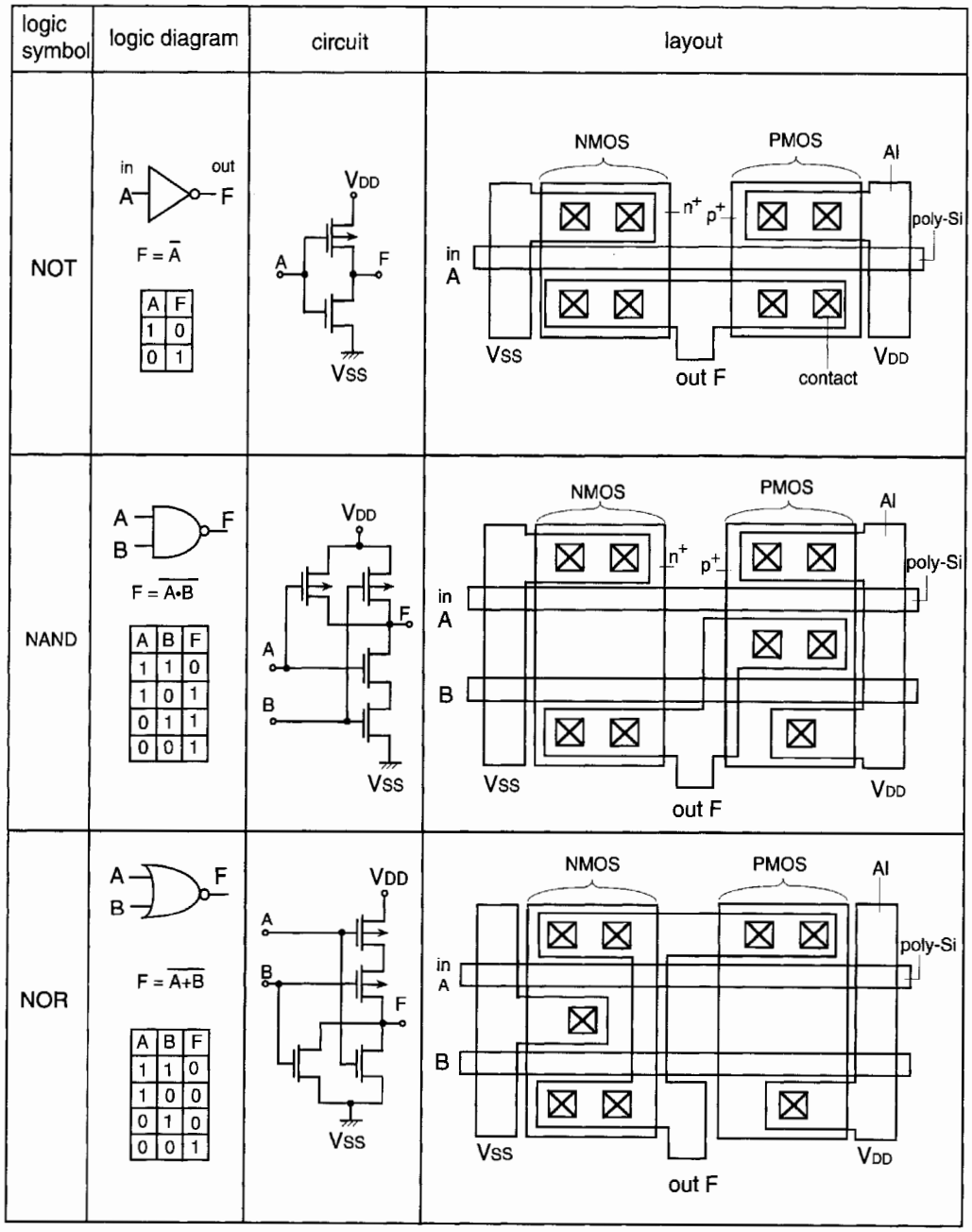


Fig. 2.34. Typical logic gates [2.1]

The Current-Mirror Amplifier. Figure 2.36 shows a basic current-mirror amplifier [2.1] with two input terminals for a differential input and one output terminal (i.e. a single-ended output). The load MOSFETs Q_3 and Q_4 form a current mirror. All MOSFETs are biased to be in saturation. Therefore, Q_5 works as a constant-current (I_S) source. If the sizes of the paired MOSFETs (Q_1 and Q_2, and Q_3 and Q_4) are the same and the differential input is 0 V (i.e. the same input voltage), a current of half of I_S flows equally to the Q_3–

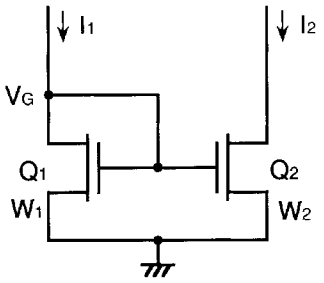


Fig. 2.35. The current mirror

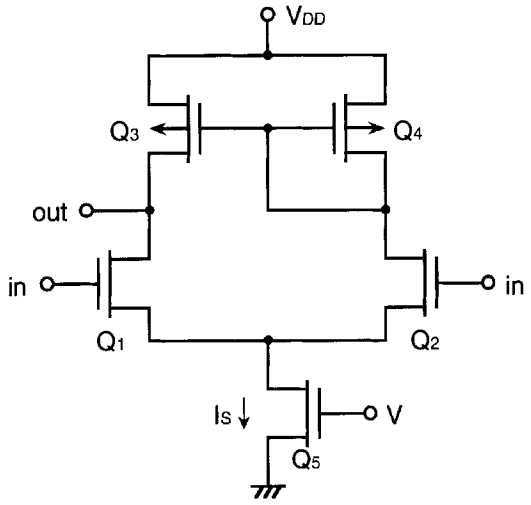


Fig. 2.36. A current-mirror differential amplifier [2.1]

Q_1 path and the Q_4–Q_2 path. The voltage gain [2.4] for a small differential input is given by

$$G = rg_m\,, \tag{2.36}$$

where r is a resistance composed of the parallel-connected output resistances of Q_1 and Q_3 or Q_2 and Q_4, and g_m is the transconductance of Q_1 or Q_2. The resistance r is derived from (2.12) as

$$\begin{aligned} r &= [1/r(Q_1) + 1/r(Q_3)]^{-1} = [(\lambda_N + \lambda_P)I_{DS}]^{-1} \\ &= [(\lambda_N + \lambda_P)I_S/2]^{-1}\,, \end{aligned} \tag{2.37}$$

where λ_N and λ_P are the channel-length modulation parameters for the NMOSFET and the PMOSFET, respectively. The transconductance g_m is derived from (2.11) as

$$g_m = \Delta I_{DS}/\Delta V_{GS} \approx \sqrt{I_S(W_1/L)\beta_{N0}} \tag{2.38}$$

where W_1 and β_{N0} are the channel width and β_0 value of Q_1. The gain G is large because both r and g_m are large. For example, $G = rg_m = 14\,\mathrm{k\Omega} \times 2.2\,\mathrm{mS} = 31$ for $\lambda_N = 0.04\,\mathrm{V}^{-1}$, $\lambda_P = 0.1\,\mathrm{V}^{-1}$, $I_S = 1\,\mathrm{mA}$, $W_1/L = 100$, and $\beta_{N0} = 5 \times 10^{-5}\,\mathrm{S/V}$.

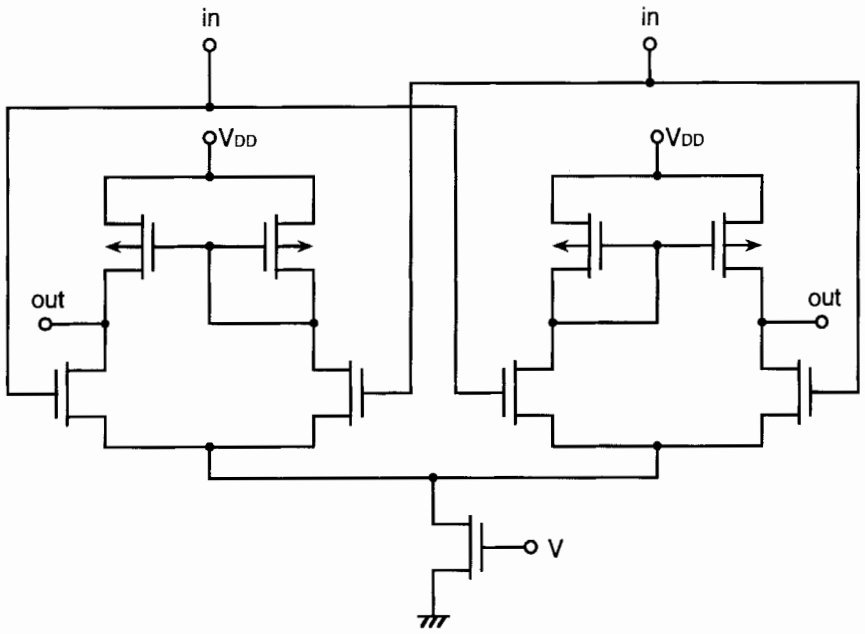


Fig. 2.37. A differential output current-mirror amplifier [2.1]

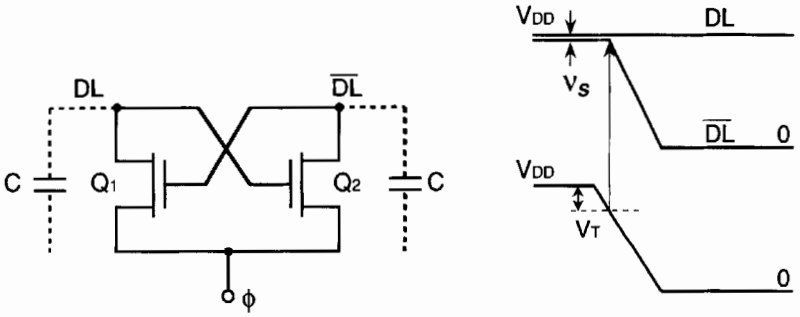


Fig. 2.38. A cross-coupled amplifier [2.1]

In the design of a differential amplifier, the channel lengths of Q_1 and Q_2 are chosen to be large to reduce the offset voltage, which is the difference in V_T between the two MOSFETs, since the offset voltage works as a noise. The resultant offset voltage is usually 20 mV [2.10]. The current-mirror amplifier has the advantages of high speed and an excellent common-mode rejection ratio. However, the voltage gain is still small, because the diode connection of Q_4 prevents the Q_4 drain from taking a large voltage swing. The single-ended output is also a problem. These problems are solved by the use of a differential-output current-mirror amplifier [2.1], as shown in Fig. 2.37. The amplifier effectively doubles the signal swing at the output and enables a differential connection to the succeeding differential circuit.

The Cross-Coupled Amplifier. Figure 2.38 shows an NMOS cross-coupled amplifier [2.1]. A differential small signal v_s, developed between a pair of data lines, on a floating V_{DD} is amplified to V_{DD} by applying an activation pulse

ϕ. Here, the amplification starts when ϕ reaches $V_{DD} - V_T$ and Q_2 is turned on. Then, $\overline{DL}$ continues to be discharged, leaving DL at V_{DD}. In practice, a CMOS amplifier consisting of cascaded NMOS and PMOS cross-coupled amplifiers has been used exclusively as a sense amplifier on each pair of data lines in modern CMOS DRAMs.

2.6.4 The Voltage Booster

Node boosting, by means of a MOS capacitor that easily offers a large capacitance, eliminates the reduction in V_T, as discussed previously, and even generates an increased quasi-dc power-supply voltage.

Gate Boosting. Gate boosting is important not only for the bootstrap inverter, but also for the DRAM word drivers [2.1] shown in Fig. 2.39. Only one of a number of word lines is selectively activated by the decoders. For example, the switches are off after applying V_{DD} only to the selected Q_1 gate and 0 V to other non-selected gates such as the Q_2 gate. The succeeding application of a V_{DD} pulse to the common terminal RX enables a V_{DD} pulse to be outputed only to WL_1, leaving WL_2 at 0 V. The full V_{DD} pulse at WL_1 is due to a sufficient boosting of the Q_1 gate, to about $2V_{DD}$, if the parasitic capacitance at the gate is negligible, due to a large gate–drain capacitance C_{GD} (Fig. 2.9). On the other hand, the small C_{GD} of Q_2 keeps the gate, and thus WL_2, at 0 V despite the application of the RX pulse.

The Increased Power Supply. An increased quasi-dc power supply from a voltage up-converter is indispensable in modern DRAMs for the generation of a high-speed, well-controlled increased pulse, as discussed in Chap. 4. The basic idea for the power-supply generation is to use a charge pump with MOS capacitors.

Figure 2.40 shows the concept of the charge pump [2.1]. Nodes N_1 and N_2 are charged to $V_{DD} - V_T$ in the quiescent state. The application of a pulse

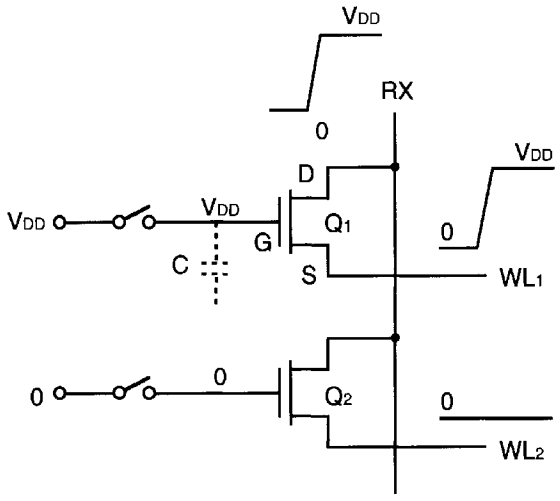


Fig. 2.39. Word-line selection by a MOS capacitor [2.1]

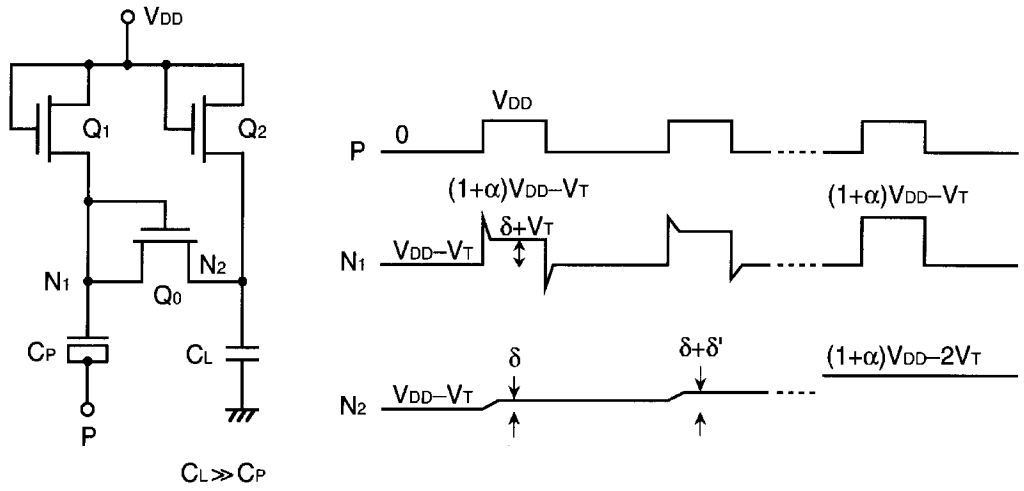


Fig. 2.40. The charge pump [2.1]

to P increases N_1 to $(1+\alpha)\ V_{DD} - V_T$ where α is the boost ratio, which is determined by C_P and the N_1 parasitic capacitance. Thus, Q_0 is turned on and N_1 is charged up. This implies that part of charges (i.e. $C_P V_{DD}$) injected to N_1 by C_P is transferred to N_1. However, the charging stops when the voltage difference between N_1 and N_2 reaches V_T. The resultant voltage change δ at N_1 is small because, usually, $C_L \gg C_P$. When the pulse is turned off, N_1 instantaneously drops to below $V_{DD} - V_T$, and then it is charged up again to $V_{DD} - V_T$ by Q_1. In this manner, successive pulse applications to P allow N_2 to continue to be charged and raised. When N_2 reaches $(1+\alpha)V_{DD} - 2V_T$, however, N_2 stops being charged, thereafter enabling to hold a stable raised voltage on C_L. If there is a pulsive load current at C_L, the increased voltage would be degraded. In this case, Q_0 would again continue to charge up C_L until the resultant voltage reached a stable level. The stable level is 7.5 V for $V_{DD} = 5\,V$, $V_T = 0.5\,V$, and $\alpha = 0.8$. The time necessary to reach the stable voltage depends on C_P, C_L, and the pulse frequency. The pulse can be generated by a ring oscillator.

2.6.5 The Level Shifter

Figure 2.41 shows a level shifter [2.1], to shift from a low-level (V_{DD}) pulse to a high-level (V_{DH}) pulse. Since either of the cross-coupled PMOSFETs may be turned on, depending on the input, there is no ratio operation with an NMOSFET. V_{DH} is generated by the charge pump.

2.6.6 The Ring Oscillator

Figure 2.42 shows a ring oscillator comprising odd-stage inverters [2.1]. It is used for charge-pumping in a voltage up-converter and in substrate-bias

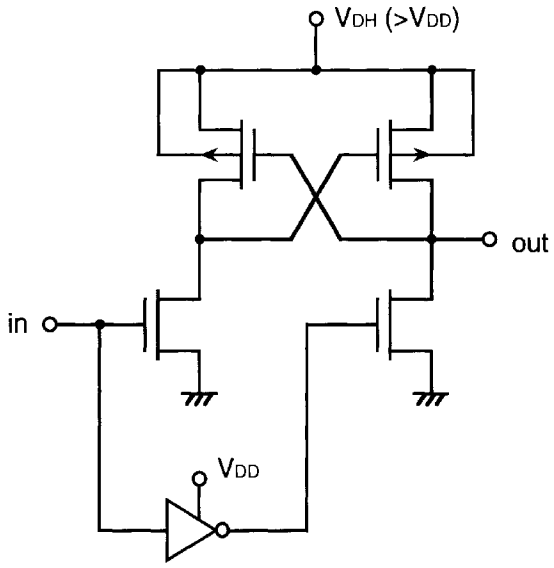


Fig. 2.41. A level shifter [2.1]

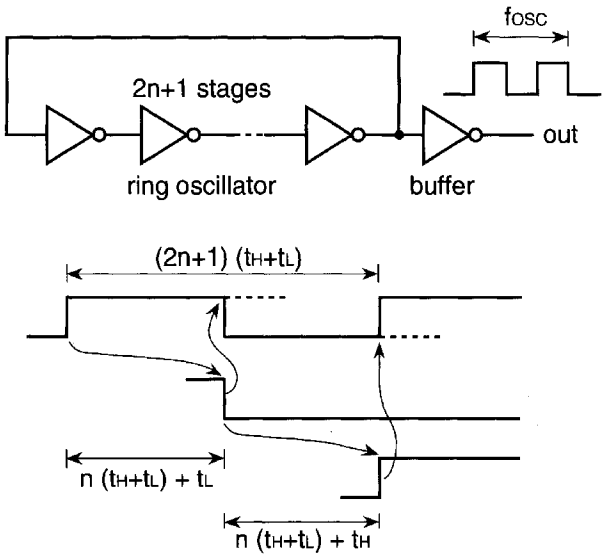


Fig. 2.42. A ring oscillator [2.1]

generators. In addition, it is indispensable in the control of the DRAM refresh time, via μs – ms interval pulses that are generated with the help of counters. The oscillation frequency f_{OSC} [2.1] is given by $[(2n+1)(t_{\mathrm{H}}+t_{\mathrm{L}})]^{-1}$, where $2n+1$ is the number of inverter stages, and t_{H} and t_{L} are the delay times of each inverter for a high-input and low-input pulse, respectively.

2.6.7 The Counter

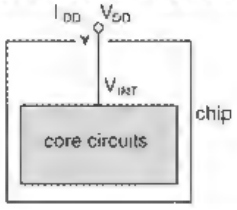
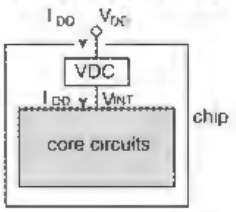
The counter monitors the number of pulses. It is well known that an address counter counts up the number of refresh operations, to select the refresh word lines in order. Another application to DRAMs is the refresh timer, which

counts the number of pulses from a ring oscillator and generates a desired pulse frequency corresponding to a refresh interval.

2.7 The Scaling Law

The performance of the LSI chips is improved by the scaling down of design parameters such as the device size, impurity and threshold voltages of the MOSFETs, and the operating voltage [2.1, 2.18]. The scaling under a constant electric field shown in Table 2.4 is well-known as ideal scaling. There have been other two scalings that are modifications of ideal scaling; constant supply-voltage scaling and the combined scaling of the above two kinds [2.1].

Table 2.4. The impact of the VDC approach on CMOS chip performance under a fixed V_{DD}

	Conventional approach	**VDC approach**
		
Performance		
External voltage, V_{DD}	1	1
Internal voltage, V_{INT}	1	$1/k$ $(k > 1)$
FET dimensions, L_i, W_i, t_{OX}, x_j	1	$1/k$
Electric field for FETs, E	1	1
FET current, I^a_{DSi}	1	$1/k$
Power dissipation, $I_{DD}V^b_{DD}$	1	$1/k$
On resistance, $R_{oni} = V_{INT}/I_{DSi}$	1	1
Delay, $\tau_c = R_{oni}C^c_{Gi}$	1	$1/k$
Circuit area	1	$1/k$

[a] $I_{DSi} = (W_i/L_i)/(V_{INT} - V_T)^2/t_{OX}$.
[b] $I_{DD} = \sum I_{DSi}$.
[c] $C_{Gi} \propto L_i W_i/t_{OX}$.

2.7.1 Constant Electric-Field Scaling

In this scaling method [2.1, 2.18], the physical size and threshold voltage V_T of the MOSFET, the wiring (i.e. interconnection) in Fig. 2.43, and the power supply V_{DD} are scaled down by a factor $k(> 1)$. The substrate doping concentration N of the MOSFET is increased by the same factor. This is due

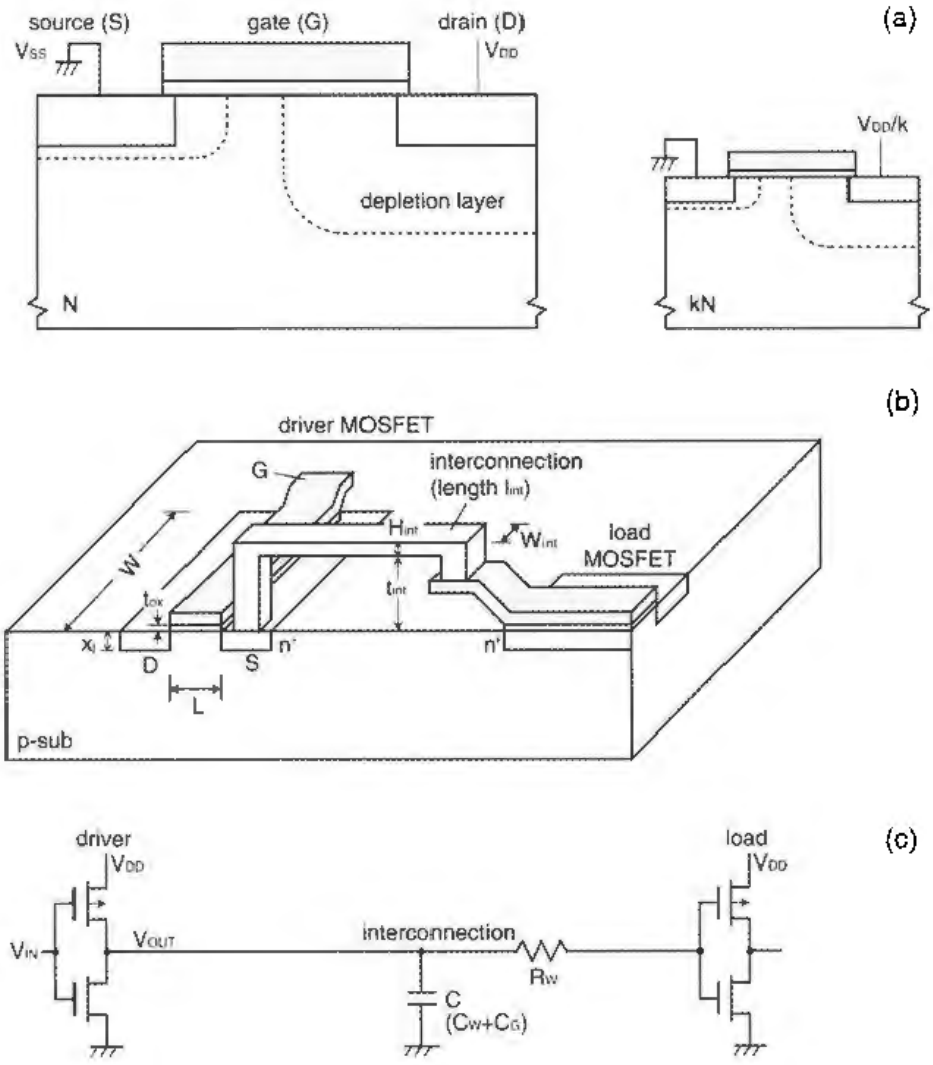


Fig. 2.43. Scaling in MOS LSIs [2.1]. (**a**) The scaling of a MOSFET ($k = 2$); (**b**) the conceptual structure of a MOS LSI; (**c**) the basic circuit

to suppression of the resulting short-channel effects, such as drain–source punch-through, by decreasing the depletion length, which is proportional to $\sqrt{V_{\mathrm{DD}}/N}$, to $1/k$. As a result, the scaling factor of the MOSFET current I_{DS}, which is proportional to $(W/L)(V_{\mathrm{DD}} - V_{\mathrm{T}})^2/t_{\mathrm{OX}}$, is also $1/k$. This scaling enables not only high density, but also high performance, while maintaining device reliability under constant electric field conditions: the delay, power dissipation, and power–delay product of a circuit, exclusive of wiring capacitances, are improved by $1/k$, $1/k^2$, and $1/k^3$, respectively. The charging power of the load is also reduced by $1/k^3$, with a reduction in the load capacitance of $1/k$. However, the drawbacks are as follows. The wiring resistance is increased by a factor of k, which makes the wiring delay more prominent as compared to the circuit delay. The voltage drop along a power-supply line cannot be scaled

down, which makes the voltage margin of the circuit narrow. The reliability of the metal wiring is degraded because of electon migration caused by the increased current density.

In practice, there are some parameters [2.1] that depart from ideal scaling. Because of the issue of V_{DD} standardization, V_{DD} cannot be scaled down, as discussed later. To partly solve the issues regarding resistance and the current density of the wiring, the thickness and width are not necessarily scaled down in accordance with the scaling law. Other non-scalable parameters, such as parasitic source and drain resistances, contact (through-hole) resistances, short-channel effects, and device-parameter fluctuations, can also degrade the MOSFET driving current. The MOSFET threshold voltage V_T is noteworthy. Excessive scaling down of V_T is not allowed because of an unacceptably large subthreshold current, as discussed in Chap. 8.

2.7.2 Constant Operation-Voltage Scaling

In the past, DRAM has maintained the power-supply voltage V_{DD} at the same level for as long as possible, to solve the V_{DD} standardization issue, as discussed in Chap. 5. For example, a 5 V V_{DD} was used for four generations, of 64 Kb to 4 Mb, despite the successive scaling down of internal devices. Even small devices for the 4 Mb chip were tailored to withstand 5 V V_{DD} operation with the help of stress-voltage-immune MOSFET structures such as LDD. In this scaling, the MOSFET delay is improved by a factor of $1/k^2$. However, the electric field and power dissipation of the MOSFET, and the current density and voltage drop along the wiring are degraded by k and k^3, respectively. The resulting serious problems are velocity saturation, conductance degradation, degraded reliability due to hot carriers, and the gate–insulator and p–n junction breakdown of the MOSFET. In addition, electromigration in metal wiring, CMOS latch-up, and noise are also problems. To reduce the above detrimental effects, other constant-V_{DD} scalings [2.1, 2.19], such as a combination of $1/\sqrt{k}$ for gate-oxide scaling and k^2 for doping concentration scaling, have been proposed.

2.7.3 Combined Scaling

This scaling necessitates an on-chip voltage-down converter (VDC approach in Table 2.4). The converter adjusts the internal supply voltage V_{INT} according to the lowering of the breakdown voltage of the scaled devices, so that the electric field of each device is held constant, while preserving V_{DD} at the same level. This scaling solves almost all of the problems involved in constant-V_{DD} scaling, although the power dissipation is larger than that of constant-electric-field scaling. Thus, this scaling has been widely used in modern DRAMs.

Based on the above discussion, here is an example of the estimated performance of a hypothetical chip [2.1], in which only the size of the MOSFET,

V_{DD}, and V_T are reduced to 1/3 (i.e. $k = 3$), while other sizes are kept constant, so that the wiring capacitances and the chip size are preserved. We assume that the wiring capacitance C_W and the input capacitance (i.e. the gate capacitance C_G) of circuits connected at the load occupy 35% and 65% of the load capacitance of an average circuit (see Fig. 2.13), respectively. Hence,

$$I_{DS} = \mu C_{OX} \frac{W}{L} \frac{(V_{GS} - V_T)^2}{2} \propto 1 \times k \times \frac{1/k}{1/k} \times (1/k)^2 = 1/k\,,$$

$$V_{DD} \propto 1/k\,,$$

$$C_G = \varepsilon_{OX} \frac{W_N L_N + W_P L_P}{t_{OX}} \propto 1/k\,,$$

$$C_W \propto 1\,, \qquad C = C_W + C_G \propto 0.35 + 0.65/k\,.$$

If the wiring resistance and the dc current of the circuit are negligible, the delay τ_D and power dissipation P at a fixed frequency are given by

$$\tau_D = V_{DD} C / I_{DS} \propto 0.35 + 0.65/k = 0.57 \qquad (k = 3)\,,$$

$$P = C V_{DD}^2 \propto (0.35 + 0.65/k)/k^2 = 0.06\,.$$

Obviously, the speed and power of an average circuit (that is, those of the chip) are improved dramatically when $k = 3$.

2.8 Lithography

Figure 2.44 shows trends in the minimum feature size of devices [2.20]. In the past, a continuing improvement in the resolution of optical lithography has enabled a size reduction of 60–70% at each generation. For the 64 Kb–256 Kb generations, contact printing or one-to-one projection printing were standard. Contact printing uses a one-to-one photo mask, which is prepared for all of the chips on a wafer. On the photo mask the patterns of each chip is reapeted by the number of chips on the wafer. Thus, the patterns of each chip are simultaneously exposed at the same size as those of the mask through the mask, which is contacted with the wafer. One-to-one projection printing reduces the mask damage caused by contact printing, because the mask is separated sufficiently from the wafer. Ultraviolet rays corresponding to the g line (wavelength of 436 nm) and the h line (405 nm), generated by a mercury lamp, were originally used. However, this method was overtaken by reduction projection printing as the patterns were miniaturized. In this printing method, a photo mask, on which the patterns are magnified in size by five times the dimensions of the actual chip, in order to relax the requirements for size accuracy of the mask pattern, is prepared for only one chip. The patterns on the mask are projected on to the wafer, with a size reduction of one-fifth. This projection for one chip is repeated one by one throughout the wafer. Reduction projection printing has improved the resolution by means of an increase

in the numerical aperture number (NA) of the optical lens and a shortening of the wavelength from the g line to the i line (365 nm), and to KrF (248 nm) and ArF (193 nm), which can be generated by excimer lasers. In addition, the method even achieves a resolution of a half-wavelength through improvements in mask fabrication and photo illumination techniques. Consequently, the 256 Mb and 1 Gb generations may be produced through the optimization of optical lithography technology. Electron-beam and X-ray lithographies for higher-resolution patterning are also being intensively developed for 1 Gb and beyond.

2.9 Packaging

High-density packaging technology has been crucial in reducing the volume of assembled memory systems with an accompanying reduction in the package volume and footprint. Figure 2.45 shows trends in high-density packaging. The rapid and remarkable progress in packaging can be summarized into two ways. One trend is toward thinner, smaller, and lighter packages. The other is toward high-pin-count packages. Traditional packages such as the DIP (Dual-In-Line Package) were followed by successively higher-density packages, such as the SOP (Small-Outline Package), the SOJ (Small Outline J-leaded Package), the SSOP (Shrink Small Outline Package), and the TSOP (Thin Small-Outline Package). It is noteworthy that the LOC (Lead-On-Chip) packaging shown in Fig. 2.46 has accommodated the ever-larger chips

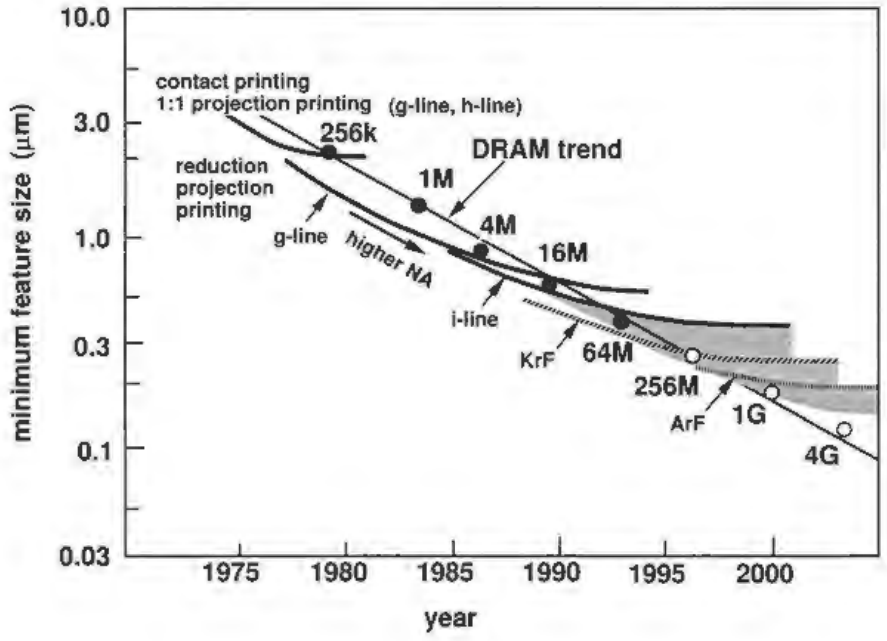


Fig. 2.44. Trends in lithography technology [2.20]

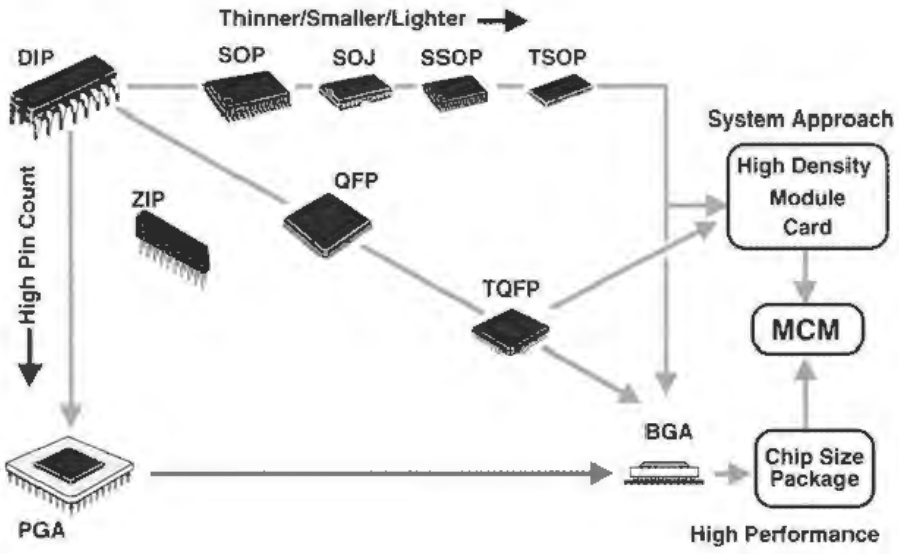


Fig. 2.45. Trends in high-density packaging

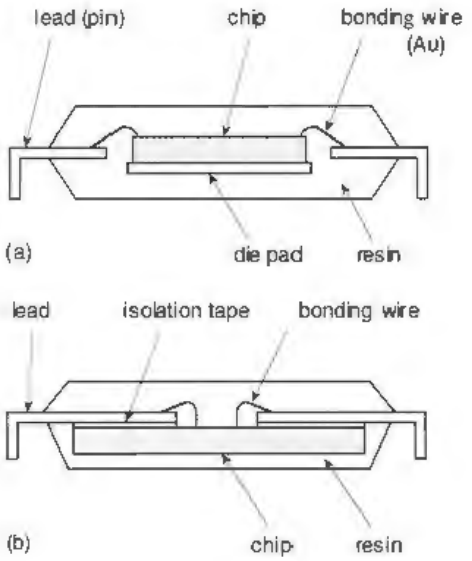


Fig. 2.46. The advancement of small packaging [2.21]. (**a**) Conventional packaging; (**b**) LOC packaging

in a small package. The traditional packages were also followed by the PGA (Pin-Grid-Array) package. The QFP (Quad Flat Package) and TQFP (Thin Quad-Flat Package) are combinations of the above two trends. Recently, the BGA (Ball-Grid-Array) package has been regarded as an important solution for high performance. Further high density is realized by the CSP (Chip-Size Package). Another trend that decreases the volume of memory systems further is toward modules that contain several of these packages. In fact, a high-density module card is now being launched on the DRAM market.

A final solution, however, might be the MCM (Multi-Chip Module), which is a combination of the CSP and the module card. Regarding high-performance packages, a low-noise and high-density package design focused on simultaneous I/O switching noise, crosstalk, and reflection noise is essential. Thus, in addition to reducing the parasitic inductance and capacitance, an accurate circuit model for packages and modules is indispensable. A low-thermal-resistance plastic package is also a key to ensuring an adequately low DRAM-cell leakage current and high reliability with a reduced junction temperature, while maintaining low cost.

3. DRAM Circuits

3.1 Introduction

The intensive research and development (R&D) directed toward DRAMs has rapidly increased memory-chip capacity by more than six orders (1 Kb to 4 Gb) at the R&D level over the past 30 years [3.1] since the advent of DRAMs in the early 1970s. As a result, 64–256 Mb DRAMs are now at the volume-production level. Such rapid progress in DRAM density has led to their playing an important role in enhancing the performance and reducing the cost of electronic systems such as large computers, workstations, personal computers (PCs), and so on. The quadrupling of memory capacity, as shown in Fig. 3.1, through high-density technology has contributed to this progress.

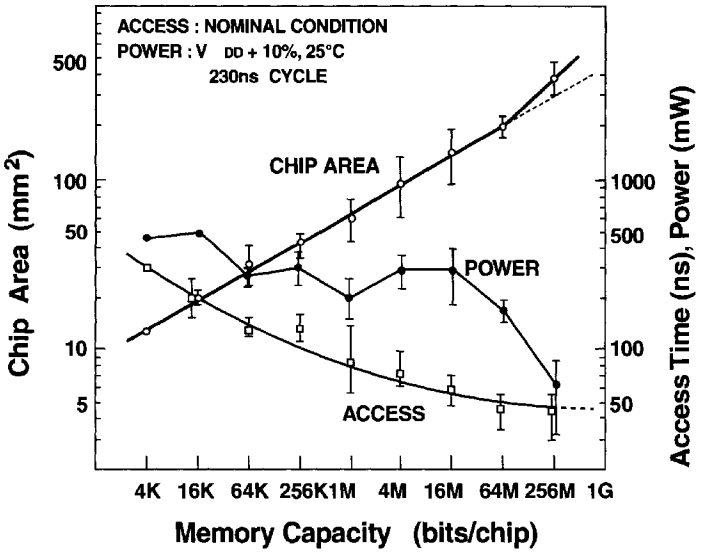


Fig. 3.1. Trends in RAM chip performance. The bit-width for the I/O pin is mostly 1 bit or 4 bit for DRAMs [3.1]

High-performance circuits aimed at higher speed and lower power dissipation, which have been achieved in the past despite the ever-increasing chip area with increased memory capacity (Fig. 3.1), have also benefited system designers.

The following summaries focus on basic technological developments in high-density and high-performance circuits, although advances in CAD/DA, testing, and high-density packaging are also noteworthy.

3.1.1 High-Density Technology

Figure 3.2 shows the trend in DRAM chip-area and wafer-area from 4 Kb to 1 Gb [3.2]. The chip area has been increasing by 1.5 times for each successive generation up to 64 Mb. This is the so called memory capacity quadrupling approach. However, experimental 256 Mb and 1 Gb DRAM chips have broken the trend, indicating difficulty in device miniaturization. Although the chip-area increase should be compensated by wafer-area enlargement, the ever increasing investment cost of fabrication facilities have hindered the change of wafer area. Thus, to reduce bit-cost, achieving smaller chip areas is an urgent task for DRAM manufacturers; this can be clearly seen in a succession of drastic chip-area-shrinks for 64 Mb DRAMs in Fig. 3.2. This is the so-called chip-shrink approach. A 64 Mb DRAM with a chip area of less than 40 mm^2 is now going into mass production using 0.18 µm technologies that were considered necessary for 1 Gb DRAM production a few years ago. In any event, both the memory capacity quadrupling and chip-shrink approaches are supported by high density technology.

Figure 3.3 shows the advances [3.4]. Fine-pattern technology has been supported by advances in photoaligner and etching technology. The photoaligner has advanced in the form of the contact aligner until the 64 Kb generation, as the one-to-one projection aligner at the 256 Kb generation, and as the five-to-one projection aligner at the 1 Mb generation and beyond. With regard

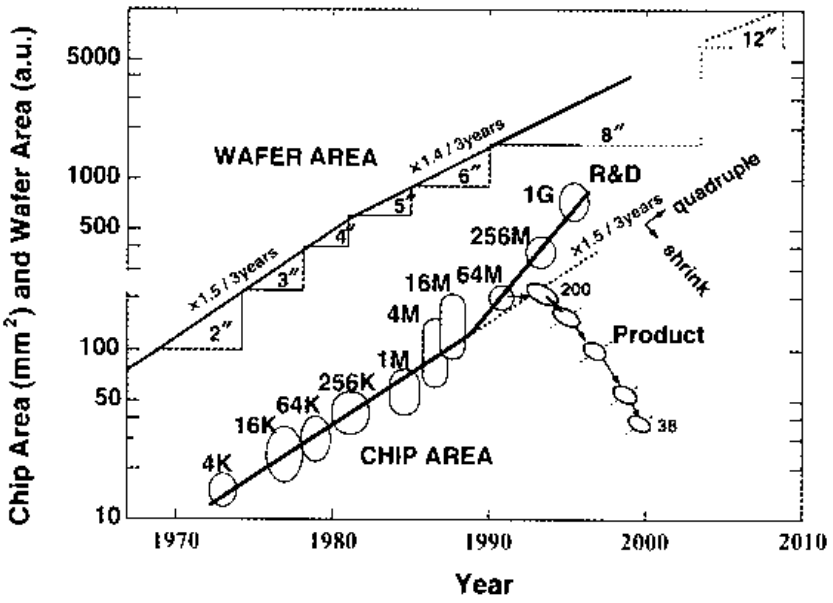


Fig. 3.2. Trends in DRAM chip and wafer areas [3.2]

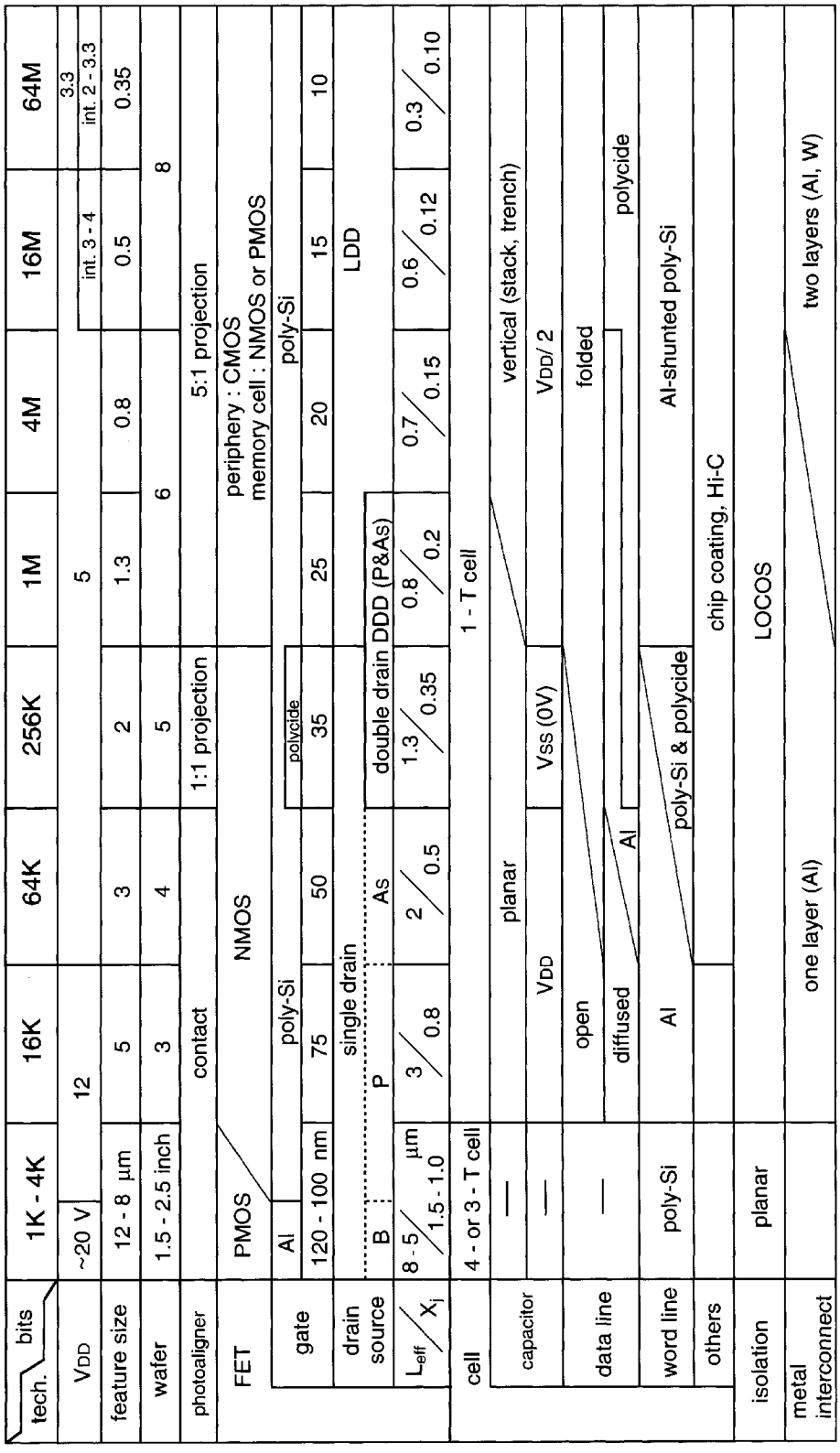


Fig. 3.3. Trends in high density DRAM technology [3.4]

to etching technology, wet etching had been popular until the 5−3 μm era. However, to overcome the issue of undercutting related to wet etching, dry etching has been widely used since the 3−2 μm era.

With regard to device technology, this has been strongly influenced by the stringent need for a high-density memory cell, a high speed and standardization of the power-supply voltage. One way of achieving smaller memory cells, as well as smaller peripheral circuits, is to add wiring layers to relax the size-reduction requirement for each layer. Good examples of this are double poly-Si layer in the 16 Kb generation and the double metal layer. The polycide layer, consisting of low-resistance silicide stacked on to high-resistance poly-Si has reduced word-line delay. Advances in the MOS transistor are noteworthy. Transistor miniaturizations based on the scaled-down theory have made a great impact on chip performance at each generation, as discussed in the previous chapter. To achieve even higher speed and lower power, the transistor structure was changed from PMOS to NMOS in the 4 Kb generation, and then to CMOS in the 1 Mb generation. The source/drain structure was also changed, depending on the size and the power-supply voltage. In the 64 Kb generation the dopant was changed from phosphorus to arsenic, which is suitable for miniaturization due to its smaller diffusion coefficient. Furthermore, the LDD structure (see Fig. 5.21) to relax the stress voltage at the drain has been widely used since the 1 Mb generation, enabling 5 V direct operation even for small transistors in the 4 Mb generation.

As for the memory cell, miniaturization has progressed while maintaining the signal charge to the greatest possible extent, in spite of reduction of the cell area to around 1/3 that of the previous generation, as discussed in Chaps. 1 and 4. The half-V_{DD} plate, three-dimensional capacitors, high-permittivity insulator films, and the folded data(bit)-line arrangement have contributed to the maintenance of signal charge. To decrease soft errors, the polyimide chip coating has been indispensable.

3.1.2 High-Performance Circuits

Circuit design has concentrated on achieving a high signal-to-noise ratio (S/N), together with low-power and high-speed operation. The absence of gain in the one-transistor, one-capacitor (1-T) cell causes a small read signal voltage, and needs a subsequent rewrite operation with a large voltage swing and refresh operations. The resultant drawbacks of the low S/N ratio, high power, and slow speed of the DRAM chip are made worse by the ever-larger chip area. Differential sensing, word bootstrapping for a full write and full read, shared sense amplifier for doubling the signal voltage while halving the data-line power, multidivided data line with the assistance of multilevel metal wiring to solve all of the above drawbacks, the folded data-line arrangement, and transposition of the data line are the well-known high-performance (especially high-S/N) circuit techniques, as summarized in Fig. 3.4 and discussed in this chapter. NMOS dynamic circuits for sense amplifiers and peripheral

tech. \ bits	1K - 4K	16K	64K	256K	1M	4M	16M	64M
V_{DD}	~20 V	12	5					3.3
V_{BB}	external (-5 V)		internal (-3 V - -1 V)					
address	separated X/Y	address multiplex						
high-speed modes	page	nibble	static column				burst	
data bits	x 1 bit						x n bits (n≥4)	
high S/N	diff. sense	word boost folded data line	shared amp		divided data line		transposed data line	
low power	dynamic NMOS (amp, driver)				CMOS V_{DD}/2 precharge shared I/O		voltage down converter	
high speed	Al-shunted poly-Si word line							hierarchical word line
others	on-chip V_{BB}	redundancy		parallel test			boosted power synchronous DRAM	

Fig. 3.4. Trends in DRAM circuit technology [3.4]

drivers had contributed to power reduction in the NMOS–DRAM era. In addition to CMOS circuits, half-V_{DD} data-line precharge combined with the CMOS sense amplifier, shared I/O, and an on-chip voltage down-converting scheme have reduced the power dissipation of multimegabit CMOS DRAMs. In addition to the use of polycide instead of poly-Si for the word line, the metal-strapped poly-Si (or polycide) word line and the multidivided word line have enhanced chip speed with reduced word-line delay.

Address multiplexing and redundancy have been indispensable for high-density packaging and low cost. Internal parallel testing, including built-in self-testing (BIST), has also become increasingly important to ensure low cost. On-chip voltage conversion circuits, such as the substrate bias-voltage generator and the voltage up-converter are also essential for the stable and low-power operation of internal circuits. The conversion circuits also help to enable ultra-low-power operations, as will be explained in Chap. 8.

In this chapter, catalog specifications exemplified by the basic operational modes of a standard 16 Mb DRAM, and the basic configuration and operation of a DRAM chip, are clarified in order to understand the relationship between the specifications and internal circuit operations. Next, design issues and solutions to meet the specifications are described conceptually. Then, basic circuit blocks consisting of a chip are discussed in terms of S/N, power dissipation, and speed. We discuss multidivision of a memory array, which is closely related to almost all aspects of the performance of the chip, read/write circuits, refresh-relevant circuits, redundancy, and on-chip testing circuits. More details and variations for high-S/N design, on-chip voltage conversion circuits, high-performance subsystem memories such as synchronous DRAM (SDRAM), and low-power and ultra-low-voltage circuits are described in other chapters.

3.2 The catalog Specifications of the Standard DRAM

Since the standard DRAM incorporates an address multiplexing scheme, it requires complicated timing specifications between the input and output signals. In addition to the timing, the voltage, current, temperature, and modes of operation are specified in detail and standardized. In the following sections, some specifications are clarified [3.4], using the example of a CMOS 16 Mb DRAM.

3.2.1 Operational Conditions

Package and Pin Assignment. Progress in memory packages toward smaller, thinner and lighter products has been remarkable given the ever-increasing chip area, as explained in the previous chapter. For the 16 Mb DRAM, the first generation used a 400-mil (1 mil = 0.001 inch) width,

28-pin SOJ (Small Outline J-leaded Package, with external dimensions of $10.2 \times 18.2 \times 3.5\,\text{mm}$), in which a chip of $127\,\text{mm}^2$ ($= 8.15 \times 15.58$) was accommodated. Due to a scaling down of the above chip to 78%, the second generation could be accommodated in a smaller 300-mil width, 26-pin TSOP (Thin Small Outline Package, 1.2 mm thick), which is the same size as in the previous 4 Mb chip, so as to be replaceable on a memory board. Figure 3.5 shows the pin assignment of a 16 Mb with a 16 MW(words)×1 b configuration [3.4]. In the case of a wider bit configuration, one pin is assigned to both D_{in} and D_{out}, to suppress the increase in the pin count (see Fig. 7.5).

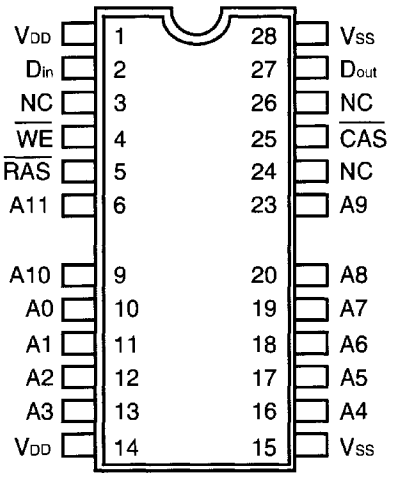


Fig. 3.5. A top view of the SOJ package and pin assignment of a 16 Mb DRAM [3.4]. NC; no connection

The Power-Supply Voltage and Current. Table 3.1 shows the catalog specifications. $V_{DD} = 5\,\text{V}$ is the standard power supply for a 16 Mb DRAM with an acceptable V_{DD} variation of ±10%, although major internal circuits, such as peripheral circuits and the memory array, operate at an internal supply V_{DL} of about 3.3 V, which is generated from V_{DD} by an on-chip voltage down-converter. The access time t_{RAC} in the worst conditions, developed across the full range of V_{DD}, temperature, cycle time, operational modes, and so on, is 60 ns at a D_{out} capacitance of 100 pF. Access time, however, ranges widely with fabrication-process variations during volume production, allowing the specifications to be sorted according to the resultant spread. The set of data in the table is for the fastest version, although there are three versions corresponding to 60, 70, and 80 ns access times. Note that there is quite a large difference between the catalog specifications and specifications presented at conferences, which are usually for a nominal condition for a typical sample. The operational current is 80 mA max. at the maximum cycle of 110 ns. The current is almost inversely proportional to the cycle time t_{RC}, because the ac current dominates it. The refresh current at a cycle time of 31 µs is 300 µA, and self-refresh current is 200 µA.

Table 3.1. The specification of a 16 Mb chip [3.4]

Items	Symbols	Specifications		Unit	Test conditions
		min.	max.		
Ambient temperature	T_a	0	70	°C	
External supply	V_{DD}	4.5	5.5	V	
Access time	t_{RAC}	–	60	ns	D_{out} load 2TTL + 100 pF
Cycle time	t_{RC}	110	–	ns	
Operating current	I_{CC1}	–	80	mA	$t_{RC} = 110$ ns, D_{out} open
Stand-by current (1)		–	300[a]	μA	CBR refresh, $t_{RC} = 31.3$ μs
Stand-by current (2)		–	200[a]	μA	Self refresh
Input logic interface	V_{IH}	2.4	6.5	V	
	V_{IL}	−1.0	0.8	V	
Output logic interface	V_{OH}	2.4	V_{DD}	V	D_{out} −5 mA
	V_{OL}	0	0.4	V	D_{out} 4.2 mA
Address pin capacitance		–	5	pF	
Clock pin capacitance		–	7	pF	
D_{out} pin capacitance		–	7	pF	
Refresh time	t_{REF}	–	64, 128	ms	
Refresh cycles	n	4096	–		

[a] Tentative.

The Ambient Temperature and the Junction Temperature. Ambient temperature T_a ranges from 0 °C to 70 °C in the catalog. However, the chip characteristics are governed by the junction temperature T_j rather than by T_a. Their relationship is expressed as follows:

$$T_j = T_a + \theta_{ja} P \tag{3.1}$$

where θ_{ja} is the thermal resistance of the package and P is the power dissipation of the chip. Here, let us estimate the range of T_j. The minimum T_j ($= T_{jmin}$) is developed when $T_a = 0$ °C, $V_{DD} = 4.5$ V, and the lowest cycle time – that is, the refresh cycle in data-retention mode – are combined. In this case, $T_{jmin} = T_a = 0$ °C because $P \simeq 0$. The maximum T_j ($= T_{jmax}$) is developed when $T_a = 70$ °C, $V_{DD} = 5.5$ V, and the cycle time is at its fastest (i.e. 110 ns), enabling a T_{jmax} as high as 100 °C by assuming that $\theta_{ja} = 60$ °C/W for a TSOP mounted to a board in still air and $P = 0.44$ W. Therefore, the variation in T_j is as great as 100 °C. Even given such a variation, the access time and refresh time (t_{REFmax}), which are quite sensitive to variation, are guaranteed.

Interfaces of Input and Output Signals. The well-known TTL logical interface is used. The minimum high-level V_{IH} is 2.4 V, while the maximum

low-level value is 0.8 V. Hence, for example, a binary "1" or "0" is discriminated for an input higher than 2.4 V or an input lower than 0.8 V, respectively. Such is the case for D_{out}. Note that a minimum V_{IL} of -1 V is catered for in case the logic input waveforms on a memory board become subject to undershoots when many memory packages are mounted on the board.

3.2.2 Modes of Operation and Timing Specifications

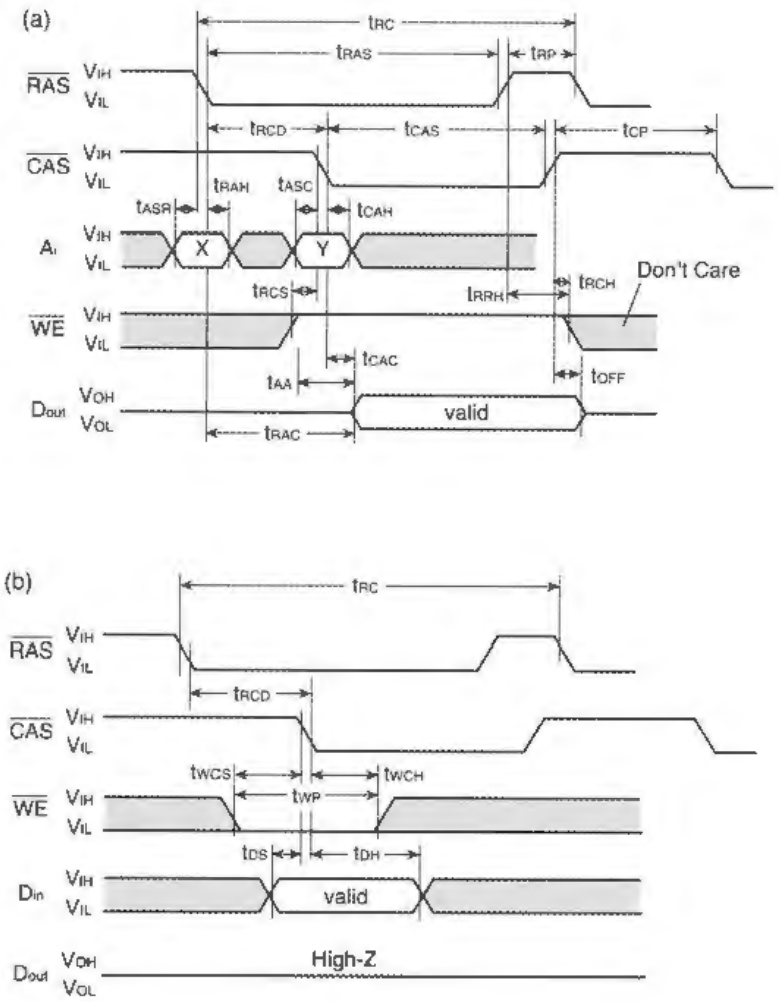
The relationship between timing specifications and chip designs will be described, and illustrated by read, write, and refresh operations, and by high-speed column modes.

The Read Operation. Figure 3.6a shows a timing diagram of the read operation. When $\overline{RAS}$ increases to a high level, the precharge operations of all the row-relevant circuits start, and complete within time t_{RP}. During this period, no memory operations are performed inside the chip. $\overline{CAS}$ stays at a high level during t_{CP} to precharge column-relevant circuits such as D_{in}/D_{out} buffers, thus inhibiting any communication external to the chip. As soon as $\overline{RAS}$ falls to a low level, memory operations start with the activation of row circuits. Subsequently, the $\overline{CAS}$ falling edge enables read or write operations to start using column interface circuits. The $\overline{RAS}$ active operations continue for t_{RAS}. The $\overline{RAS}$ cycle time t_{RC}, which is usually the cycle time of the chip, is $t_{RAS} + t_{RP} + 2t_T$ where t_T is the rise and fall time of $\overline{RAS}$, 5 ns in the catalog. According to the catalog the cycle time can vary from 110 ns to 10 µs. The read operation starts by increasing $\overline{WE}$ to a high level before $\overline{CAS}$ low-level-activation, and holding it for the $\overline{CAS}$ activation period t_{CAS}. The resultant D_{out} is held until $\overline{CAS}$ is increased. Note that there are three kinds of access time: the $\overline{RAS}$ access time t_{RAC}, which is the $\overline{RAS}$ to D_{out} delay; the $\overline{CAS}$ access time t_{CAC}, which is the $\overline{CAS}$ to D_{out} delay; and the address access time t_{AA}, which is the column address to D_{out} delay.

The Write Operation. Figure 3.6b shows a timing diagram of the write operation. The write operation starts by reducing $\overline{WE}$ to a low level, unlike the read operation, before $\overline{CAS}$ low-level-activation. The other timing relationships are same as for the read operation. During the write operation, D_{out} is held to a high impedance (high-Z).

Refresh Operations. The standard refresh operations can be categorized into normal refreshings, such as the $\overline{RAS}$-only refresh and the CBR ($\overline{CAS}$ before $\overline{RAS}$) refresh, which are carried out by interrupting random operations mixed with read and/or write, and self-refreshing only for data retention in the battery back-up mode.

Figure 3.7a shows a timing diagram of the $\overline{RAS}$-only refresh. By applying the refresh addresses (A_0–A_{11}) to the package address pins, each memory cell on the selected word line is simultaneously refreshed with a read–rewrite operation. There are two specifications of t_{REFmax}, 64 ms and 128 ms with

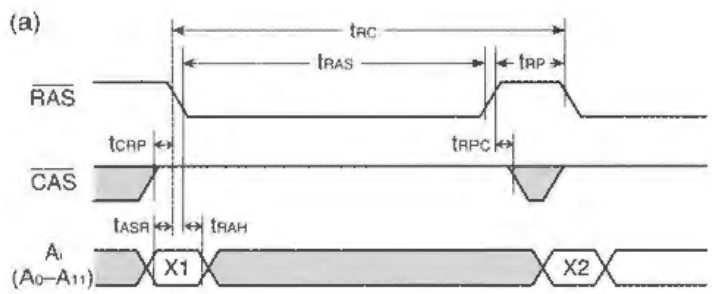


symb.	spec. (ns)	
	min.	max.
tRC	110	—
tRP	40	—
tRAS	60	10 µs
tCP	10	—
tCAS	15	10 µs
tRCD	20	45
tASR	0	—
tRAH	10	—
tASC	0	—
tCAH	15	—
tRAC	—	60
tCAC	—	15
tAA	—	30
tRCS	0	—
tRRH	5	—
tRCH	0	—
tOFF	—	15
tWCS	0	—
tWCH	15	—
tWP	15	—
tDS	0	—
tDH	15	—

Fig. 3.6. Timing diagrams for a 16 Mb DRAM [3.4]. (**a**) Read cycle; (**b**) write cycle

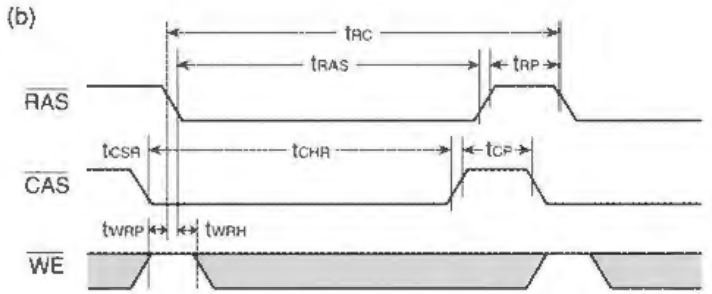
same number of refresh cycles, $n = 4096$. The number n refers to the number of word lines in a logical memory array, as will be explained later. Thus, 4096 word lines must be sequentially selected within t_{REFmax} by combinations of the 12-bit refresh addresses above. This can be done by lumped selection or by distributed selection. Lumped selection repeats the 4096-word-line selections at the fastest cycle time (i.e. 110 ns). During a resulting period of 0.4 ms, the chip cannot accept any random operation from outside. Instead, during the remaining period, $t_{\mathrm{REFmax}} - 0.4$ ms, it can accept any operation without needing to be refreshed. Distributed refresh distributes the 4096 refresh operations uniformly within t_{REFmax}, enabling a refresh time interval of t_{REFmax}/n. Thus, for example, the interval is 16 µs for $t_{\mathrm{REFmax}} = 64$ ms. Since distributed refresh has become popular, one refresh cycle usually interrupts random operations every 16 µs.

Figure 3.7b shows a timing diagram for the CBR refresh operation. $\overline{\mathrm{CAS}}$-before-$\overline{\mathrm{RAS}}$ timing is internally detected, and the resulting pulse activates



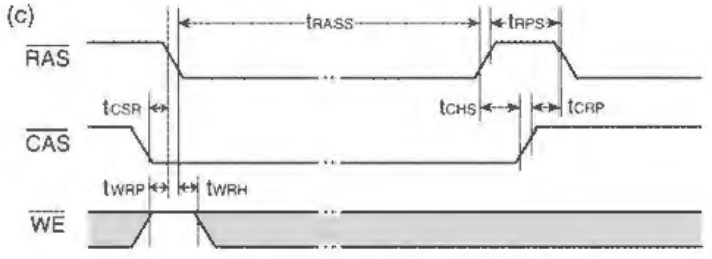
symb.	spec. (ns)	
	min.	max.
t_{RC}	110	—
t_{RP}	40	—
t_{RAS}	60	10 μs
t_{CRP}	5	—
t_{RPC}	0	—
t_{ASR}	0	—
t_{RAH}	10	—
t_{REF}	—	64/128 ms

WE : Don't care
n = 4096



symb.	spec. (ns)	
	min.	max.
t_{RC}	110	—
t_{RP}	40	—
t_{RAS}	60	10 μs
t_{CSR}	10	—
t_{CHR}	20	—
t_{CP}	10	—
t_{WRP}	10	—
t_{WRH}	10	—
t_{REF}	—	64/128 ms

address : Don't care
n = 4096



symb.	spec. (ns)	
	min.	max.
t_{RPS}	110	—
t_{RASS}	100 μs	—
t_{CSR}	10	—
t_{CHS}	−50	—
t_{CRP}	5	—
t_{WRP}	10	—
t_{WRH}	10	—

address : Don't care

Fig. 3.7. Timing diagrams for the refresh operations of a 16 Mb DRAM [3.4]. (**a**) $\overline{\text{RAS}}$-only refresh; (**b**) CBR ($\overline{\text{CAS}}$-Before-$\overline{\text{RAS}}$) refresh; (**c**) self refresh

an on-chip refresh address counter, so that a word line is activated by the refresh address. The refresh is performed without external address signals. The cycle time of external clocks is 16 μs for $t_{\text{REFmax}} = 64$ ms.

Figure 3.7c shows a timing diagram for the self-refresh operation. The width of the $\overline{\text{RAS}}$ pulse, t_{RASS}, after normal operations is extended to over 100 μs with CBR timing. After time t_{RASS}, a refresh operation starts by using an on-chip refresh-address counter and a refresh timer, as explained later. Self-refreshing continues as long as $\overline{\text{RAS}}$ and $\overline{\text{CAS}}$ are held at a low

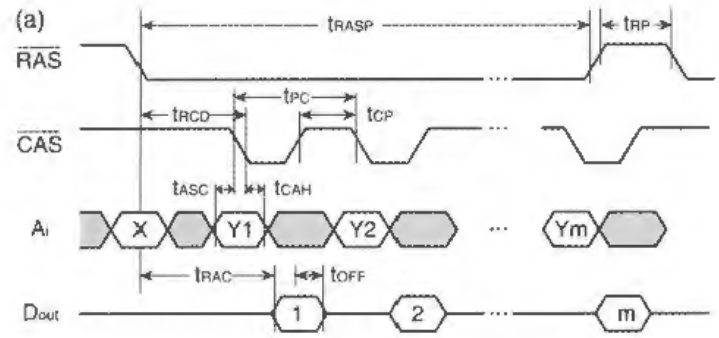
level. The refresh interval, which is very sensitive to T_j, is automatically determined by a timer that detects T_j. After the $\overline{\text{RAS}}$-precharge time t_{RPS}, the self-refresh mode is released to a normal random mode.

High-Speed Column Modes. Column modes provide a higher speed, and in particular a higher throughput for some systems, such as cache-application systems and graphics, as discussed in Chap. 6. The high speed stems from sequential or random accessing of only column addresses, with row addresses fixed. Thus, the modes exclude the inherently slow $\overline{\text{RAS}}$-relevant circuit operations, such as word-line activation and sensing, from the access and cycle paths. In traditional DRAM designs, there are three versions of column modes. The page mode can supply random data bits from the data that are latched on to the data lines as a result of row access and the subsequent cell-signal amplification. The random bits can be accessed by changing the column address until the next row access operation. The nibble mode can supply three extra data bits from sequential locations for every row access operation. The static column mode is similar to the page mode, except that it is not necessary to toggle $\overline{\text{CAS}}$ every time the column address changes. The details will be explained below, and extremely high-speed versions developed recently will be discussed in Chap. 6.

Figure 3.8a shows a read-timing diagram of the page mode. Each data bit latched on to a pair of data lines is outputted to the D_{out} pin, under the controls of the column address and other signals. The major column-relevant circuits are precharged every cycle using ATD, which is discussed later. A random column-address cycle time of 40 ns is obtained. However, the cycle time is eventually limited by tight specifications on the address set-up time t_{ASC}, address hold time t_{CAH}, and the $\overline{\text{CAS}}$ precharge time t_{CP}. Although the mode has been popular, users must design the memory board through managing these rigid specifications. A narrow data-valid time t_{OFF} is also a drawback.

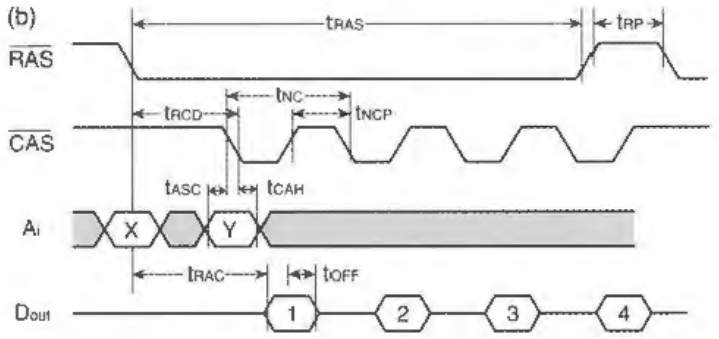
Figure 3.8b shows a read-timing diagram of the nibble mode. A four-bit shift register at the D_{in}/D_{out} buffers manages communications between the memory arrays and the buffers, so that the first data bit can be randomly selected by two address signals, followed by the remaining three data bits, which are serially outputted only by toggling $\overline{\text{CAS}}$. The shift register consists of four data-latch circuits, and a four-bit decoded ring counter/shift register [3.6]. Four data bits from four subarrays are latched in parallel at the latch circuits, and are converted into serial data for synchronous outputting with $\overline{\text{CAS}}$. The shift register can be designed to be fast enough to enable a 35 ns cycle time. Unlike page mode, no timing margin is necessary between $\overline{\text{CAS}}$ and the addresses. This enables ease of use. However, the narrow data-valid time and the fixed number of data-bits are drawbacks in some system designs.

Figure 3.8c shows a read-timing diagram of the static column mode. While $\overline{\text{CAS}}$ stays at a low level, random column selections can be realized. Note

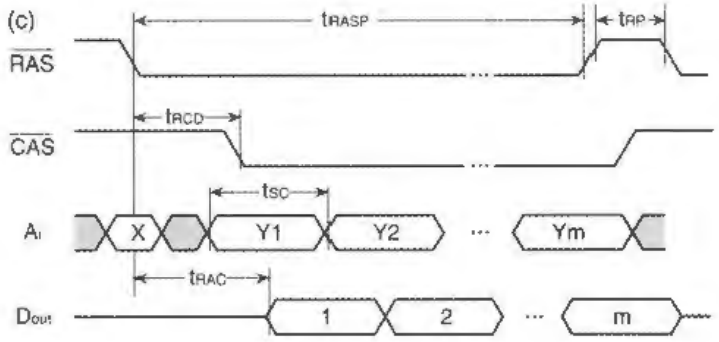


symb.	spec. (ns)	
	min.	max.
t_{RP}	40	—
t_{RASP}	—	—
t_{RCD}	20	100 μs
t_{PC}	40	—
t_{CP}	10	—
t_{ASC}	0	—
t_{CAH}	15	—
t_{RAC}	—	60
t_{OFF}	—	15

t_r, t_f = 5ns



symb.	spec. (ns)	
	min.	max.
t_{RP}	40	—
t_{RAS}	60	10 μs
t_{NC}	35	—
t_{NCP}	10	—
t_{ASC}	0	—
t_{CAH}	15	—
t_{RAC}	—	60
t_{OFF}	—	15



symb.	spec. (ns)	
	min.	max.
t_{RP}	40	—
t_{RASP}	—	100 μs
t_{RCD}	20	45
t_{SC}	40	—
t_{RAC}	—	60

Fig. 3.8. Timing diagrams of high-speed column modes of a 16 Mb DRAM, exemplified by read cycles ($\overline{WE}$: H) [3.4]. (**a**) High speed page mode; (**b**) nibble mode; (**c**) static column mode

that any "don't care" is not included in the address timing, since the address signals cannot be latched by $\overline{CAS}$. The cycle time t_{SC} is determined only by the address transition cycle. ATD detects the transitions so that succeeding column circuits operate. If there is an address skew (i.e. a timing irregularity between address signals), the cycle time becomes slower.

3.3 The Basic Configuration and Operation of the DRAM Chip

3.3.1 Chip Configuration

Figure 3.9 shows the major circuit blocks in a standard DRAM chip [3.4]. The input signals are categorized as N-bit address signals (Ai), two clocks ($\overline{\text{RAS}}$ and $\overline{\text{CAS}}$), a data input D_{in}, and a write enable $\overline{\text{WE}}$. The output signal is only a data output, D_{out}. The Power-supply voltages are V_{DD} and V_{SS} (0 V), but are usually referred to as a single V_{DD} power supply. The input signals and power-supply voltages are inputted to alminum bonding pads, which are usually located at the edge of the chip, through package pins, wiring on to the package, and bonding wires. The output signal is routed opposite to the package pin. There are buffers near the pads corresponding to each input/output signal: these are address buffers (ABs), $\overline{\text{RAS}}$ and $\overline{\text{CAS}}$ buffers, a D_{in} buffer,

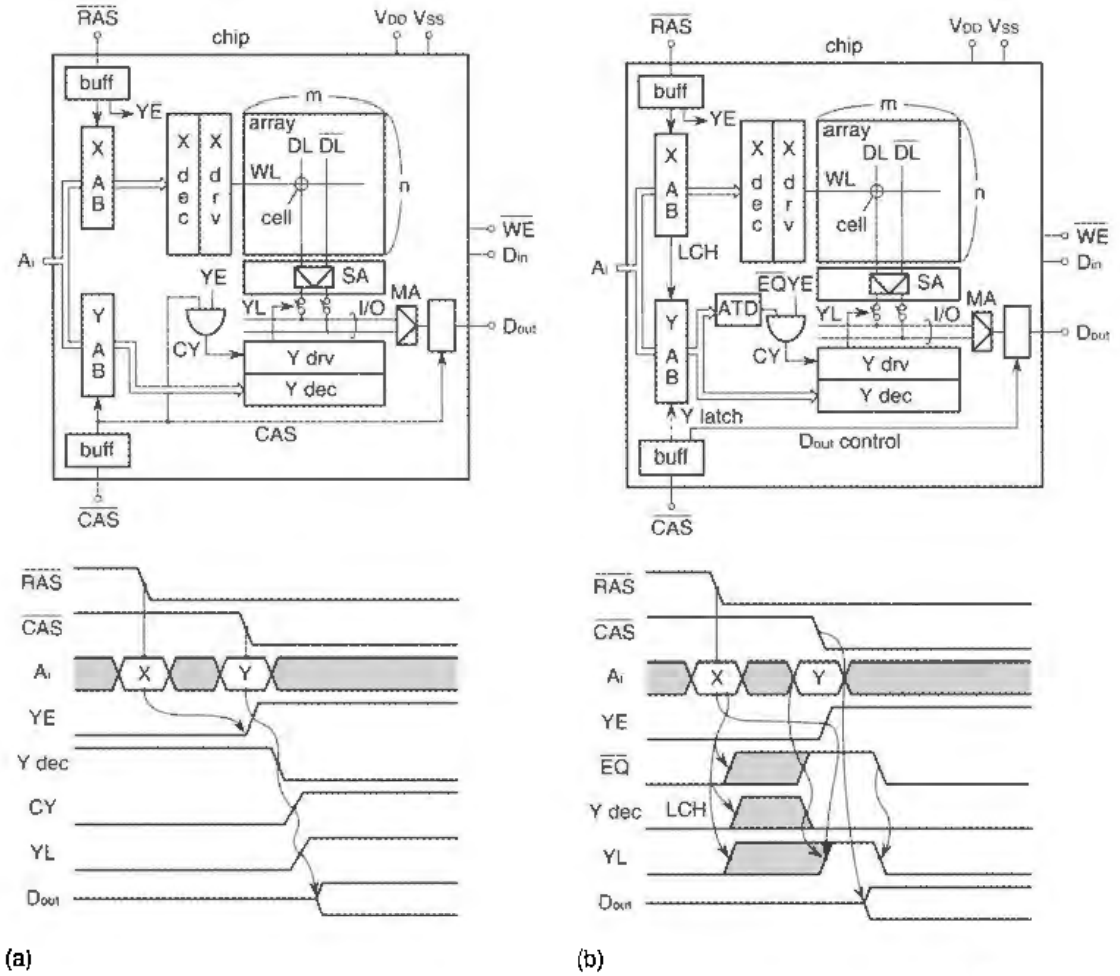


Fig. 3.9. Basic configuration of DRAM chip using address multiplexing [3.4]. (a) NMOS; (b) CMOS

a $\overline{\text{WE}}$ buffer, and a D_{out} buffer. A memory array is composed of n rows (i.e. n word lines, WLs) and m columns (i.e. m pairs of data lines, DLs and $\overline{\text{DL}}$s). Communication between the buffers and the memory array is carried out through peripheral circuits which consist of two kinds of circuits, direct and indirect peripheral circuits. The direct peripheral circuits, which are directly connected to the memory array, comprise row/column decoders, row/column drivers, sense amplifiers, and so on. The indirect peripheral circuits, which are placed between the buffers and the direct peripheral circuits, are composed of many control logic circuits, refresh-related circuits, redundancy circuits, voltage converters, and so on.

3.3.2 Address Multiplexing

Basic Operations. Each address buffer converts an external address signal Ai to internal complementary addresses a_i, and $\overline{a_i}$, and sends them to decoders. The address buffers (ABs) are categorized as row ABs for selecting a row (word line) and column ABs for selecting a column (a pair of data lines). One of the serious problems is the larger package, which causes a lower packing density on a memory board, due to the increased address pin count of the package with increasing memory capacity. For example, a 16 Mb chip needs 24 address pins. However, an address multiplexing scheme [3.5] halves the necessary address-pin count, resulting in 12 pins. The scheme allows one address pin to be used twice, so that a row address and a column address are successively outputted as a result of strobing the input address signal with two successive clocks, $\overline{\text{RAS}}$ (the Row Address Strobe) and $\overline{\text{CAS}}$ (the Column Address Strobe). This scheme has been unique in standard DRAMs since the 16 Kb generation. The operation of address multiplexing (Fig. 3.9) is as follows. Here, each dashed region shows the state of don't care.

At first, a set of external address signals (Ai's) is strobed by $\overline{\text{RAS}}$ and converted into a set of complementary address signals, which are then sent to row decoders. Consequently, one of 2^{12} decoders is selected by 12 sets of complementary addresses, and the succeeding row (word) driver and word line are activated, so that all of the memory cells on the word line are read and then amplified by the corresponding amplifiers. After amplification, a column enable YE is outputted by the $\overline{\text{RAS}}$ buffer. If a column decoder has been selected by another 12 sets of column address signals before YE generation, YE activates a column activation signal CY so that a column select line YL is selected to connect a pair of data lines and a pair of I/O lines. Thus, the amplified signal on the data lines is outputted on the I/O lines, so that a data output D_{out} is available through a main amplifier MA. The write operation is performed by giving the I/O lines a differential V_{DD} voltage corresponding to a data input D_{in} under the control of $\overline{\text{WE}}$. A refresh operation is performed by a read operation for each word line.

The address buffers must be equipped not only with a function that is synchronized with $\overline{\text{RAS}}$ and $\overline{\text{CAS}}$, but also with a latch function, as in the

NMOS DRAM shown in Fig. 3.9a. Hence, as soon as an external row-address signal strobed by $\overline{\text{RAS}}$ is latched into the row-address buffer, the address signal can be changed thereafter to the succeeding external column-address signal, so that the $\overline{\text{RAS}}$–$\overline{\text{CAS}}$ address timing margin is widened. The CMOS address buffers shown in Fig. 3.9b enable high-speed control through the elimination of a synchronous function with $\overline{\text{CAS}}$. Without any regulation by $\overline{\text{CAS}}$, the column address buffers can accept address signals at any time after the row addresses have been latched, and the resultant latch signal LCH has been outputted. Thus, a selected decoder activates one YL, so that the amplified signal on the data lines is transferred to the D_{out} buffer. During these processes, $\overline{\text{CAS}}$ makes the D_{out} buffer ready to accept the processed data in advance. Thus, the need to synchronize the address signals with $\overline{\text{CAS}}$ at the first stage of the address buffers is eliminated, and thus the column access time from column address to D_{out} is shortened without any access penalty. However, the column address must be latched by $\overline{\text{CAS}}$ at the buffers. This is realized by using ATD (Address Transition Detector), again without any access penalty. ATD outputs a pulse $\overline{\text{EQ}}$ every time a set of column addresses is changed (Figs. 3.43 and 7.18). The resultant $\overline{\text{EQ}}$ generates various pulses to control the column-relevant circuits, instead of the $\overline{\text{CAS}}$ pulse. The controls include equalization of I/O lines and activation of the main amplifier, which consumes a dc current, for only a short period of time, necessary for amplification. Here, the address buffers must be composed of CMOS static circuits, so that they can accept address signals at any time at low power.

Differences between Row- and Column-Relevant Circuits. Chip operation using address multiplexing is governed by both row circuits, which are relevant to the path of the $\overline{\text{RAS}}$ input to cell signal amplification, and column circuits, which are relevant to the output path of the amplified signal to the I/O lines to data output (D_{out}). However, there is an essential difference between both row and column circuits; row circuits must be dynamic circuits, while column circuits can be static circuits, as follows. During time t_{RP} in Fig. 3.6, every word-line activation is followed by a data-line precharge operation. However, the data-line precharging must be started after the word line has been completely turned off, to 0 V. If it starts while the word line is still at a high or intermediate level, a voltage that is developing toward the final precharge voltage (e.g. $V_{DD}/2$) on each data line may be written to each memory cell. This implies loss of the previous write or rewrite data. Also, the next word-line activation must be started after each pair of data lines has been completely equalized, as a result of their precharging. Otherwise, the insufficient equalization is a source of noise for the succeeding read operation. Thus, a succeeding word-line activation must be accomplished after the time period t_{RP}. This timing sequence imposes precharge operations on the row-relevant circuits. Consequently, they must be constructed using dynamic circuits, or circuits that are suitable for dynamic operations. On the other hand, even successive column-line (YL) activations without a precharge operation

could be performed without any detrimental effect on the memory cells, as in SRAM. A precharge operation of the I/O lines is not essential for basic memory operation, but only for shortening of the column cycle time.

3.4 Fundamental Chip Technologies

In this section, first, design issues for a larger memory-capacity chip are clarified [3.4] in relation to the theory of scaled-down devices discussed in Chap. 2. Next, the solutions are described [3.4] in detail, in terms of the S/N ratio, power dissipation, speed, and chip area.

3.4.1 A Larger Memory Capacity and Scaled-Down Devices

Figure 3.10 shows a schematic internal chip configuration [3.4]. A memory array consists of n' word lines (WL_0,..., $WL_{n'-1}$) and m' pairs of data lines (DL_0, $\overline{DL_0}$,..., $DL_{m'-1}$, $\overline{DL_{m'-1}}$). X and Y denote the row and column decoders and drivers, while SA and MA are a sense amplifier and a main amplifier, respectively. The aspect ratio of the length to the width of standard DRAM chips has usually been about 2, and results from the package size, the pin assignment of package, and the aspect ratio of the memory-cell size.

Figure 3.11 shows trends in memory-cell area and cell-signal charge [3.3]. The cell area has been reduced at a rate of 38% per generation, due to advancements in high-density technology. The reduction in the area is due to a size reduction to about 62%. However, despite such a rapid reduction in cell size the memory array has become larger by about 1.52 ($= 0.38 \times 4$) with each successive generation, causing the roughly 1.5 times increased chip area that has been traditionally established, as shown in Fig. 3.2. The ideal scaling rule discussed in Chap. 2 is for a fixed memory capacity, in which all dimensions, including the chip dimensions, can be reduced. For the quadrupled memory capacity, however, the scaling rule must be modified. In addition to the chip dimensions, other parameters – such as a thickness of wiring metals – differ from the ideal case, causing ever-increasing capacitances, resistances, and thus wiring delays. In general, the scaling ratio for the thickness of the wiring metal must be larger than that for the areal dimensions. Otherwise, a thinner interlevel dielectric would reduce the breakdown voltage, and a thinner metal film would degrade reliability giving an increased current density and stress-induced voiding.

Here, let us evaluate how these deviations from ideal scaling influence the performance of the chip. In Fig. 3.12, our major concerns are with the global metal wiring (interconnection) that runs from the edge of the chip to a peripheral circuit block, B_2, at another edge, and with the word lines and data lines in a memory array, B_3. Under ideal scaling, the performance of the chip is improved, since the dimensions of the chip and thus the length

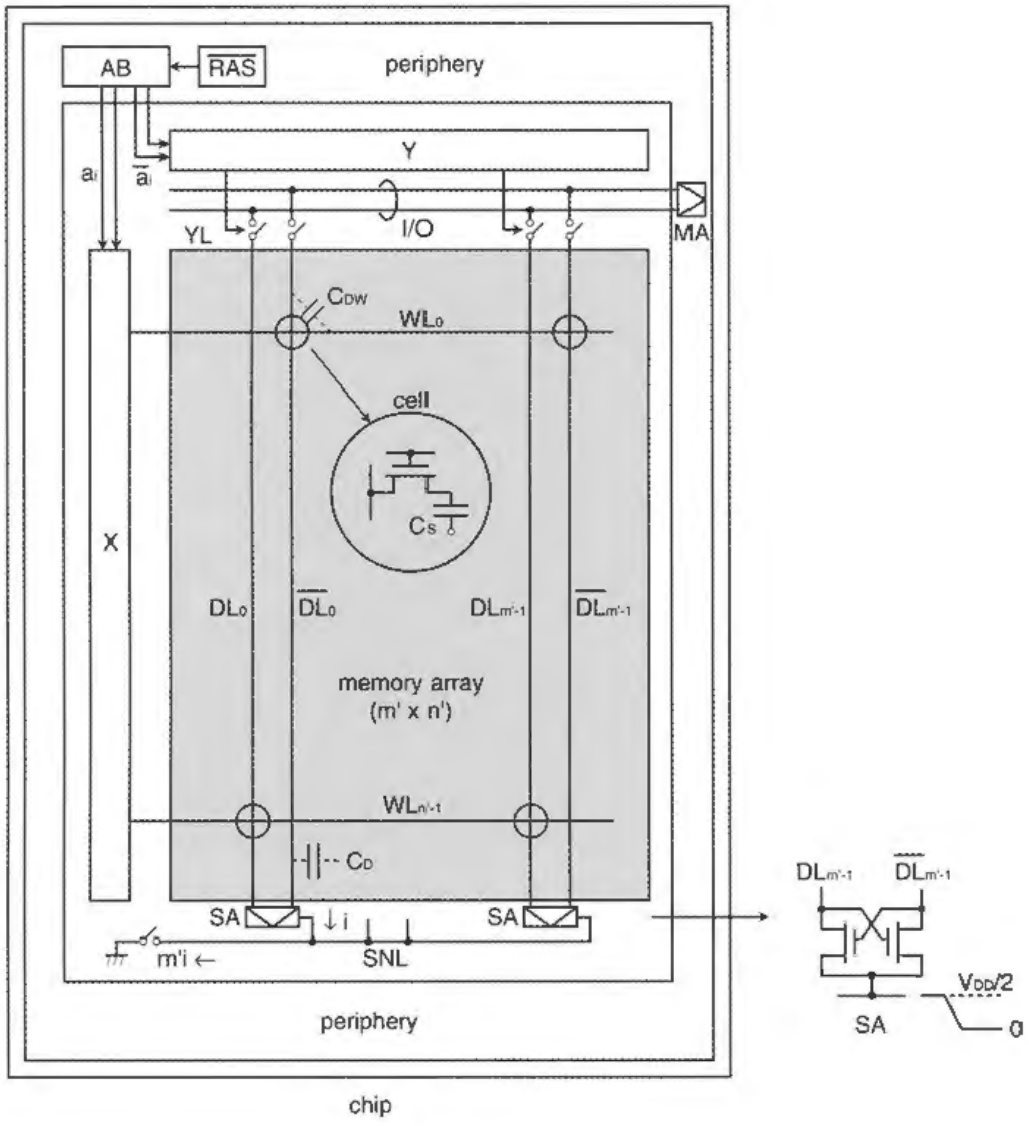


Fig. 3.10. The structure of a DRAM chip [3.4]

of the local metal wiring (l_{L}) in a sub-block $\mathrm{B_I}$ and the global metal wiring (l_{G}) are reduced by the same scaling factor k^{-1} ($k > 1$). On the contrary, however, the dimensions of the chip have actually increased by a factor of k_{a} ($k_{\mathrm{a}} > 1$), and thus the length of the global wiring has increased to $k_{\mathrm{a}} l_{\mathrm{G}}$. If the scaling factors of the areal dimensions of the wiring, the thickness of the interlevel dielectric, and the thickness of the wiring are assumed to be k, k_{d}, and k_{c}, respectively, the capacitance, resistance, and delay of the global wiring increase at rates of $k^{-1} k_{\mathrm{a}} k_{\mathrm{d}}$, $k_{\mathrm{a}} k k_{\mathrm{c}}$ and $k_{\mathrm{a}}^2 k_{\mathrm{d}} k_{\mathrm{c}}$ with each successive generation. If we assume [3.4] $k^{-1} = 0.62$, $k_{\mathrm{a}} = 1.22$, $k_{\mathrm{d}}^{-1} = 0.735$, and $k_{\mathrm{c}}^{-1} = 0.81$, these rates are as large as 1.03, 2.43 and 2.50, respectively. Even in this case, the performance of the local wiring is improved, because the

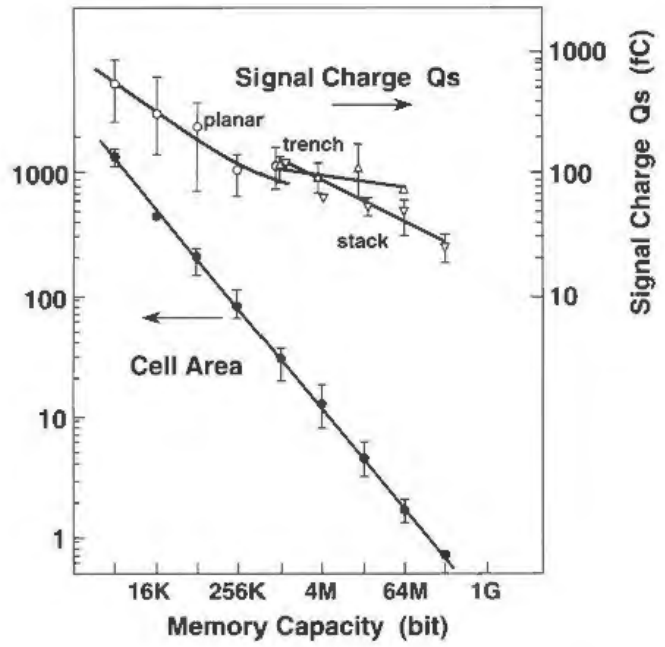


Fig. 3.11. Trends in memory cell area and signal charge [3.4]

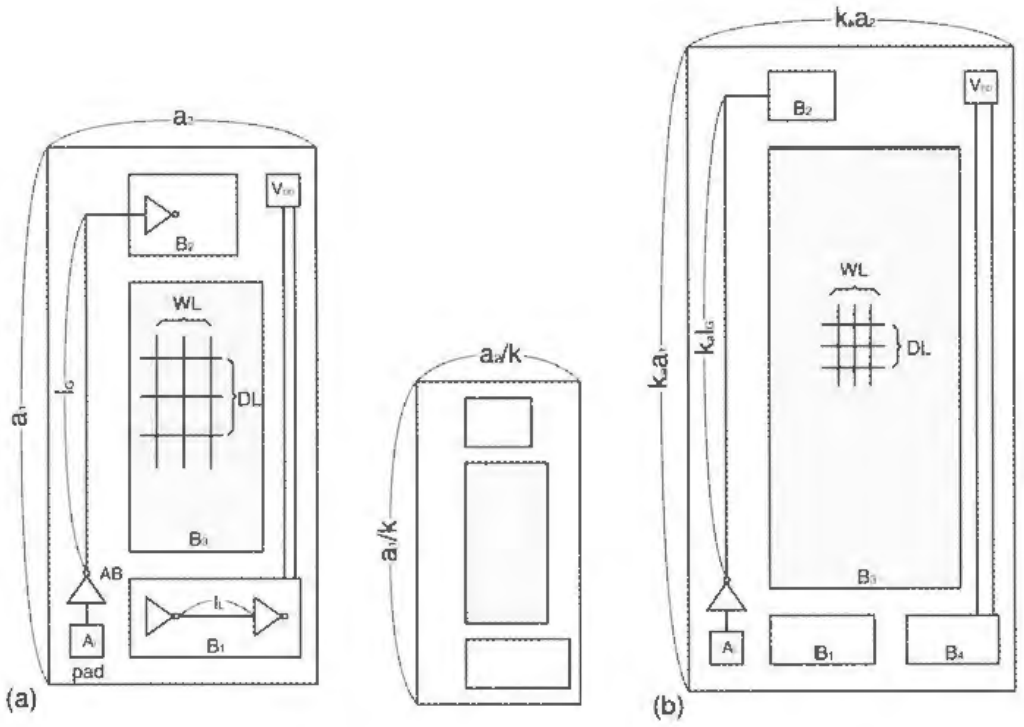


Fig. 3.12. Chip scaling [3.4]. (a) Ideal ($k > 1$); (b) actual ($k_a > 1; k_a \approx 2/k$)

rates are $k^{-2}k_{\mathrm{d}}(=0.52)$, $k_{\mathrm{c}}(=1.23)$ and $k^{-2}k_{\mathrm{d}}k_{\mathrm{c}}(=0.64)$. Therefore, the delays caused by the global wiring have a strong impact on the speed of the peripheral circuits.

Word lines and data lines, which are a kind of global wiring, and which are made of resistive materials, increase the delays further. As a result of making the higher density of a memory cell the top priority, word lines and data lines are made of relatively highly resistive materials such as poly-Si and polycide, which are suitable for a self-aligned and fine patterning process, and even a diffused layer. The resultant large delays are enhanced at the above-calculated rate of 2.5. Moreover, the charging power of data lines increases rapidly since the number of data lines charged up at any one time is doubled in each successive generation, even if the increase in the data-line capacitance C_{D} is assumed to be as small as a factor of 1.03. Furthermore, even C_{D} reduces the signal voltage of the memory cell, because the cell capacitance C_{S} usually decreases. Increases in voltage drops at power-supply lines and the occurrence of crosstalk, when the chip area increases in smaller devices, also make the voltage margin of the chip narrower. Therefore, the S/N ratio, power dissipation, and delay issues become more prominent.

3.4.2 High S/N Ratio Circuits

To realize stable memory-cell operation, the signal-voltage characteristics and noise-generation mechanisms during cell activation and signal amplification must be investigated. A summary of the S/N ratio issue is given here. The full details are explained in Chap. 4.

The Signal Voltage of a Memory Cell. The signal voltage v_{S}, which is expressed as

$$v_{\mathrm{S}} = Q_{\mathrm{S}}/(C_{\mathrm{D}} + C_{\mathrm{S}}),\ Q_{\mathrm{S}} = C_{\mathrm{S}}V_{\mathrm{DD}}/2\,, \tag{3.2}$$

is maintained by increasing the signal charge Q_{S} as much as possible through three-dimensional capacitors, such as stacked and trench capacitors, a half-V_{DD} capacitor plate, high-permittivity capacitor insulators, and word boots-trapping. Multidivision of the data line and a shared sense amplifier have also contributed to a larger v_{S} with reducing values of C_{D}.

Noise at Cell-Read Timing. Various forms of noise, coupled to a pair of data lines from non-selected word lines and adjacent data lines during cell activation and signal amplification, are canceled by the folded data-line arrangement combined with a dummy cell and/or a half-V_{DD} data-line precharge. The ever-increasing offset voltage of the sense amplifier, which is a difference in V_{T} values between cross-coupled transistors, can be reduced by enlargement of their feature sizes.

Noise Coupled from an Array to Peripheral Circuits. The simultaneous charging or discharging of many data lines with a large swing of voltage V_{DD} is a noise source for peripheral circuits. The voltage swings couple noise to non-selected word lines and the relevant power lines via the data-line to word-line capacitances. They also couple substrate noise to peripheral circuits via the data line to substrate diffused capacitances. The folded data-line arrangement, combined with a half-V_{DD} data-line precharge, also cancels noise, achieving a quiet memory array.

Noise in Peripheral Circuits and D_{in}/D_{out} Buffers. Noise increases not only with an increasing chip area, but also with the ever-higher speed. The multi-data-bit I/O configuration, which is essential to realize a larger memory capacity, is also a source of noise at the D_{out} buffers. For example, simultaneous voltage bounces of 3–5 V at 16 buffers, each of which has a load capacitance as great as 100 pF, surely generate noise on the powerlines of buffers. Careful attention to layout, with multilevel metal layers, and adjustment of the rise and fall times of the D_{out} waveform, suppresses the noise. The isolation and shielding of noise sources from noise-sensitive circuit blocks, such as main amplifiers, I/O lines, and comparators in voltage down-converters, are also well-known noise-suppression techniques.

3.4.3 Low Power Circuits

In this section, the sources of power dissipation in a DRAM chip are discussed first. Next, it is made clear that the data line is a major power source. Then, various power reduction circuits are discussed, with an emphasis on data-line power reduction.

Sources of Power Dissipation. The power dissipation P of a DRAM chip that operates with a cycle time t_{RC} and a power-supply voltage V_{DD} can be expressed as follows:

$$P \simeq \Sigma C_j(\Delta V_j/\Delta t)V_{DD} + (I_{DC} + I_{PH})V_{DD} \tag{3.3}$$

$$\simeq \Delta Q_T \cdot V_{DD}/t_{RC} + (I_{DC} + I_{PH})V_{DD} \tag{3.4}$$

$$\simeq (C_{DT} \cdot \Delta V_D + C_{PT} \cdot \Delta V_P)V_{DD}/t_{RC} + (I_{DC} + I_{PH})V_{DD} \tag{3.5}$$

$$\simeq (\Delta Q_{DT} + \Delta Q_{PT})V_{DD}/t_{RC} + (I_{DC} + I_{PH})V_{DD} \tag{3.6}$$

where Cj and ΔV_j are the capacitance and voltage swing at node j; ΔQ_T, ΔQ_{DT}, and ΔQ_{PT} are the total charges of the chip, data lines, and peripheral circuits, which are charged up from the V_{DD} power supply; the I_{DC} are the major dc currents such as the ratio current at common I/O circuits and the constant-current sources of main amplifiers; the I_{PH} are the quasi dc currents from circuits, that are always in operation, such as on-chip voltage converters and relevant refresh circuits; and C_{DT} and ΔV_D are the total charge of data lines that are simultaneously charged up, and the data-line voltage swing. Hence, the average current I_T in the chip is as follows:

$$I_T = (\Delta Q_{DT} + \Delta Q_{PT})/t_{RC} + I_{DC} + I_{PH} \ . \tag{3.7}$$

The array current I_A and peripheral current I_P, when the chip is operating at the minimum cycle time t_{RCmin}, are thus given by

$$I_A = \Delta Q_{DT}/t_{RCmin} = C_{DT} \cdot \Delta V_P/t_{RCmin} = m'C_D \cdot \Delta V_D/t_{RCmin} \ , \tag{3.8}$$

$$I_P = \Delta Q_{PT}/t_{RCmin} = C_{PT} \cdot \Delta V_P/t_{RCmin} \ ; \tag{3.9}$$

$$\therefore \ I_T = (I_A + I_P)(t_{RCmin}/t_{RC}) + I_{DC} + I_{PH} \ . \tag{3.10}$$

Note that only one of a pair of data lines contributes to power dissipation, because a charging current always flows from the V_{DD} power supply to only one data line during one cycle. In V_{DD}-precharge, one data line is charged up to V_{DD} during precharging, while the other remains at V_{DD}. In a half-V_{DD} precharge, a charging current flows to raise one data line from a half-V_{DD} to V_{DD} at amplification. Currents that do not flow from the V_{DD} power supply, such as discharging and equalizing currents, do not contribute to power dissipation. Hence, ΔV_D is equal to V_{DD} and $V_{DD}/2$ for V_{DD} and half-V_{DD} precharging, respectively. The cycle time t_{RC} has an acceptable slowest value, which is t_{REFmax}/n' for the array shown in Fig. 3.10. Thus, the total current I_T is maximized to be I_{Tmax} at t_{RCmin} and minimized to be I_{Tmin} at t_{REFmax}/n', as shown in Fig. 3.13. The reduction of both I_{Tmax} and I_{Tmin} is important, since reduction of I_{Tmax} reduces T_j, while reduction of I_{Tmin} extends the battery back-up time. For a non-divided array (Fig. 3.10) I_A is overwhelmingly larger than the other components, as explained later, and the expression for I_T is therefore simplified to

$$I_T \simeq m'C_D \cdot \Delta V_D/t_{Rcmin} \ . \tag{3.11}$$

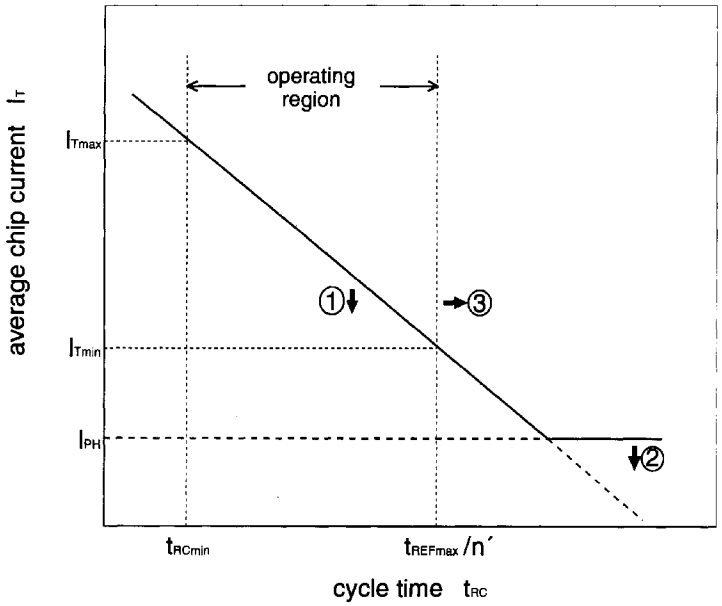


Fig. 3.13. Current versus cycle time [3.4]

If we assume a 20% reduction in ΔV_D, an improvement in t_{RCmin} of 10% and a 3% increase in C_D (i.e. a factor of 1.03, as previously discussed) in each successive generation, I_T increases by a factor of over two every generation.

Reductions in V_{DD}, ΔQ_T and I_{DC} are the keys to power reduction, as given by (3.4). Thus, in addition to reductions of the standard power supply V_{DD} from 12 V to 5 V, and then to 3.3 V, partial activation of a multidivided data line, an increase in t_{REFmax}, low-voltage operations through on-chip voltage down-converters and low-signal voltage transmission, CMOS circuits, and power-down operations using ATD have all been effective, as briefly discussed below.

Partial Activation of a Multidivided Data Line. If each data line is divided into q subdata lines and, for example, only one subdata line is activated, the array current I_A is reduced to $1/q$, with C_D/q. If each subdata line is divided into two (the shared I/O is explained later), I_A is again halved, with a halved value of C_D.

An Increase in t_{REFmax}. The reduction of m' in (3.11) cannot be achieved without increasing the number of refresh cycles n and the maximum refresh time of the cell, t_{REFmax}. Here, if we introduce a logical memory array that uses n (i.e. an $m \times n$ matrix) instead of a physical memory array with an $m' \times n'$ matrix for the same memory capacity M, the complicated relationship between m' or n and t_{REFmax} becomes easier to understand. As a result of q divisions of each data line, a physical memory array can be composed of q physical subarrays, each of which comprises an $n_0 \times m'$ matrix, and operates independently, as shown in Fig. 3.14a. Here, n_0 is the number of memory cells connected to one subdata line, which is determined by the S/N issue and usually ranges from 256 to 1024. If the number of subarrays that operate simultaneously is j, we can consider that one word line logically activates jm' memory cells, enabling a logical memory array with a matrix of $m = jm'$ and $n = qn_0/j$. On the other hand, the refresh busy rate, γ, which expresses the percentage of the time which is not accessible from outside the chip, is given by

$$\begin{aligned}\gamma &= t_{RCmin}/(t_{REFmax}/n) \\ &= (M/m)(t_{RCmin}/t_{REFmax}) \\ &= (q/j)n_0(t_{RCmin}/t_{REFmax})\,. \end{aligned} \tag{3.12}$$

Figure 3.15 shows two extreme cases of logical organization. Figure 3.15a is for simultaneous operation of all subarrays, with $n = n_0$, $m = qm'$, $C_{DT} = qm'C_D$, and $\gamma = n_0 t_{RCmin}/t_{REFmax}$. Figure 3.15b is for only one subarray operation, with $n = qn_0$, $m = m'$, $C_{DT} = m'C_D$, and $\gamma = qn_0 t_{RCmin}/t_{REFmax}$. Obviously, the former maximizes power dissipation with a maximized C_{DT} but minimizes the refresh busy rate, which is contrary to the operation of the latter. In practice, both n and t_{REFmax} are specified in the catalogs (Table 3.1), through compromising the power with the cell leakage current

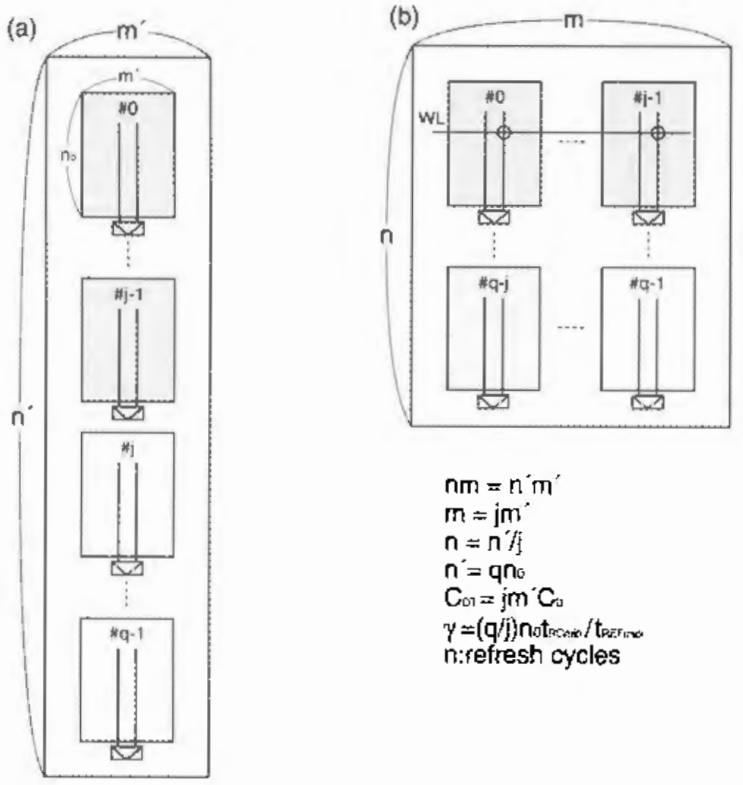


Fig. 3.14. The physical memory array (a) and the logical memory array (b) [3.4]

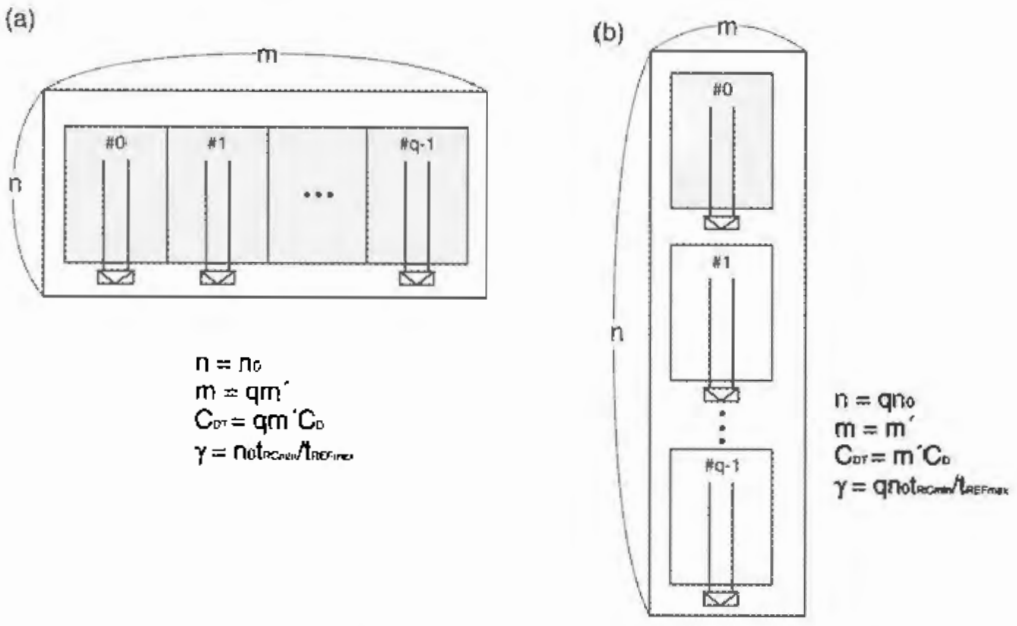


Fig. 3.15. The configuration of the logical memory array, depending on the number of subarrays that operate simultaneously [3.4]. (a) simultaneous operation of all subarrays; (b) operation of only one subarray

for almost-constant γ and t_{RCmin}. For fixed M, it is necessary to maintain mt_{REFmax} to keep γ constant, assuming a fixed t_{RCmin}. This implies a reduced m accompanied by an increased t_{REFmax}. Moreover, to quadruple M, mt_{REFmax} must be quadrupled. This has been achieved by almost doubling both m and t_{REFmax}, since quadrupling m increases the array power greatly while quadrupling t_{REFmax} requires an unacceptably low cell-leakage current. As a result, n $(= M/m)$ has gradually increased with each successive generation, with an increased t_{REFmax} for commercial DRAMs (Fig. 3.16). The alternative choices for n and m have eventually been rejected by the adoption of this compromise.

Low-Voltage Operations. In addition to the reduction of V_{DD}, a half-V_{DD} data-line precharge, using a CMOS sense amplifier, and on-chip voltage down-converters that enables the use of small devices have been effective in reducing the power, by halving ΔV_{D} and reducing ΔV_{D} and ΔV_{P} by the ratio of $V_{\mathrm{DL}}/V_{\mathrm{DD}}$ given in (3.5).

The Low-Power CMOS Decoder. To reduce the ever-increasing power of the decoder, CMOS circuits have been used instead of NMOS circuits since the 1 Mb generation: this will be discussed in detail below. The NMOS dynamic NOR decoder allows the output nodes of all the decoders except one selected decoder to be discharged, while leaving the output node of the one selected decoder at the precharge voltage of V_{DD}. On the contrary, the CMOS NAND decoder allows the output node of the one selected decoder to be discharged, leaving the remaining output nodes at V_{DD}. Thus, the discharging power of the CMOS decoder is extremely low, almost indepen-

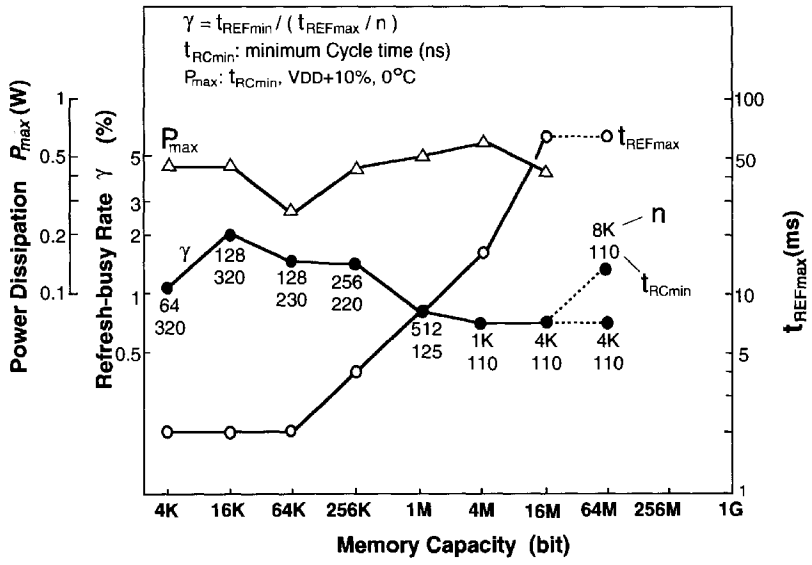


Fig. 3.16. Trends in power dissipation (P_{max}), the refresh-busy rate(γ), and the maximum refresh time(t_{REFmax}) [3.3, 3.4]

dently of memory capacity, while that of the NMOS decoder is increased with increasing memory capacity, despite a dynamic circuit configuration.

The Reduction of dc Currents. The large dc currents, that were consumed in some peripheral and sensing circuits, have been reduced by the use of dynamic circuits and CMOS-circuits.

Low-Power Peripheral Circuits. The NMOS dynamic circuits discussed in Chap 2 reduced the power of drivers and decoders in the NMOS era. The succeeding CMOS circuits have further reduced their power with a wider voltage margin. Without CMOS static address buffers, a low-power, high-speed page mode could never have been realized.

Low-Power Sensing Circuits. In the 4 Kb generation, a cross-coupled NMOS static circuit was used for each sense amplifier. In the 16 Kb generation, however, to eliminate a dc current during amplification this was replaced by an NMOS dynamic circuit, combined with V_{DD} data-line precharge. In the 1 Mb generation, to halve the data-line charging power, this circuit was replaced again, by a CMOS cross-coupled circuit combined with half-V_{DD} data-line precharge. The reduction of dc currents for I/O-relevant circuits has also been indispensable. However, a small signal voltage needed for a fast signal transmission on a heavily capacitive I/O line inevitably develops a dc current that flows from the loads on the I/O lines to a CMOS sense amplifier on the selected data line. In addition, as in the current-mirror amplifier discussed in Chap. 7, a dc current source is needed for a main amplifier, to detect the small signal voltage and convert to a large logic voltage swing. ATD techniques cut the dc currents during periods when they are unnecessary.

A Comparison of Power between the NMOS DRAM and the CMOS DRAM. Figure 3.17 shows a comparison of the power [3.7] between a 1 Mb NMOS DRAM and a 1 Mb CMOS DRAM that are using the same 1.3 μm process technology. Obviously, the use of CMOS circuits almost halves the power of the NMOS DRAM. Figure 3.18 shows how V_{DD} and the number of refresh cycles n affect the power of a CMOS 64 Mb DRAM [3.4], in which a multidivided data line, a half-V_{DD} data-line precharge, and a shared sense

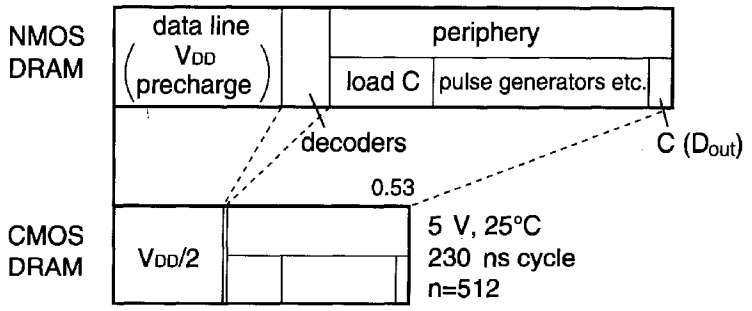


Fig. 3.17. Power comparison between a 1 Mb NMOS DRAM and a 1 Mb CMOS DRAM [3.7]

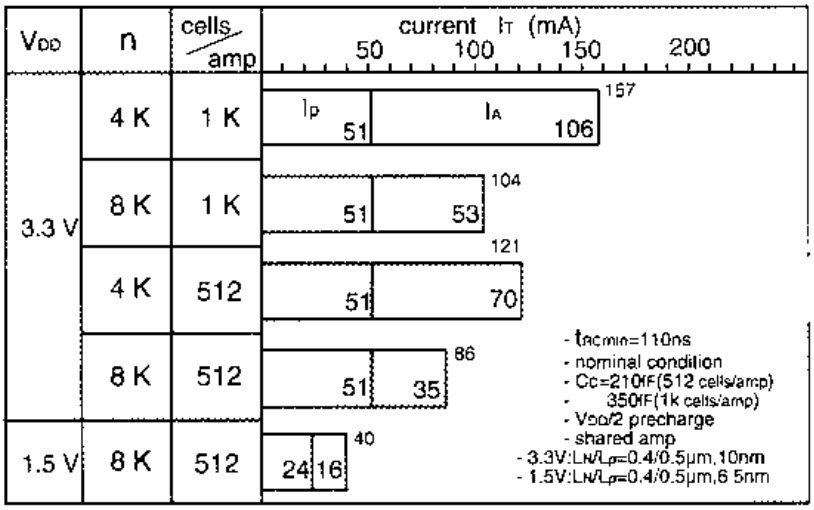


Fig. 3.18. The current components of 64 Mb DRAMs [3.4]. I_{P}, the peripheral current; I_{A}, the array current, including the dc current of the sense amplifiers

amplifier and I/O are used. The power dissipation decreases with lowering V_{DD}, increasing n and decreasing the number of memory cells n_0 connected to a sense amplifier (i.e. decreasing C_{D}). Here, let us cite an example of $V_{\mathrm{DD}} = 3.3\,\mathrm{V}$, $n = 8\,\mathrm{K}$, and $n_0 = 512$. Obviously, the number of data lines or sense amplifiers that operate simultaneously is 8 K, because $m = 8\,\mathrm{K}$. Thus, the number of data-line divisions is 16 $(= n/n_0)$ for a logical memory array. A shared I/O scheme again halves the length of the resulting subdata line. Without any data-line division I_{A} would increase from 35 mA to approximately 1120 mA $(= 35\,\mathrm{mA} \times 32)$ and the total chip current would thus increase from 86 mA to as much as 1171 mA, making the data-line charging power dominate with an occupancy of 96%. Thus, a multidivided data-line scheme drastically reduces the chip power.

A reduction in the current I_{Tmin} in the battery back-up (or stand-by) mode is realized by reducing the average current ① and the quasi-static current I_{PH} ② and increasing t_{REFmax}/n' ③, as shown in Fig. 3.13. The above-described low-power techniques, and switching to ultra-low-power circuits dedicated to the stand-by mode, are effective ways of reducing the average current and I_{PH}, respectively. An on-chip timer automatically extends t_{REFmax} according to the junction temperature, as will be explained later.

3.4.4 High-Speed Circuits

The access time of the chip, which is the sum of the delay times of major circuit blocks on the critical path, is given by

$$t_{\mathrm{RAC}} \simeq \tau_{\mathrm{RA}} + \tau_{\mathrm{DEC}} + \tau_{\mathrm{WL}} + \tau_{\mathrm{S}} + \tau_{\mathrm{I/O}} + \tau_{\mathrm{OUT}} \tag{3.13}$$

where τ_{RA} is the delay time from $\overline{RAS}$ low to the validating complementary addresses (a_i and $\overline{a_i}$); τ_{DEC} is the delay time of the decoder block (X); τ_{WL} is the delay time from the decoder output to activation of a word line up to a sufficient level; τ_S is the delay time from word-line activation to activation of a column select line (YL), which is almost equal to the time needed for cell-signal amplification; $\tau_{I/O}$ is the delay time of the amplified signal from the data line to the main amplifier (MA) input; and τ_{OUT} is the delay time from the main amplifier to the data output (D_{OUT}). All of the delay components except τ_{DEC} and τ_{OUT} increase as the memory capacity is quadrupled, because the chip area is increased despite scaled-down devices, as mentioned previously. τ_{RA} increases because all sets of complementary addresses run along an increased length of chip. τ_W, τ_S, and $\tau_{I/O}$ also increase with increased word-line, data-line, and I/O-line delays. τ_S increases further, since the driving speed of the sense amplifiers is more degraded. This is because as soon as m' pairs of NMOS sense amplifiers (Fig. 3.10) are driven by activation of the drive line SNL from $V_{DD}/2$ to 0 V, the resultant total discharging current $m'i$ raises the SNL voltage due to the resistance of the SNL-line, so that the sensing speed suddenly slows down with a lower gate to source (i.e. data-line to SNL) voltage of the sense amplifier, as will be explained later.

To reduce these delay components, the line delay and circuit speed of each circuit block must be reduced. Typical examples are shown below.

Reduction of the Line Delay. Figure 3.19 summarizes various fast driving schemes [3.4] for a resistive line shown in Fig. 3.19a. They are all based on multidivision of the line. Figure 3.19b shows center driving of a two-divided line, which enables a quarter of the original delay, with a halved resistance and capacitance. Figure 3.19c shows a metal-strapped line which has been widely used for DRAM word lines. Figure 3.19d shows a combination of (b) and (c) which has been indispensable for multidivided word-line and data-line structures, which will be discussed later. Figure 3.19e shows a buffer (a so-called

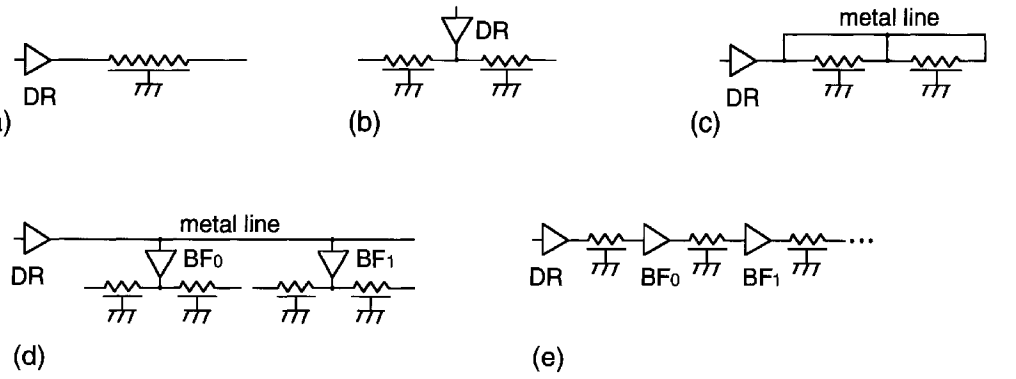


Fig. 3.19. Delay reductions for a resistive line [3.4]. DR, driver; BF, buffer. (**a**) Original resistive line; (**b**) center driving; (**c**) metal strapping; (**d**) hybrid driving; (**e**) insertion of repeaters (BFs)

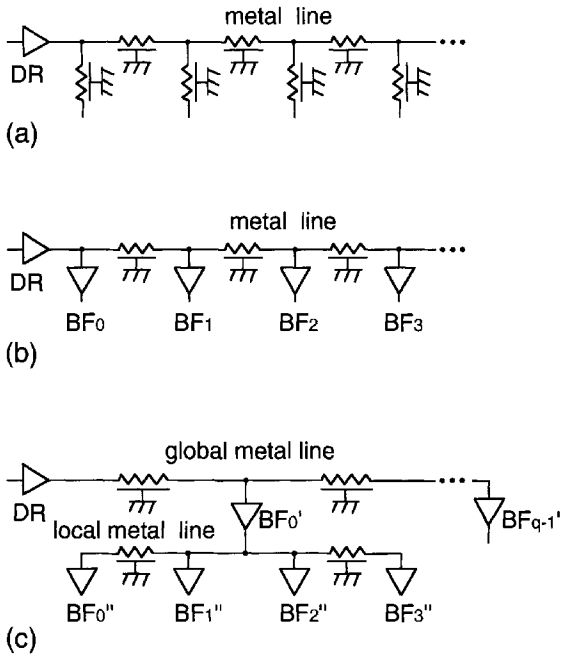


Fig. 3.20. Delay reduction of a long metal line with heavily capacitive branch lines [3.4]. (**a**) Heavily capacitive line with branches; (**b**) insertion of buffers; (**c**) hierarchical metal line with buffers

"repeater") inserted at each division of the line. In this scheme, the total delay is reduced from $l_{\mathrm{G}}^2 r_{\mathrm{int}} c_{\mathrm{int}}$ to $l_{\mathrm{G}}^2 r_{\mathrm{int}} c_{\mathrm{int}}/q$, if the buffer is fast enough, where l_{G}, r_{int}, c_{int}, and q are the total length of the line, the resistance and capacitance per unit length, and the number of divisions, respectively. This scheme has been proposed for the word lines of an experimental chip [3.8].

Figure 3.20 shows delay-reduction schemes for a long, low-resistance metal line (such as a global line), which has many heavily capacitive branch lines, as shown in Fig. 3.20a. In Fig. 3.20b the total delay is reduced by cutting the branch lines by buffers. The resultant delay τ is given by

$$\tau \simeq l_{\mathrm{G}} r_{\mathrm{int}} \cdot n_{\mathrm{C}} c_{\mathrm{in}} \ , \qquad \text{if} \quad l_{\mathrm{G}} c_{\mathrm{int}} \ll n_{\mathrm{C}} c_{\mathrm{in}} \tag{3.14}$$

where n_{C} and c_{in} are the number of buffers and the input capacitance of a buffer. Figure 3.20c shows a combination of a global metal line and a number of local metal lines. The delay time from the driver DR to a buffer on a local line is given as follows, assuming that the buffers are sufficiently fast. The delay of the global line τ_1 is $l_{\mathrm{G}} r_{\mathrm{int}} \cdot q c_{\mathrm{in}}$, and that of the local line τ_2 is $(l_{\mathrm{G}}/q) r_{\mathrm{int}} \cdot (n_{\mathrm{C}}/q) c_{\mathrm{in}} \cdot (1/4)$ because of center driving. Hence, the total delay τ_{12} of the far end of the local line is

$$\tau_{12} = l_{\mathrm{G}} r_{\mathrm{int}} c_{\mathrm{in}} \left[q + (1/4) n_{\mathrm{C}}/q^2 \right] \ ; \tag{3.15}$$

$$\therefore \ \tau_{12}/\tau = (q/n_{\mathrm{C}}) + (1/4q^2) \ . \tag{3.16}$$

Note that $\tau_{12}/\tau \simeq 0.015$ for $n_{\mathrm{C}} = 1024$ and $q = 4$, giving a large reduction in the delay. This stems from the combined effect of a reduction in the loading

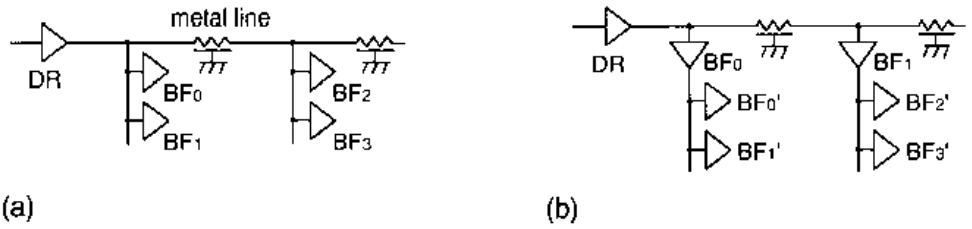


Fig. 3.21. Delay reduction of distributed connections of branched buffers [3.4]. (**a**) Distributed and branched buffers; (**b**) insertion of buffers

for the global line due to the insertion of the buffers and a reduction in the delay for the local line due to the divisions. The scheme has been widely used in circuit blocks with large values of l_G and n_C, such as decoders, as explained later. The example shown in Fig. 3.19d is regarded as a variation of this scheme.

Figure 3.21 shows groups of buffers branched and distributed on a line. The insertion of more buffers, as in Fig. 3.21b, reduces the total delay time in the same manner.

Line delays also depend on how bonding pads and peripheral circuits are located in a chip. To reduce the delays, each bonding pad and its relevant I/O buffer should be laid out as closely as possible, to shorten the I/O signal line. Furthermore, circuits on the critical path should be planned so as to localize the relevant signal lines, leaving non-critical signal lines to be global lines. There are two kinds of location that meet the requirements, as shown in Fig. 3.22. In Fig. 3.22a they are localized along two edges of the chip. However, the longest signal line might be $a_1 + a_2$, where a_1 and a_2 are the length and width of the chip, respectively. In Fig. 3.22b they are almost localized in one cross-shaped area, which allows the longest signal line to be $(a_1 + a_2)/2$.

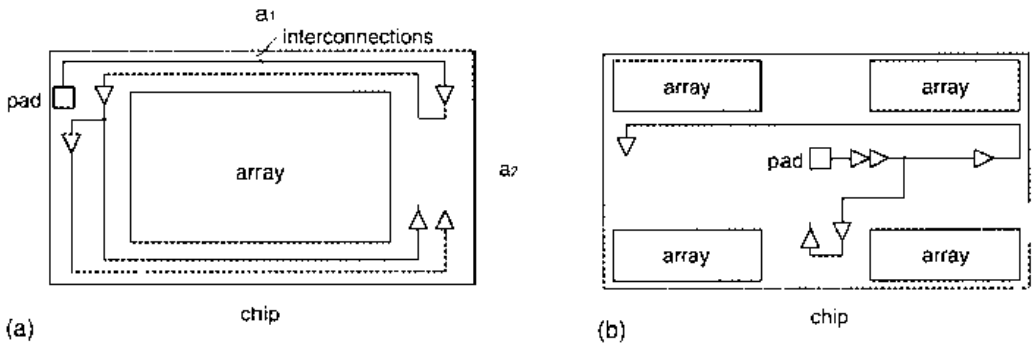


Fig. 3.22. Locations of bonding pads and peripheral circuit blocks [3.4]. (**a**) Location at two edges of chip, maximum signal length $\approx a_1 + a_2$; (**b**) location at a cross-shaped area, maximum signal line length $\approx (a_1 + a_2)/2$

Distributed Driving of Sense Amplifiers. The accumulation of a small current in a line would cause an unexpectedly slow speed, with a resulting voltage drop across the line even if it is a low-resistance metal line. Such is the case for the common drive line SNL of sense amplifiers, as shown in Fig. 3.23a. The raised source voltage of the sense-amplifier MOSFET in the far-end portion degrades the sensing speed. Multidivision of SNL combined with parallel driving of the resultant sub-SNLs, as shown in Fig. 3.23b, improves the speed.

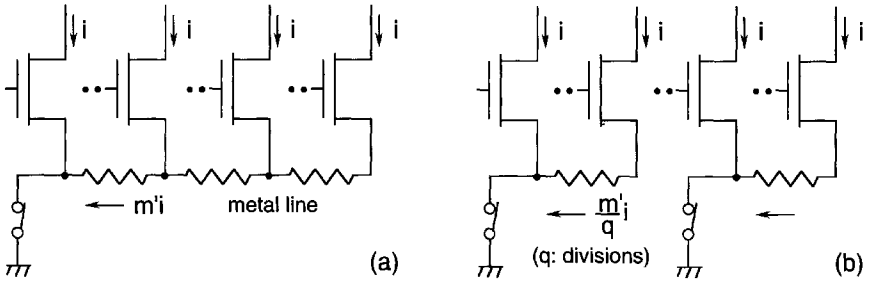


Fig. 3.23. Reduction of the voltage drop on an SA drive line by current distribution [3.4]. (**a**) Current accumulation; (**b**) current distribution

Low-Voltage Operation. The operational speed of a driver is proportional to the voltage swing at the load for a fixed load capacitance and driver current. Hence, to improve the speed when the load capacitance is heavy and/or the driver current is small, low-voltage operation is effective. If the succeeding voltage up-converter from the low voltage to the original high voltage is fast enough, the overall speed is improved. Small-voltage signal transmission on I/O lines or current driving of I/O lines (i.e. where the voltage swing is effectively zero) are typical examples. Because of the small driving capability of a sense amplifier and a heavy I/O capacitance, low-voltage I/O operations are very effective. High-speed direct sensing, in which a small memory-cell signal voltage is directly transferred to the I/O lines without the help of a sense amplifier, is regarded as a low-voltage data-line operation: this is discussed later. A half-V_{DD} data-line precharge combined with a CMOS sense amplifier is also a low-voltage data-line operation. The CMOS amplifier works as a high-speed level converter.

The Reduction of the Number of Circuit Stages. The static operation of a word driver through a combination of a boosted power supply V_{DH} and a CMOS circuit improves the speed with reduced number of circuit stages, because the pulse operation of a word driver line is eliminated. Direct sensing also reduces the number of stages because data-line driving by a sense amplifier is eliminated. Since both lines are heavily capacitive, these reductions are quite effective.

3.4.5 The Multidivision of a Memory Array

To realize stable operation with a higher S/N ratio, a lower power dissipation, and a higher speed, multidivisions of both the data line and the word line are essential, as discussed previously. The resulting multidivided memory array provides high performance despite an increase in memory capacity. When the number of divisions increases, however, the total memory array area increases, because an additional area is necessary at each division. Figure 3.24 shows an example, in which the memory array shown in Fig. 3.10 is simply divided into 16, with four divisions of data line and the word line. Obviously, the bundles of row and column decoders increase at every division, causing an increased

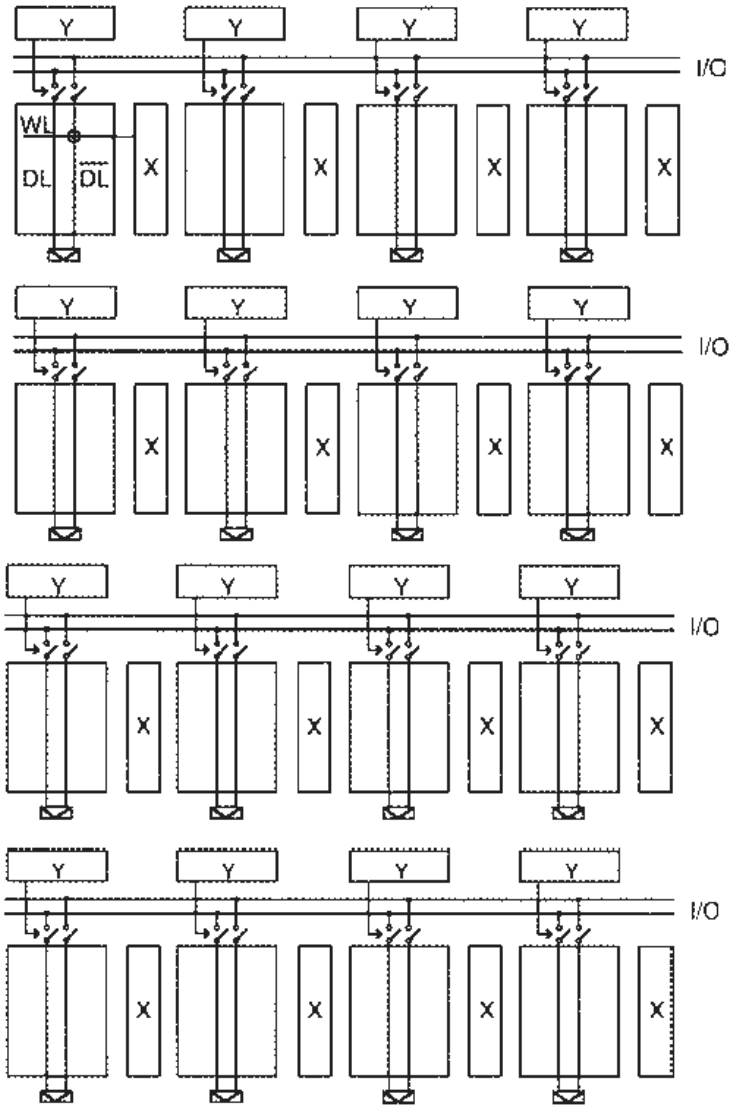


Fig. 3.24. Simply divided memory cell arrays [3.4]. X, row decoders and drivers; Y, column decoders and drivers

chip area. In addition, the resulting increased load capacitance of the address line reduces the speed of the address buffer. Up to the 4 Mb generation, such a simple division had been popular, as a result of giving the first priority to a simpler fabrication process. Recently, however, multilevel metal wiring, using Al or W, has been widely used to suppress the increase in the area due to the division and to obtain a higher S/N ratio and a higher speed. Multilevel metal wiring offers new multidivided-array architectures to minimize the number of decoders, and realizes low-resistance layouts for power and signal lines. Figure 3.25 shows one of the most advanced multidivided arrays [3.4], although many kinds of array divisions using multilevel metal wiring have been proposed. Their features are summarized as follows.

Shared Decoders. A memory array is constructed by repeating the unit subarray shown in Fig. 3.26 in both the row and column directions. A bundle of decoders for the row or column is shared by a number of subarrays. Thus, the load capacitances of the address buffers are lowered. A number of subdata lines (DLs and $\overline{\text{DL}}$s) in the column direction are commonly controlled by a column select line (YL) from the bundle of column decoders. A number of subword lines (SWLs) in the row direction are also controlled by a row select line (i.e. the main word line, MWL) from the bundle of row decoders. Each subarray is selected by activating both bundles of row and column decoders. Here, the row-subarray select line (RX) is necessary to select one of the subarrays in the row direction. Otherwise, information stored in the memory cells of non-selected subarrays, whose sense amplifiers are not activated, would be lost if only MWL was activated, because of the destructive read out characteristics of the DRAM cell. If the MWLs and YLs are laid out over the subarrays with two additional metal layers, there are no area or speed penalties. Examples of the materials used are poly-Si for SWL, the first-level aluminum for MWL, polycide for DL, and the second-level aluminum for YL and RX.

Distributed Driving of Sense Amplifiers. Distributed driving of the sense amplifiers discussed earlier, is achieved by the scheme shown in Fig. 3.26. In this scheme, each common source line (SNL) of sense amplifiers running in the row direction is divided into a number of lines. Each resulting sub-SNL belonging to one subarray is connected to an orthogonally running ground (V_{SS}) line, through a discharging transistor that continues to be turned on during amplification by SDR activation. The first-level aluminum layer is for SNL and SDR, while the second-level aluminum layer is for the V_{SS} line.

The Cross-Shaped Layout Area for the Periphery. Since the peripheral circuits and bonding pads are laid out in a cross-shaped area, the lengths of signal lines such as MWL and YL are shortened, as shown in Fig. 3.22.

Others. The selection function of RX is not needed as long as all subarrays in the row direction are activated by activating their sense amplifiers, despite the increased data-line charging power. Even in this case, the requirement for

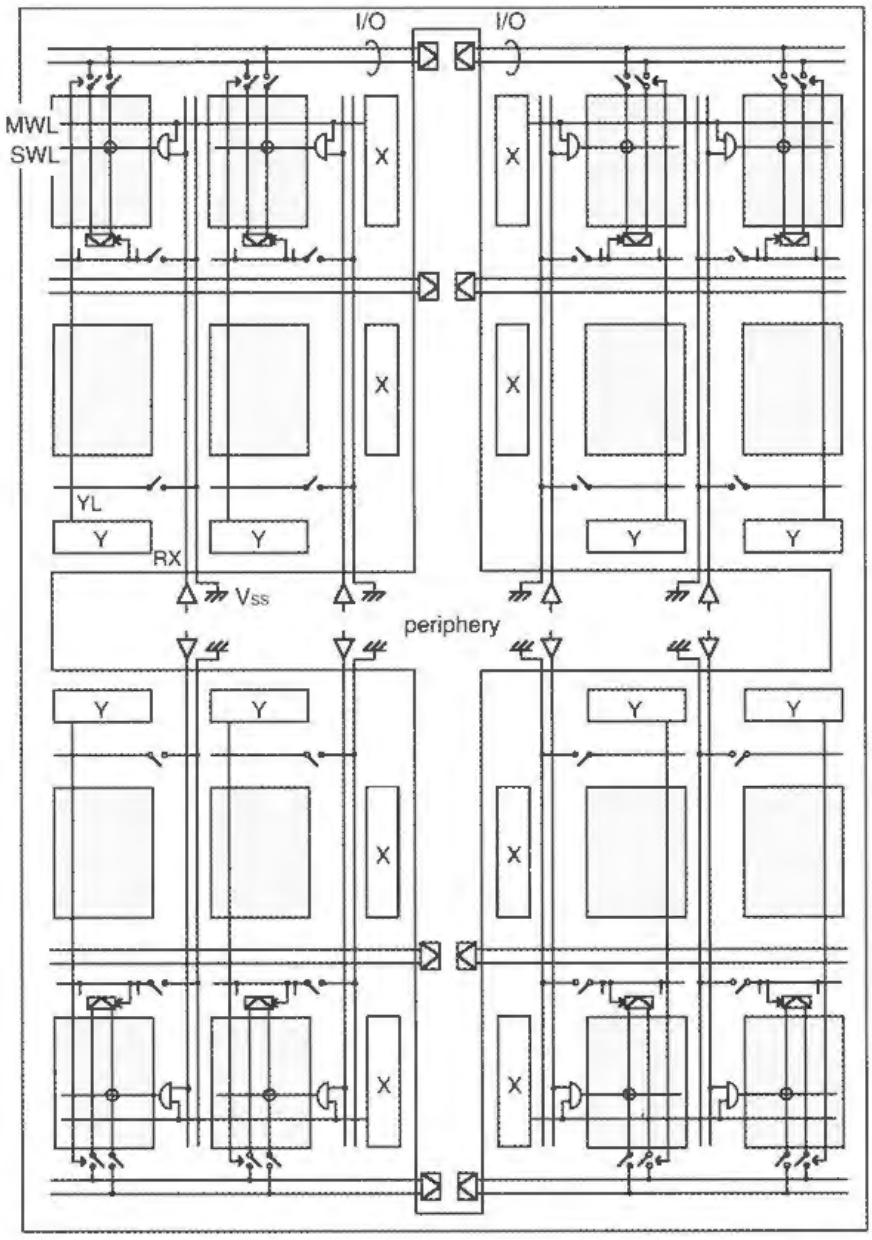


Fig. 3.25. Multidivided memory-cell arrays featuring shared decoders, distributed SA driving and cross-shaped layout areas for peripheral circuit blocks [3.4]

fine patterning for MWLs would be relaxed if a configuration of one MWL shared by a number of SWLs, as discussed later, was adopted. Note that a fine patterning of MWLs at the same pitch as for the SWLs is difficult, because the MWL layer is placed almost at the top surface, as suggested by the structure shown in Figs. 1.18, 1.20, and 1.21. However, the selection function of RX enables flexible designs. For example, the number of subarrays operable at any one time could be changed, depending on the operational modes. If

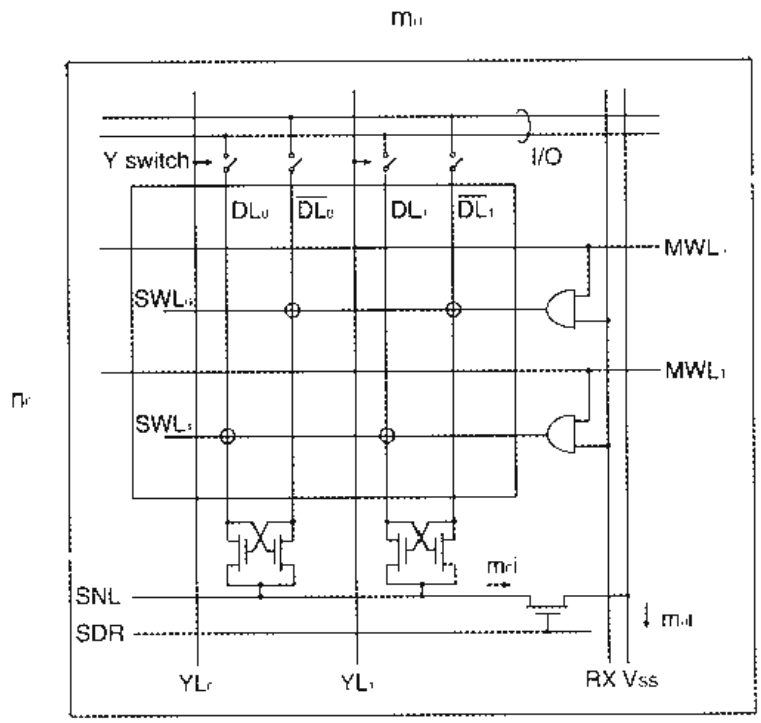


Fig. 3.26. The basic unit of the subarray [3.4]

all subarrays in the row direction are simultaneously activated in a refresh operation, while only one subarray – including the selected memory cell – is activated in a normal operation, the array current I_A in Fig. 3.18 could be negligible, which would allow the chip current to be almost halved.

3.5 The Multidivided Data Line and Word Line

The number of memory-array divisions has been increasing with memory capacity (see Fig. 3.78), to cope with the ever-degrading S/N ratio and the increasing power dissipation and delay time. It depends on the aspect ratios of the chip and memory-cell sizes, and on the memory-cell arrangement, as explained below. In the past, in order to use a small and cheap dual-in-line package (DIP), which must accommodate a rectangular chip with a small number of package pins, a rectangular chip with an aspect ratio of over two has been popular since the advent of address multiplexing in the 16 Kb generation. The folded data-line arrangement memory cell easily realizes a rectangular chip, because the ratio of the data-line pitch p_y to the word-line pitch p_x is about 2.2. Here, the ratio of 2.2 is derived from an assumption that p_y is the sum of a two data-lines pitch and an additional size caused by a data-line to cell contact, while p_x is only one word-line pitch, and the pitches of the data line and the word line are equal. In this case,

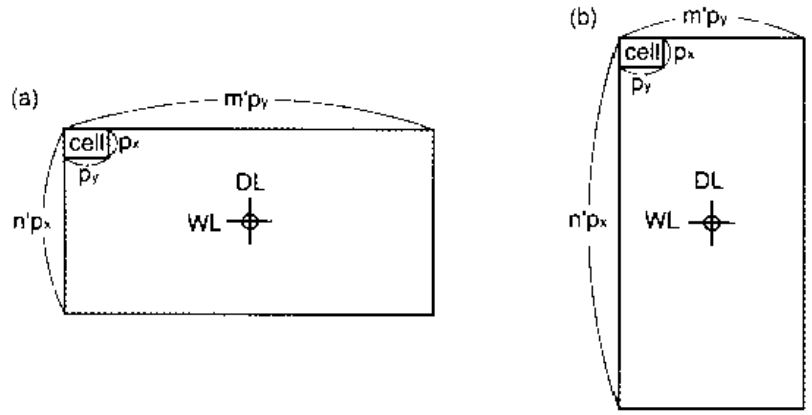


Fig. 3.27. Memory-cell arrangements realizing a rectangular chip ($p_y \fallingdotseq 2.2p_x$) [3.4]. (**a**) $n' = M'$; (**b**) $n' = 4m'$

however, there are two kinds of memory-cell arrangement in a chip, in terms of the number of memory cells in both the row and the column directions; a one-to-one arrangement with $n' = m'$, and a one-to-four arrangement with $n' = 4m'$, as shown in Fig. 3.27. The data lines run along the width of the chip in Fig. 3.27a, while they run along the length of the chip in Fig. 3.27b. The chip performances of the two arrangements are almost equal if the lengths of the sub-data-line and the sub-word-line are equal in both cases.

When the number of array divisions is simply increased, the resulting chip area may become large. Thus, many array divisions that minimize the area penalty by sharing circuit blocks, and/or through the use of multilevel metal wiring, have been proposed.

3.5.1 The Multidivided Data Line

Although many attempts have been made so far, only few data-line arrangements and their divisions have been shown to be practical. An open data-line arrangement was used in the 16 Kb generation, in which the one transistor, one-capacitor cell (1-T cell) began to be widely used, and in the 64 256 Kb generations. On the other hand, the folded data-line arrangement began to be used to some extent in the 64 Kb generation, and since then has become the standard for the 1-T cell.

Divisions of the Folded Data-Line Arrangement. Figure 3.28 shows the basic array configuration of the folded data-line arrangement. A small cell signal developed on a pair of data lines (DL and $\overline{\text{DL}}$) is amplified by a sense amplifier, SA. The amplified signal is outputted on a pair of common I/O lines (I/O and $\overline{\text{I/O}}$) through column (Y) decoders. Figure 3.29 shows two data-line divisions with a shared Y decoder, which was standard in the 64 Kb and 256 Kb generations [3.9]. A Y switch, YS, which is controlled by a Y decoder, connects the selected subdata lines and the corresponding I/O

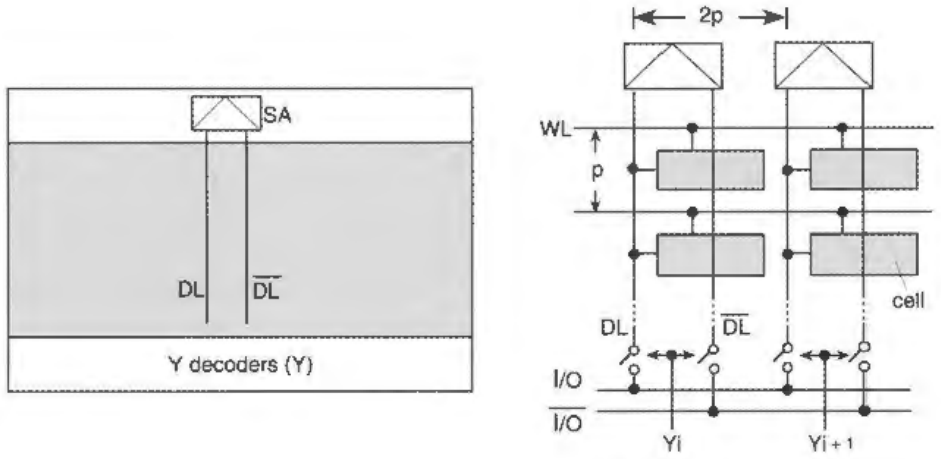


Fig. 3.28. The folded data line arrangement [3.4]. p, Minimum pitch

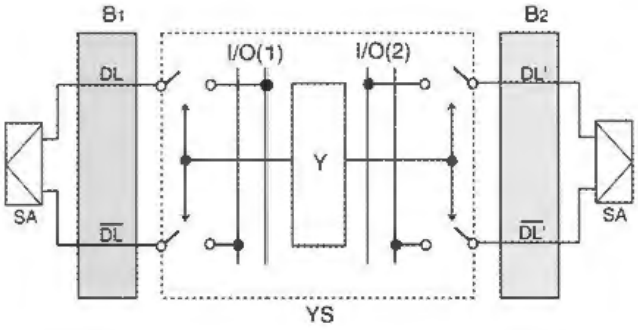


Fig. 3.29. The basic unit for the multidivided data-line scheme [3.4, 3.9]

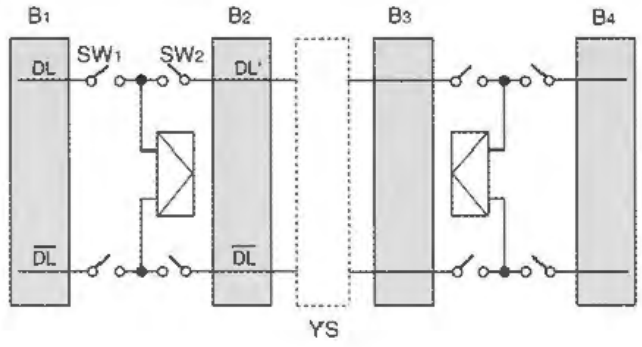


Fig. 3.30. The basic unit for the shared-amplifier multidivided data-line scheme [3.4, 3.6]

lines. Figure 3.30 is the well-known shared amplifier scheme [3.6] widely used in the 256 Kb to 4 Mb generations. This allows two subdata lines to share one amplifier. A small read signal from the subdata line labeled B_1 is fed into the amplifier input by turning on switch SW_1, and then amplified with switch SW_2 turned off. Then, an amplified signal is transferred to subdata line B_2

and to the I/O line by successively turning on switches SW_2 and YS. For a given number of amplifiers and YS switches, this scheme offers a doubled signal voltage. This is because the C_D that contributes to the cell signal voltage is half that of the conventional one. Note that the C_D that contributes to the power dissipation is maintained at the same level. This scheme, however, requires a complicated timing sequence, which results in a slower speed. For a given C_D contributing to the cell signal voltage, this scheme provides double the power dissipation with half the number of amplifiers and YS switches. This results in a smaller chip area with a larger power dissipation. Such is the case for the actual designs shown in Fig. 3.37. The multidivided data line is realized by repeating the basic unit in Fig. 3.29 or Fig. 3.30. To further improve the performance, two types of shared I/O schemes, combined with a shared amplifier, have been proposed. Figure 3.31a features shared I/O lines parallel to the subdata lines [3.10] while Fig. 3.31b shows the one across to the subdata lines. Both schemes inevitably introduce two-level metal wiring to avoid an increase in the chip area due to the additional wiring of the I/O lines and the Y control lines, the YLs, on the memory cell array. However, progress in the development of materials and fabrication processes favors these schemes. Figure 3.31a allows each selected subdata

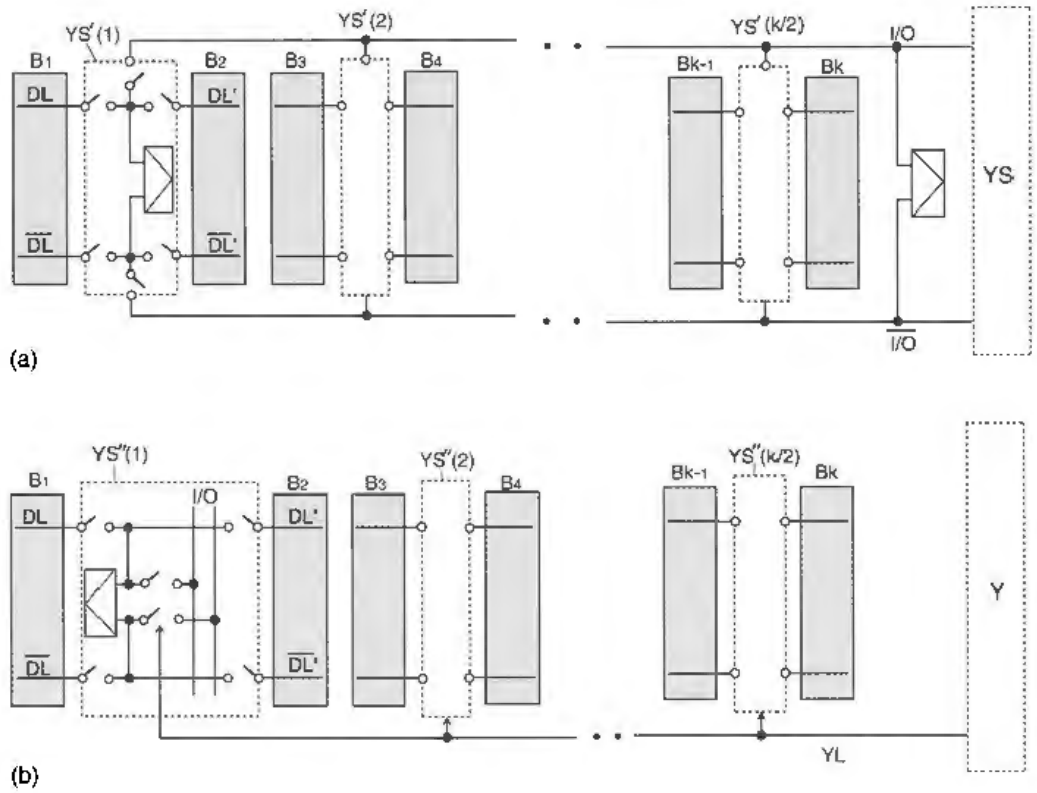


Fig. 3.31. The multidivided data-line scheme for sharing of a sense amplifier, I/O, and a Y decoder [3.4, 3.10, 3.11]. (**a**) I/O parallel to data line; (**b**) I/O orthogonal to data line

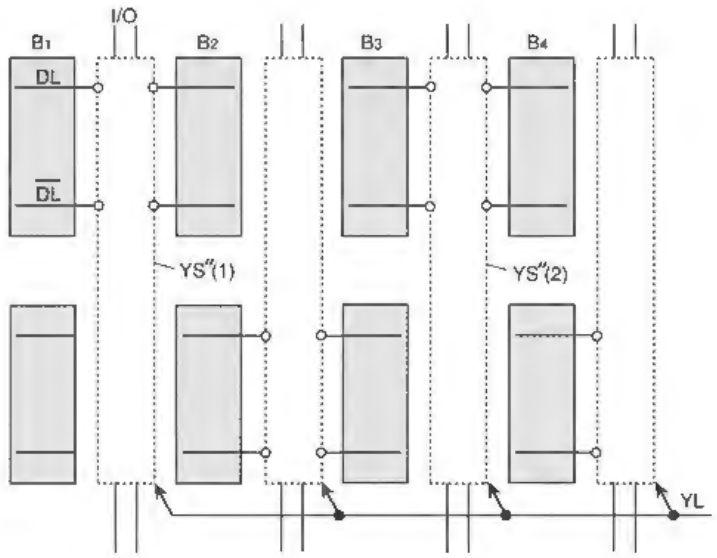


Fig. 3.32. Alternate placement of the Y-switch (YS″) [3.4, 3.12, 3.13]

line to communicate with the common I/O line. Only one YS, placed at the end of the common I/O line, can control the connection between it and the secondary I/O line in YS. This minimizes the data-line capacitance for both the cell signal and power dissipation while enabling a smaller chip area, even for further subdivision, than with the shared amplifier scheme in Fig. 3.30. However, this scheme still suffers from the disadvantage of an additional parasitic capacitance on the I/O line. All the capacitances on the I/O lines are charged and discharged simultaneously, which causes a substantial increase in power dissipation. The circuit shown in Fig. 3.31b overcomes this problem. A selected YL can control the connection between each subdata line and each I/O line. Hence, simultaneous connections are accomplished at the divisions. The lower power dissipation is brought about by a single activation of YL and the smaller number of I/O lines. This scheme has been standard since the 16 Mb generation because it is seemingly the best way to reduce power dissipation as well as to reduce the chip area. In practice, in order to double the layout pitch of the YS″ switches in Fig. 3.31b, an alternate placement of the switches [3.10, 3.12, 3.13], as shown in Fig. 3.32, has been adopted.

Note that the small value of C_D that results from multidivision contributes to high speed because of the shortened RC delay on the data lines.

The Divisions of an Open Data-Line Arrangement. Figure 3.33 shows the basic array configuration of an open data-line arrangement [3.14]. It necessitates two subarrays, each of which is composed of one of two paired data lines. Both the sense amplifier and the Y decoder are laid out at the center of two subdata lines. The data-line and word-line pitches of this arrangement are almost the same as those of the folded data-line arrangement for the standard memory cells shown in Fig. 3.56a. A multidivided data line is realized by

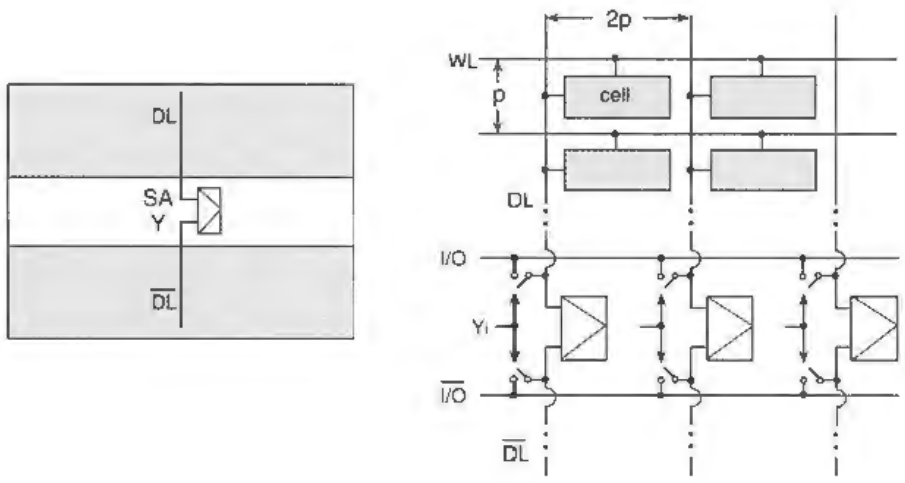


Fig. 3.33. The open data-line arrangement and its multidivision [3.4, 3.14]

repeating this basic unit in the column direction. The arrangement suffers from drawbacks of greater noise generation and a doubling of the number of Y-decoder bundles for multidivision of the data line, compared with a folded data-line arrangement. It is very difficult for each Y decoder and sense amplifier to be laid out within a tight pitch of data lines at the center of the subdata lines, which is subject to noise sources. On the contrary, the folded data-line arrangement features layout flexibility without noise generation. The decoder and amplifier can easily be laid out together at the end of the subarray, or even separately, as shown in Fig. 3.28. Moreover, the total number of Y decoders necessary for multidivision is halved, because of capability of the shared Y-decoder.

Figure 3.34 shows a hybrid data-line arrangement that was proposed in the 1 Mb generation [3.15]. Each data line is constructed of a hierarchy of local open-subdata lines (DL and $\overline{DL}$, etc.) and global folded-subdata lines (d and $\overline{d}$), enabling a separated layout of the Y decoder and the sense amplifier. The memory cell uses a metal word-line and an open diffused data-line arrangement (see Fig. 3.56) and the global data line is made of aluminum. The total data-line capacitance, which is determined by a selected local data line and a global data line, can be made small enough to generate a sufficiently large cell signal. Thus, the need for a sense amplifier at each local data line is eliminated, allowing a small chip area. The short local data line despite the quite large capacitance per unit length of the diffused layer, and the small capacitance per unit length of the aluminum layer despite the long global line, are responsible for the small overall capacitance. A sense amplifier located at each division, combined with an alternate placement of the sense amplifiers, has been also proposed [3.16]. However, hybrid arrangements have not been used for DRAM products because of the inherently large noise generation of the open data-line arrangement.

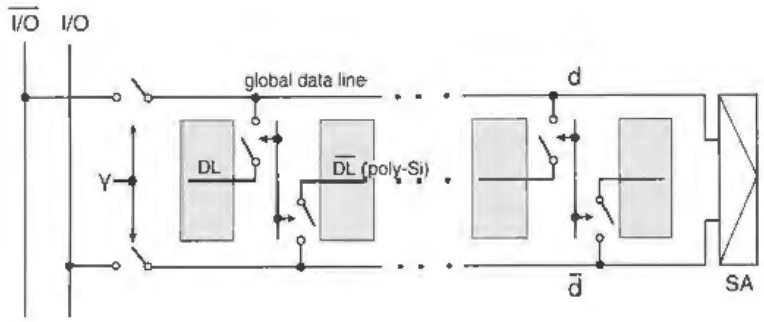


Fig. 3.34. The hybrid data-line arrangement [3.4, 3.15]

Summary. With increasing memory capacity, the ever-increasing number of memory cells connected to one data line causes an increased C_D and thus an increased power, as well as a poor signal-to-noise ratio. One practical solution is to divide a data line into several sections and to activate only one section, as explained previously.

In the early days, the methods of data-line division applied to commercial chips changed with every generation. In the 16 Kb generation the open data-line arrangement with two data-line divisions shown in Fig. 3.33 was standard. In the succeeding 64 Kb and 256 Kb generations, however, two data-line arrangements coexisted. For the open data-line arrangement, two and four divisions (i.e. a simple duplication of the two divisions) were popular. For the folded data-line arrangement, the two divisions shown in Fig. 3.29 were standard in the 64 Kb generation, and then a duplicate of this arran-

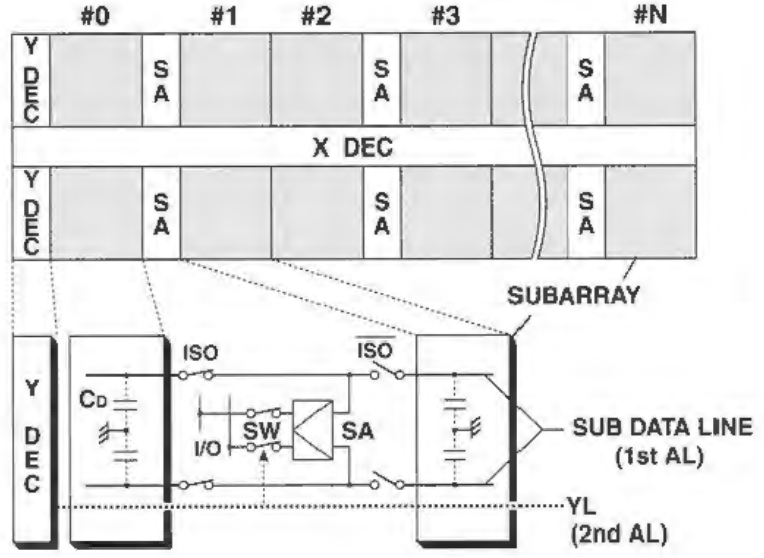


Fig. 3.35. Multidivided data-line architecture with a shared SA, a shared I/O, and a shared Y decoder scheme [3.11]

gement and a shared sense amplifier, shown in Fig. 3.30, were added in the 256 Kb generation. However, the division process has been almost completed in the 16 Mb generation with the combination of a shared SA, shared I/O, and a shared Y decoder, as shown in Fig. 3.35, this being the same as in Fig. 3.31b [3.3]. The shared I/O further divides a multidivided data line into two parts, which are selected by the isolation switches, ISO and $\overline{\text{ISO}}$. The shared SA provides an almost doubled cell signal with a halved value of C_{D}. A shared Y decoder can be constructed without any increase in area by using the second-level metal wiring for the column selection line (YL). The partial activation is performed by activating only one sense amplifier along the data line.

Figure 3.36 shows trends in the resulting C_{D} and in other parameters that contribute to the cell-signal voltage [3.4]. In spite of reductions in C_{S}

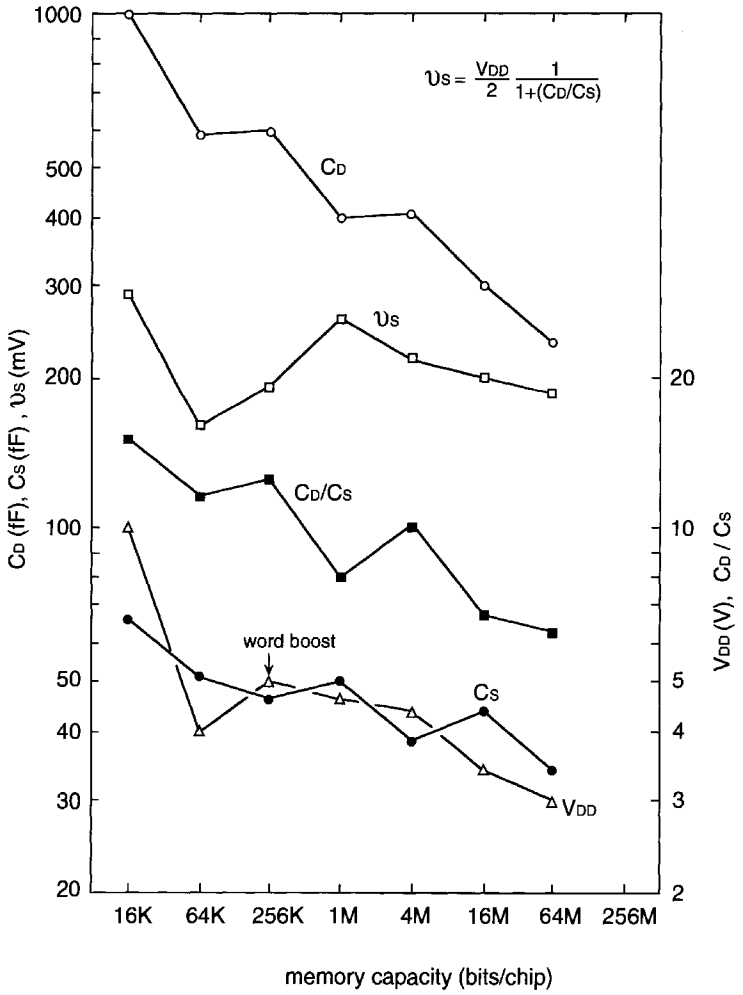


Fig. 3.36. Trends in signal voltage and relevant parameters [3.4]. Data are from ISSCC

and V_{DD} in each successive generation, the signal voltage has been maintained at about 200 mV by reducing C_D.

Figure 3.37 shows trends in the total data-line charging capacitance, C_{DT} ($= mC_D$). C_{DT} has been minimized by sharing the SA, I/O, and Y decoder, and by increasing the number n, as described later. Figure 3.38 shows trends in the total data-line dissipating charge, Q_{DT} ($= mC_D\Delta V_D$). Q_{DT} has been suppressed as much as possible with the help of the ever-reducing operating voltage, as shown in Fig. 7.11.

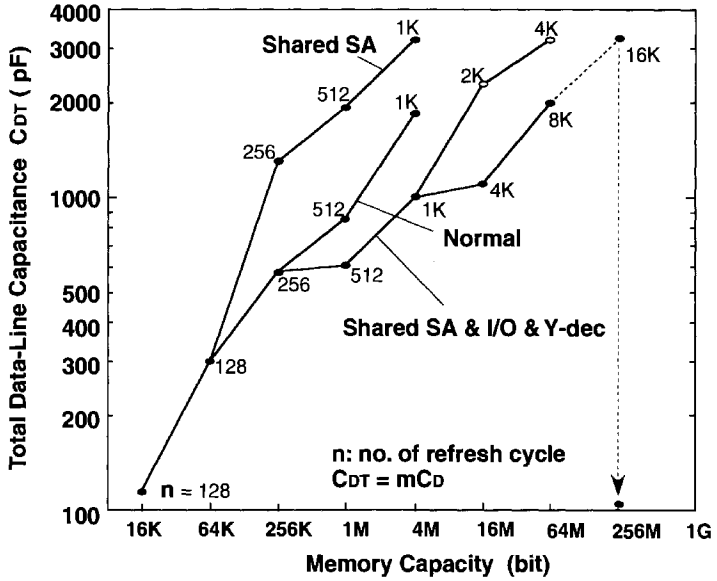


Fig. 3.37. Trends in the total charging capacitance of data lines [3.3, 3.4]. A C_D of 200 fF and a ΔV_D of 1 V are assumed for 256 Mb

A drawback of the multidivided data line is the increased number of sense amplifiers, and thus an increased chip area despite the use of multilevel metal wiring. Figure 3.39 shows trends in the number of sense amplifiers [3.4]. The number of sense amplifiers has increased rapidly, because the number of memory cells connected to one sense amplifier is limited in order to reduce C_D. Note that a 1 Gb chip needs as many as one million sense amplifiers. The reason why the memory array has occupied only 65% at most of chip area stems from this increase.

3.5.2 The Multidivided Word Line

The open data-line arrangement using a low-resistance metal word line (Fig. 3.56) achieves a high speed without multidivision of a word line. Instead, it generates large array noises, as explained in 4, and thus it was replaced in the 1 Mb generation by the folded data-line arrangement. For the folded data-line arrangement cell (Fig. 3.56), however, multidivision of a word line

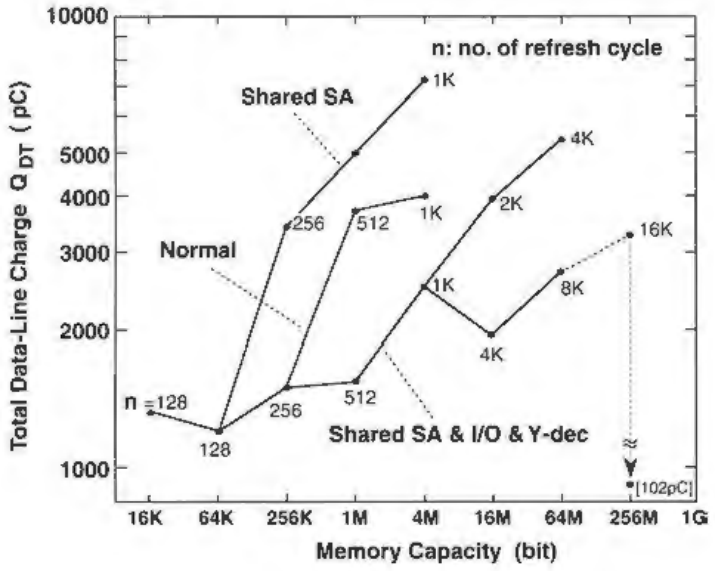


Fig. 3.38. Trends in the dissipating charge of data lines [3.3, 3.4]. A C_D of 200 fF and a ΔV_D of 1 V are assumed for 256 Mb

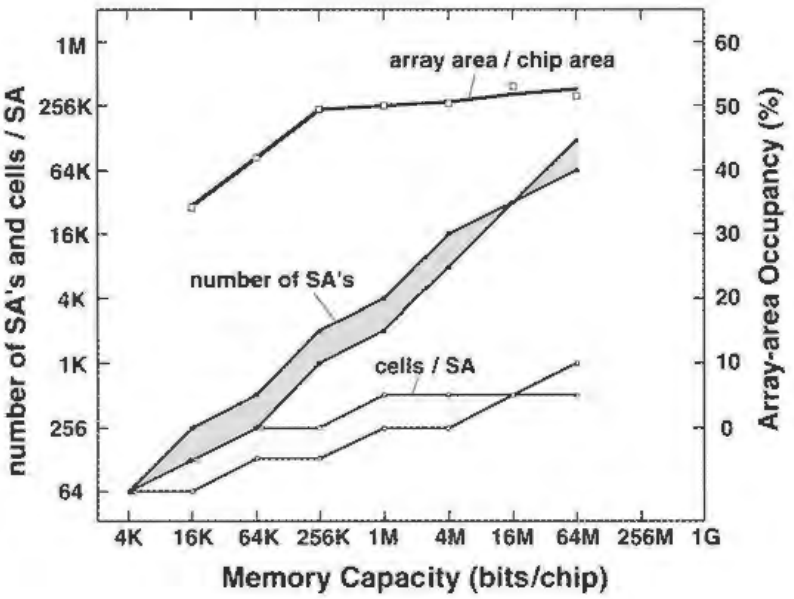


Fig. 3.39. The ever-increasing number of sense amplifiers in a DRAM chip [3.4]

is needed to reduce the quite large RC delay of a word line made of a resistive poly-Si or polycide. In the past, the number of divisions has been determined by compromising the delay, with the resultant area penalty. In the 64 Kb generation, in which the word line was still short and the speed requirement was not so severe, two divisions of a word line, as shown in Fig. 3.40a [3.9], was enough. In the 256 Kb generation, however, poly-Si was replaced by a lower-

resistance polycide, in order to double or quadruple the number of memory cells to be driven. To further reduce the *RC* delay, a poly-Si or polycide word line strapped with a low-resistance aluminum line in each string of 64–128 cells [3.17], as shown in Fig. 3.40b, has been widely accepted in commercial chips since the 1 Mb generation. In the 64 Mb generation, even a hybrid division of the two types shown in Fig. 3.40 has been proposed. However, the aluminum-strapped word-line structure suffers from some drawbacks. It becomes difficult to achieve fine patterning of aluminum at a tight pitch of word lines on a hilly surface, while still connecting to the poly-Si line at the bottom, as the memory cell is miniaturized. In addition, even the word-line structure starts to create quite a large *RC* delay as many memory cells are connected to a word or subword line. A multidivided word-line scheme using a hierarchical word-line structure [3.18, 3.19], as shown in Fig. 3.25, solves these problems, and is discussed later in detail.

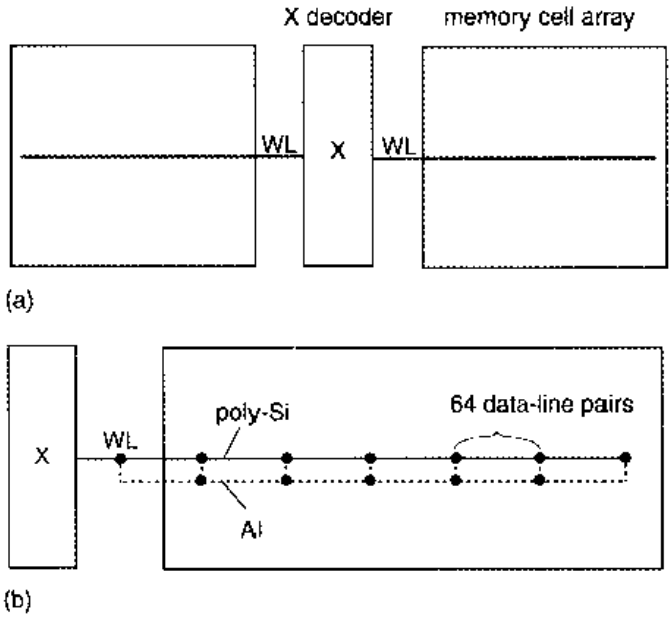


Fig. 3.40. Reductions in word-line delay [3.4, 3.9, 3.17]. (**a**) 64 Kb–256 Kb. 64 Kb: poly-Si (30 Ω/□), 64 data-line pairs per WL. 256 Kb: polycide (1–5 Ω/□), 256 data-line pairs per WL. (b) 1 Mb and beyond: poly-Si (50 Ω/□), Al (0.1 Ω/□)

3.6 Read and Relevant Circuits

3.6.1 The Address Buffer

In the NMOS era of the 16–256 Kb generations, a differential address buffer using an on-chip reference voltage [3.14] was widely used. Since the 1 Mb

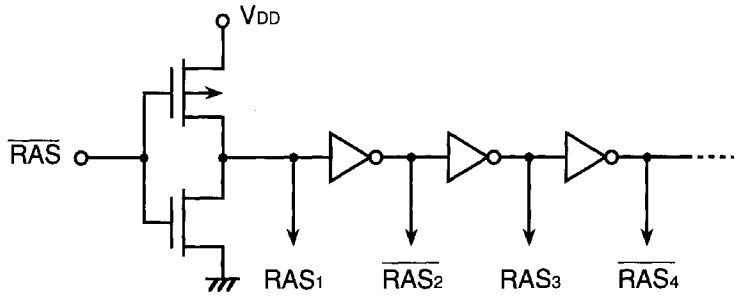


Fig. 3.41. The $\overline{\text{RAS}}$ clock buffer [3.21]

generation, various simple CMOS address buffers have been proposed. In principle, the same circuit configuration is applicable to both the row and column address buffers. Recently, however, a high-speed column address buffer has been especially important to enhance the data throughput by shortening the column address access time t_{AA}, which is dominated by the address buffer (about 40% of t_{AA} [3.20]).

Figure 3.41 shows a typical $\overline{\text{RAS}}$-clock buffer [3.21] to control the row address buffers. To discriminate between the TTL logic levels of over 2.4 V or below 0.8 V for an input $\overline{\text{RAS}}$ signal, the logical threshold voltage is adjusted to be about 1.6 V by tuning the channel-width ratios of the NMOS and the PMOS at the first input stage. A multistage CMOS circuit controls the row internal circuits, using pulses with differing polarities and delays.

Figure 3.42 shows an address buffer that features a cross-coupled differential amplifier [3.21]. It enables an almost constant speed due to a differential circuit configuration, independently of the power-supply noise. Address buffers consisting of inverters, similar to the above $\overline{\text{RAS}}$ buffer, have also been widely accepted. In general, however, the operation of inverter-type buffers is susceptible to power-supply noise [3.20]. For example, as soon as the input logic signals to many buffers are simultaneously switched to the other logic state so that the ground (V_{SS}) line voltage is instantaneously raised and maximized, the speed of each buffer is degraded, with a reduced NMOS gate–source voltage. For a 64 Mb design [3.20], a simultaneous switching of 13 address buffers caused a V_{SS} noise of 0.4 V and a speed difference of 2.3 ns between the inverter and cross-coupled types, although the difference depended on the quality of the V_{SS} layout. In the figure, the input voltage of address Ai is compared with a reference V_{REF} (1.6 V) to discriminate between a high logic level (H) and a low logic level. The resultant differential signal developed between N_1 and N_2 is quickly amplified to a full V_{DD} by the cross-coupled amplifier, as a result of the application of $\overline{\text{RAS}}_2$. After that, the complementary addresses a_i and $\overline{a_i}$ are generated by the application of $\overline{\text{RAS}}_4$ to address latches (ALCs). Note that during standby periods (i.e. $\overline{\text{RAS}}$: H) both a_i and $\overline{a_i}$ are kept low and there is no current path in the buffer.

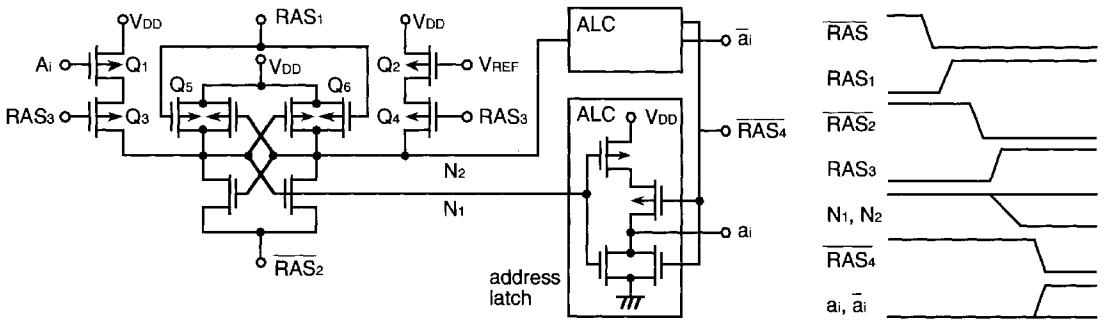


Fig. 3.42. The cross-coupled address buffer [3.21]

Figure 3.43 shows a typical address transition detector (ATD) [3.22], although the variations are shown in Fig. 7.18. Exclusive OR of a'_0 and the address delayed by τ generates a short pulse every a'_0 transition. All the short pulses generated from all the address input transitions are summed up to one ATD pulse, $\overline{EQ}$. Thus, an ATD pulse is generated at any address transition so as to control internal circuits instead of external clocks. If any address transition is quickly detected by ATD and the resultant ATD signal precharges the I/O line in advance, a data line will be selected using addresses just after the transition, so that a data on the data line is outputted on the I/O line without waiting for the I/O precharging. Thus, a long I/O precharging time can be concealed. The ATD signal reduces the power dissipation of main amplifier by cutting the dc current during periods when it is not needed.

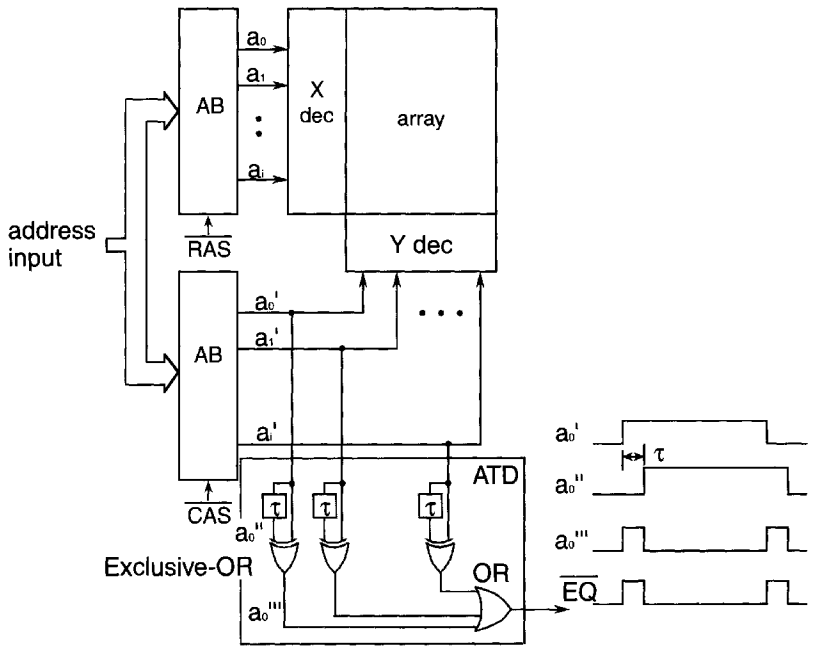


Fig. 3.43. The address transition detector (ATD) scheme [3.22]. AB, address buffers

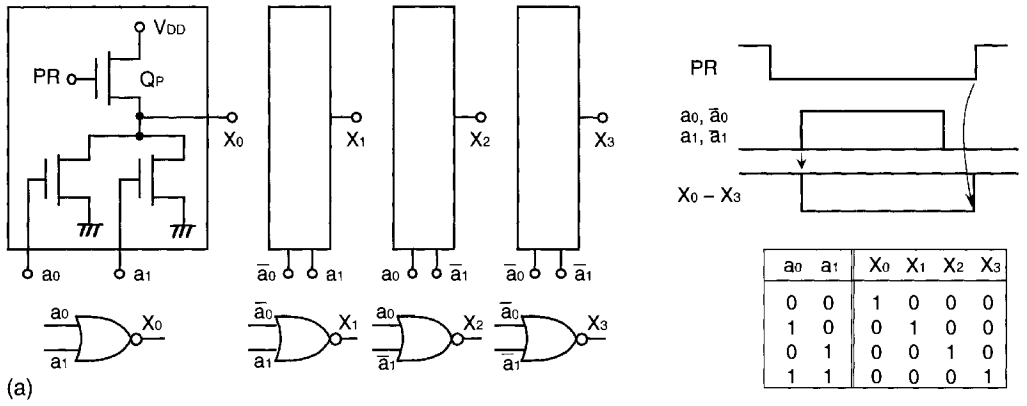
3.6.2 The Address Decoder

Major concerns for the address-decoder block are power dissipation, speed, and area, because the block includes a huge number of circuits and occupies quite a large segment of the chip.

There are two kinds of decoder; row decoders and column decoders. In DRAM design, unlike SRAM designs, the circuit configurations of the two are totally different. Each row decoder must be a dynamic circuit, while each column decoder can be a static circuit, as explained previously. Note that to precharge all the row decoders without any dc current path, all of the complementary addresses are fixed at a low level during a precharge period, as shown in Fig. 3.42.

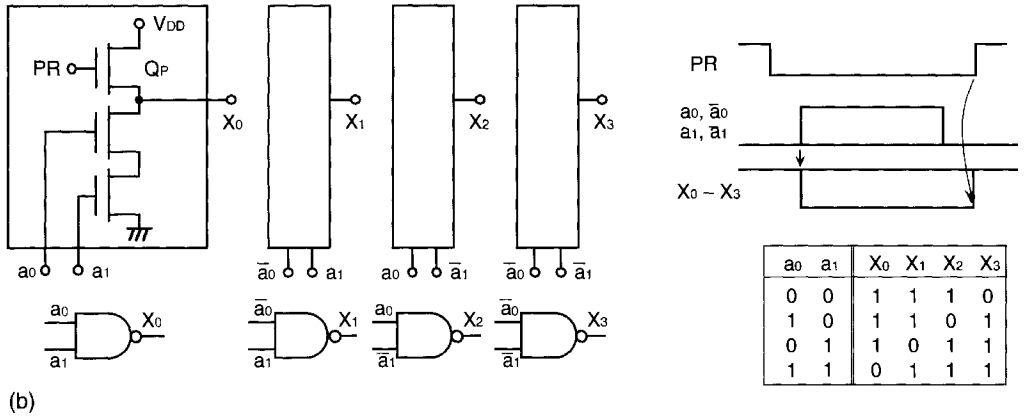
Figure 3.44 shows dynamic and static decoders, exemplified by two-bit address signals. There are two kinds of dynamic decoder in DRAM applications; NOR and NAND decoders. First, all of the output nodes (X_0–X_3) are precharged to V_{DD} by transistors Q_P, while keeping all addresses low. Then, according to the succeeding valid address signals, the output nodes are discharged or kept high. Obviously, in NOR decoders all output nodes except for a selected one are discharged, while in NAND decoders all output nodes except for a selected one are kept high. Thus, NOR decoders suffer from a drawback of the large charging and discharging power. The power increases with memory capacity, because an increased number of the nodes – for example, a few thousands, for multimegabit DRAMs – is involved. On the other hand, in NAND decoders only one node is discharged or charged up, independently of memory capacity. NAND decoders, however, suffer the drawback of a slower speed, because the node is discharged by serially connected (i.e. stacked) transistors. The number of stacked transistors is also limited by the body effect of the transistor. Static NAND decoders for the column are simple, as shown in Fig. 3.44c.

Figure 3.45 shows applications of dynamic decoders to the row and static decoders to the column. In dynamic decoders, a word line WL is activated by an RX pulse that is applied after the decoder output Xi has been settled. In the selected decoder, the Q_W gate–drain (i.e. RX terminal) capacitance C_{GD} is large, because the Q_W gate stays at the high level of $V_{DD} - V_T$. Thus, an RX pulse positively going from 0 V to V_{DD} can boost the Q_W gate voltage. The boost ratio is large, because a diode Q_D isolates the Q_W gate from the node Xi capacitance. Due to the resulting boosted gate voltage, to higher than $V_{DD} + V_T$, the word line is quickly driven to V_{DD}. In the non-selected decoders, an RX pulse application never raises the Q_W gate voltages, since the gate voltages are 0 V and thus their C_{GD} values are almost zero. For NOR decoders, even the heavily capacitive Q_W gate is quickly discharged by at least one transistor of the decoder. For NAND decoders, however, to accomplish rapid decoding, Q_W is driven with the help of a small CMOS inverter, whose input capacitance is small enough to be quickly driven even by stacked transistors. Despite the area penalty, the inverter added to each



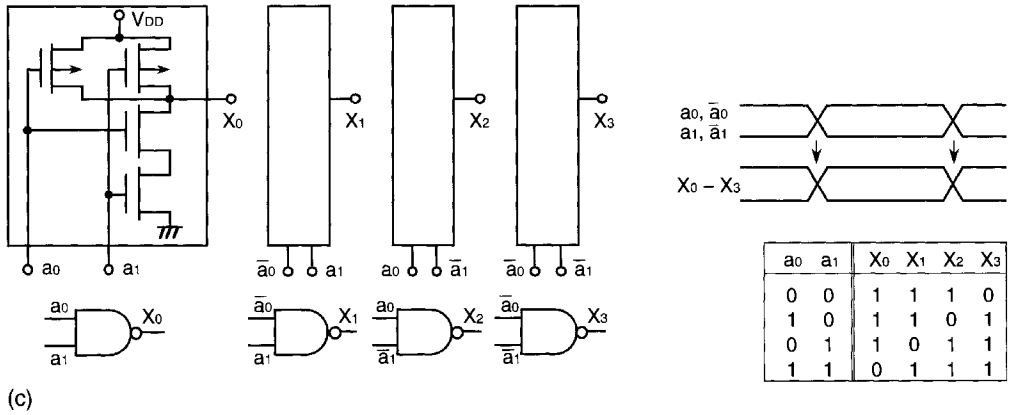
(a)

a_0	a_1	X_0	X_1	X_2	X_3
0	0	1	0	0	0
1	0	0	1	0	0
0	1	0	0	1	0
1	1	0	0	0	1



(b)

a_0	a_1	X_0	X_1	X_2	X_3
0	0	1	1	1	0
1	0	1	1	0	1
0	1	1	0	1	1
1	1	0	1	1	1



(c)

a_0	a_1	X_0	X_1	X_2	X_3
0	0	1	1	1	0
1	0	1	1	0	1
0	1	1	0	1	1
1	1	0	1	1	1

Fig. 3.44. Decoders and operations, exemplified by two address bits [3.4]. (**a**) Dynamic NOR; (**b**) dynamic NAND; (**c**) static NAND

decoder never increases the decoder power because it is a CMOS inverter. The reason why NMOS NOR decoders have been replaced by CMOS NAND decoders since the 1 Mb generation is just to reduce the ever-increasing decoder

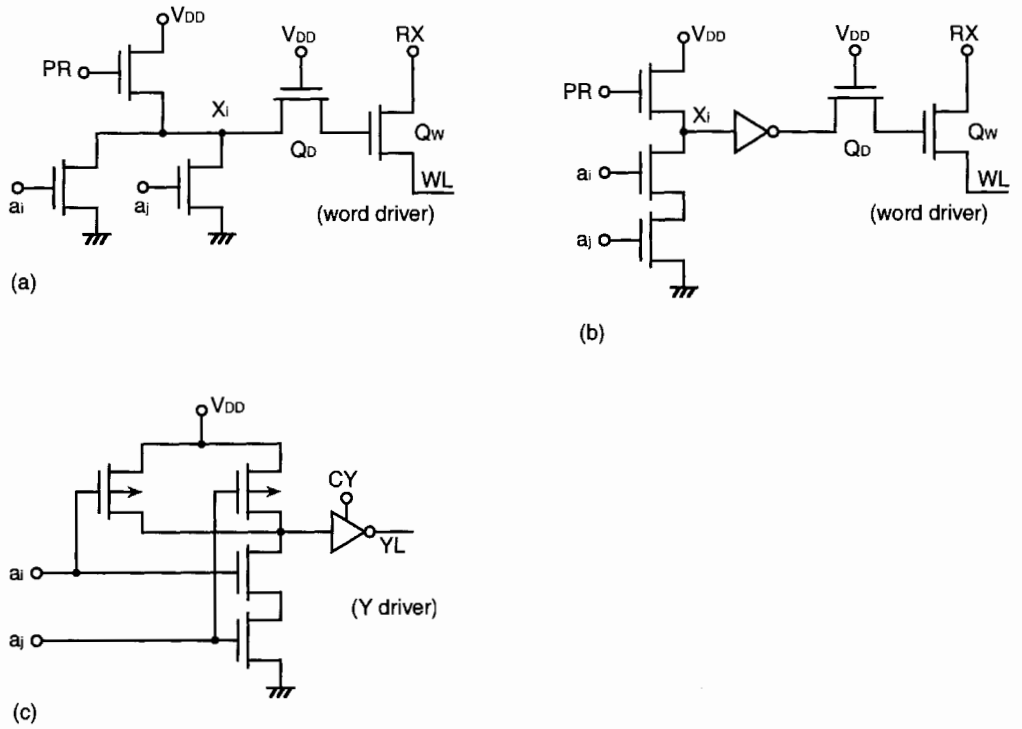


Fig. 3.45. Applications of decoders to word and column drivers [3.4]. (**a**) Dynamic NOR; (**b**) dynamic NAND; (**c**) static NAND

power. Even for a small memory capacity chip of 1 Mb, CMOS NAND decoders have reduced the decoder power down to 4% of that needed for NMOS NOR decoders [3.7, 3.23] as shown in Fig. 3.17. However, to improve the performance of CMOS NAND decoders further, it is essential to reduce the number of stacked transistors. This is realized by predecoding schemes [3.21], as follows.

A predecoding scheme achieves a faster decoding and area reduction of a decoder while reducing the number of stacked transistor in a CMOS NAND decoder. In addition, it reduces the input capacitance and the necessary address input lines of the decoder. Figure 3.46 compares predecoding schemes [3.4]. Direct decoding, 2 bit predecoding, and 3 bit predecoding are shown in Figs. 3.46a–c, respectively. A circle in the figures denotes a transistor connection. For example, when a high level is applied to a_0 and a_1 in Fig. 3.46a, decoders X_0, X_4, and X_8 are selected as a result of NAND decoding. Each 2 bit predecoder can select one of four address input lines coming to the decoders by using two sets of complementary addresses from two address buffers, while each 3 bit predecoder can select one of eight address input lines by using three sets of complementary addresses. Here, let us cite an example of a total external address bits of 6 bits (A_0–A_5). The numbers of address lines to the decoders for direct decoding, 2 bit decoding, and 3 bit decoding are 12, 12, and 16, respectively. The numbers of transistors

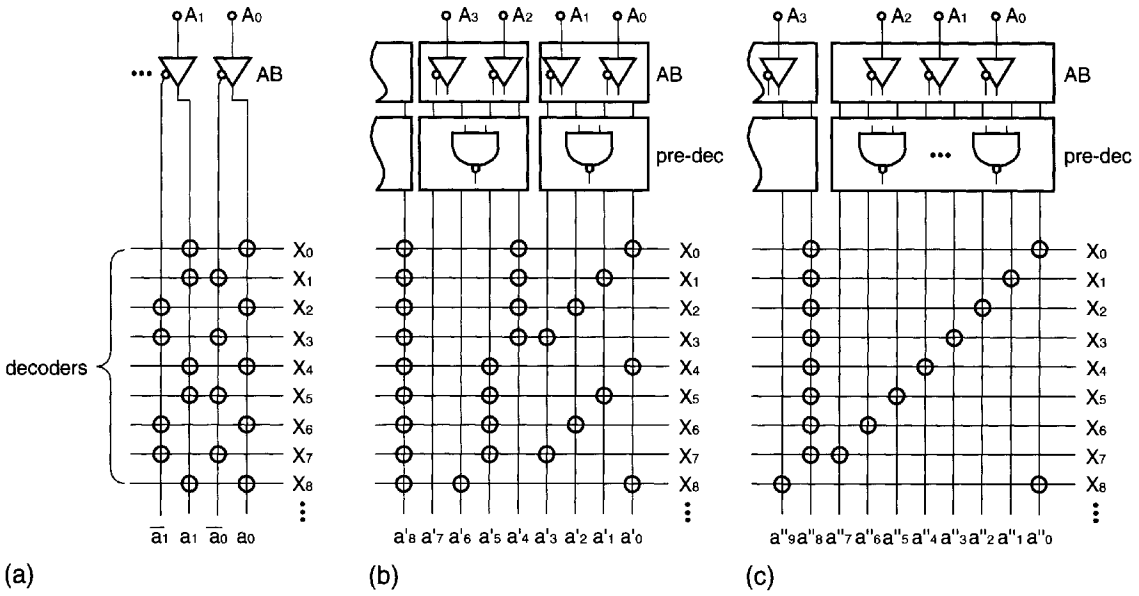


Fig. 3.46. Predecoding [3.4]. (**a**) no predecoding; (**b**) 2-bit predecoding; (**c**) 3-bit predecoding

connected to each address line are 32, 16, and 8, and the numbers of transistors consisting of each decoder are six, three, and two, in the same order. Hence, the predecoding schemes achieve a higher speed with reduction of the address-line capacitance and the number of stacked transistors needed for each NAND decoder, given an acceptable number of address lines. They also reduce the decoder area. The resulting improvement in speed offsets an additional delay caused by the predecoders. Here, predecoding schemes with more than 4 bits are not practical because of a rapid increase in the number of address lines.

Figure 3.47 shows a reduction in the delay of an address line [3.24], which is an example of the buffer insertions shown in Fig. 3.20. Quite a long delay time is developed, despite the aluminum address line, because the line becomes resistive and capacitive due to fine patterning, and is loaded by the distributed gate capacitances of the decoder transistors, as shown in Fig. 3.47a. However, the delay is reduced by the multidivided decoder shown in Fig. 3.47b. The resulting block decoder is driven by a buffer. Each block is constructed so as to correspond to a subarray and only one block is selectively activated by the subarray activation pulse Φ_i.

3.6.3 The Word Driver

A word driver needs to be designed carefully – more so than a column driver – because its load has the following unique features:

1. *A Boosted Word Voltage.* The need for a boosted word voltage, for full write and read operations, means that row decoders and word drivers

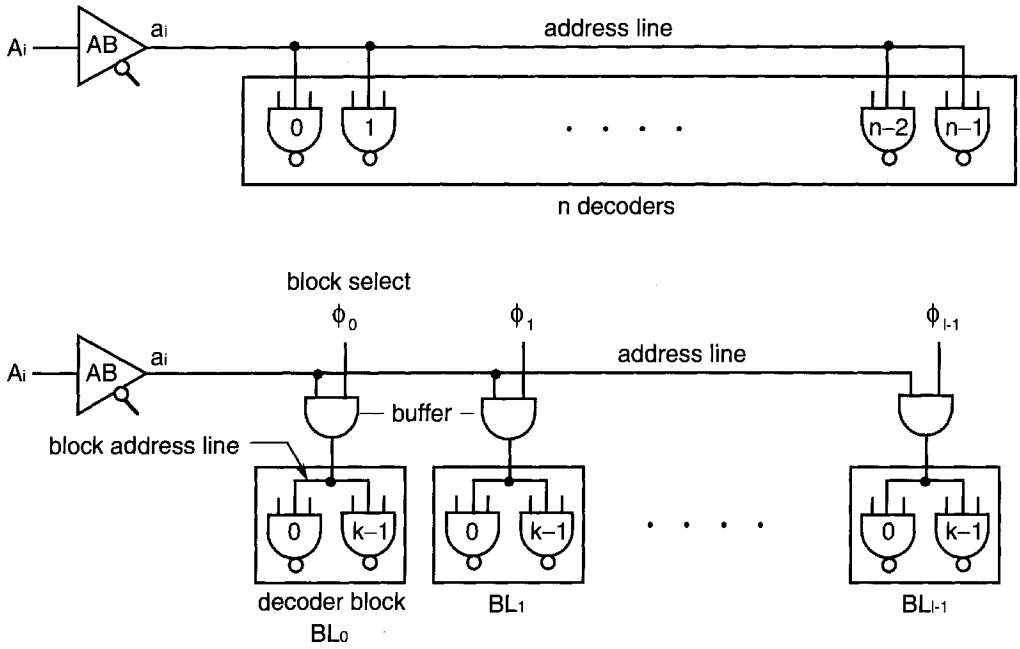


Fig. 3.47. The delay reduction of an address line running on decoders (*upper*) by insertion of buffers (*lower*) [3.4, 3.24]

have complicated designs. On the other hand, column-relevant circuits, such as column decoders and drivers, do not need any boosting. Even without boosting, an amplified signal voltage from a data line can be transmitted to the I/O line, and a data-input voltage of V_{DD} can be fully transmitted from the I/O line to the data line with the help of the CMOS sense amplifier.

2. *A Large Word-Line Capacitance and Resistance.* The electrical characteristics of a word line differ from those of column line YL in the shared Y decoder scheme. Word-line capacitance is quite heavy, because of connections with many memory cells. On the other hand, column-line capacitance is light, because a small number of transistors, equal to the number of data-line divisions (i.e. q in Fig. 3.14), are connected to a YL line. Thus, a larger word driver is needed. As for line resistance, there is also a large difference between the two. A word line made of poly-Si or polycide is resistive for a folded data-line cell, while a column line is made of aluminum.
3. *Loss of Stored Information.* If the stored charges of a non-selected cell are allowed to escape to a data line through the transfer transistor, the refresh time or S/N ratio of the cell is degraded, as discussed in Chap. 4. Thus, noise suppression is essential on non-selected word lines. Moreover, to avoid loss of stored information, data-line precharging must be started after completely turning off the word pulse. Such considerations are not

needed for the column. Here, an explanation of the column driver is omitted in what follows, because the driver is almost the same as in Fig. 3.45.

A Basic Word Driver. Figure 3.48 shows the basic unit of conventional word drivers [3.13, 3.21, 3.23]. Each word line is divided into two, to reduce word-line delay (see Fig. 3.40), and the resulting divided word line has its own word driver, Q_W. Since a CMOS NAND decoder cannot be placed at a tight word-line pitch, it is shared with two sets (left and right) of four word drivers, although only the right section is shown in the figure. Address signals X_j and X_k are inputted from 2-bit predecoders to the decoder. Each of the four word drivers selected by the decoder is selectively driven by decoded row select lines RX (RX_0–RX_3), enabling the corresponding word line to be driven. The RX drivers in Fig. 3.48b provide a boost word voltage to one of RXs as a result of two-address bit (a_0, a_1) decoding, as follows. At first, node WDL is precharged almost to V_{DD} during the precharge period (i.e. $\overline{RP}$: H) and all address signals and thus all RX lines are fixed at 0 V. When the addresses have been valid after starting activation with $\overline{RP}$: L, a clock Φ_B generated by the $\overline{RAS}$ buffer is applied, so that only one selected RX line is driven at a high enough voltage, boosted by C_B.

The latch transistor Q_L in the word driver suppresses noise on each non-selected word line while fixing the word line at 0 V as follows. During precharge period all of the non-selected word lines are fixed at 0 V because the Q_Ls are turned on. When one set of four word lines is selected by a decoder, the gate voltages of the corresponding four Q_Ls are changed from V_{DD} to 0 V and the Q_Ls are thus cut off. At this moment, the gate voltages of the corresponding four word-driver transistors Q_W are increased from 0 V to $V_{DD} - V_T$. During this process, the word lines are not at any floating voltage that easily couples noise to the lines, because the four word lines WL_0–WL_3 are fixed at the voltages (i.e. 0 V) of RX_0–RX_3 through the respective Q_Ws. After that, for example, RX_0 is selected and a V_{DH} pulse is sent to WL_0, fixing the remaining non-selected word lines at 0 V. Note that all of the word lines belonging to the non-selected decoders continue to be fixed at 0 V, because the Q_Ls are turned on. Thus, the noise coupled to each non-selected word line, even during signal amplifications performed at a large voltage swing of V_{DD} or a half-V_{DD} on data lines, can be sufficiently suppressed. To further reduce noise on the word lines, another scheme of an additional transistor, which is controlled by address signals, on each word line has been proposed [3.25].

High-speed driving of the RX line is also important, because a long delay is developed by the heavy loading of the large Q_Ws and the long RX line running along a memory array. An RX driver placed at the end of the subarrays in Fig. 3.49a increases the line delay. However, a RX driver for each subarray in Fig. 3.49b [3.26], which is similar to the scheme in Fig. 3.47, shortens the delay. In this scheme, only one subarray is selected by the address signals and a subarray selection signal Φ_i. The node WDL of the RX line in Fig. 3.48 corresponds to a line WDL that is common to a number of subarrays in

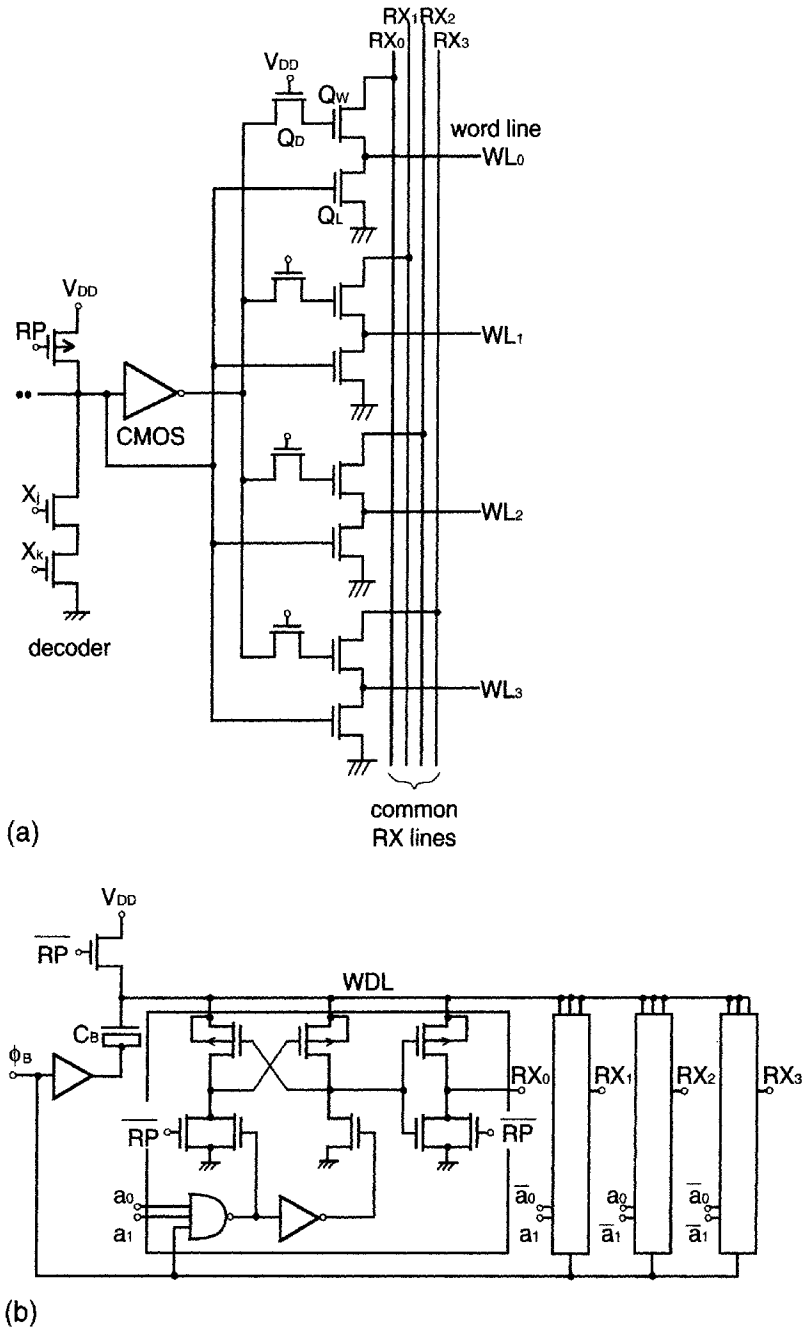


Fig. 3.48. The configuration of word drivers and relevant circuits [3.4, 3.13, 3.21, 3.23]. (**a**) Word drivers; (**b**) RX drivers

Fig. 3.49b. As soon as $\overline{\text{RAS}}$ activation starts, the heavily capacitive WDL line is boosted. After that, the addresses are valid and a subarray is thus selected so that a V_{DH} pulse is applied to a common RX line belonging to the subarray. Consequently, the boosting time of the WDL line can be concealed and the RX line that the RX driver must drive is shortened to one subarray. Thus, the driving speed is improved.

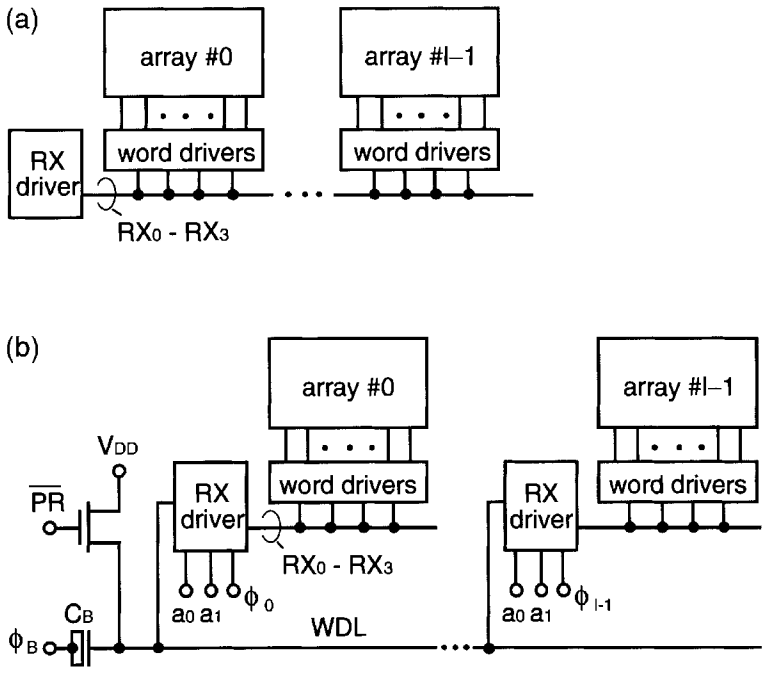


Fig. 3.49. Driving schemes of RX lines [3.4, 3.26]. (**a**) Direct driving of RX lines; (**b**) selective driving of multidivided RX lines

The Voltage-Stress Relaxed Word Driver. This is the word driver that must operate at the highest voltage in a DRAM chip. Therefore, many circuits have been proposed to relax the voltage stress applied at normal and/or burn-in tst operations. They are categorized as the use of PMOS transistor in the word driver instead of NMOS transistor, the changing of the boost ratio according to V_{DD}, and the use of a well-regulated boosted dc voltage.

The PMOS Output Transistor [3.4]. The stress voltages applied to the word-driver output transistor are high, because the word-line voltage must be higher than $V_{DD} + V_{TM}$ (where V_{TM} is the threshold voltage of the memory-cell transistor). However, the PMOS transistor can relax the voltage stress, ensuring high reliability, as follows. For the NMOS transistor in Fig. 3.50a, in order that an increased voltage V_{DH} ($> V_{DD} + V_{TM}$) boosted by C_B at the drain (i.e. RX) is available at the source (i.e. word line), the gate voltage must be boosted to a level that is somewhat higher than $V_{DH} + V_{TD}$ by utilizing C_{GD}. Here, V_{TD} is threshold voltage of the output transistor, which is usually smaller than V_{TM} due to the narrow channel effect, and so on. This dual boosting raises the gate voltage to a considerably high voltage. The PMOS output transistor in Fig. 3.50b does not need dual boosting, since the activation is performed only by lowering the gate voltage from V_{DH} to 0 V. Hence, the PMOS transistor reduces the gate voltage by at least V_{TD}. In actual practice, the difference in gate voltage between the PMOS and NMOS transistors becomes larger, due to variations in the boost ratio at the

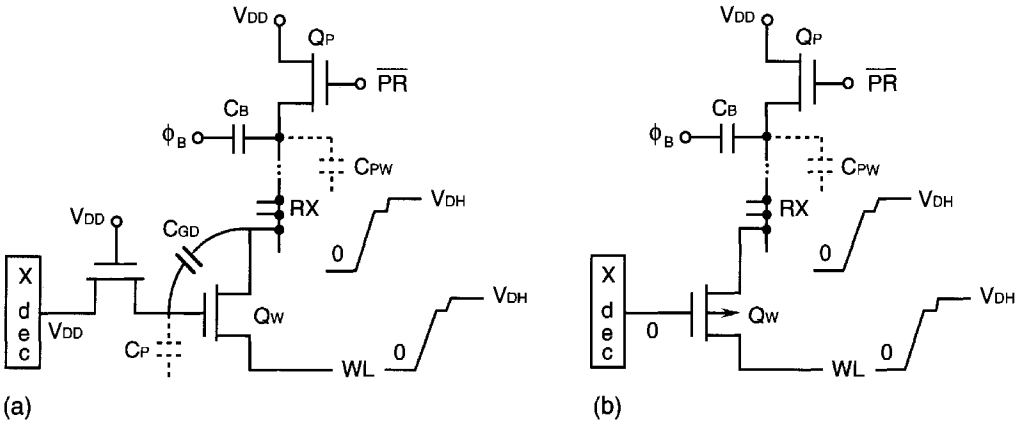


Fig. 3.50. Word boosting for NMOS (**a**) and PMOS (**b**) drivers [3.4].

NMOS transistor caused by variations in the V_{DD} and fabrication-process. The PMOS transistor may allow the gate–source voltage to be large enough to achieve a faster speed despite the lower conductance of the PMOS, compared to the NMOS transistor, in which the gate–source voltage is usually insufficient. Moreover, for lower-V_{DD} operation, the boost ratio must be larger, since there is a minimal threshold voltage for V_{TD} to prevent a degradation of the V_{DH} level caused by a subthreshold current; this is discussed in Chap 8. Thus, the PMOS transistor has been widely used instead of the NMOS transistor.

The Varied Boost-Ratio Driver. There are some operating modes in which MOS devices must not break down even when the operating voltage varies widely. These are the burn-in test mode, which ensures device reliability by applying an increased voltage, and battery operations, which require a wide voltage margin. Figure 3.51a shows a word driver for boosting a well-regulated low-voltage V_{DL}, which is lowered from V_{DD} (= 3.3 V) using an on-chip voltage down-converter [3.27]. The node WDL, which corresponds to WDL in Figs. 3.48 and 3.49, is boosted by activating $\overline{\Phi}_B$ to a low level after it has been precharged to V_{DL}. An excessively high stress voltage is applied to devices in the burn-in test mode, since the boost ratio is almost constant. A voltage down-converter dedicated to the test mode relaxes the stress voltage in a high-voltage region as follows. The reference voltage V_{RN} increases when V_{DD} is increased. Eventually, however, it is clamped at a voltage of two MOS-diode drops and thus V_{DL} becomes equal to $r_1 V_{RN}$. Here, r_1 is a resistance division ratio. Although V_{DL} increases slightly with V_{DD} because of a slight increase in the diode drops caused by the increased diode current, V_{DL} is determined in turn by the other voltage down-converter for the burn-in mode. The converter generates a boosted voltage, monitoring the threshold voltage V_{TM} of the memory-cell transistor as follows. The input voltage V_{RB} of the comparator is given by

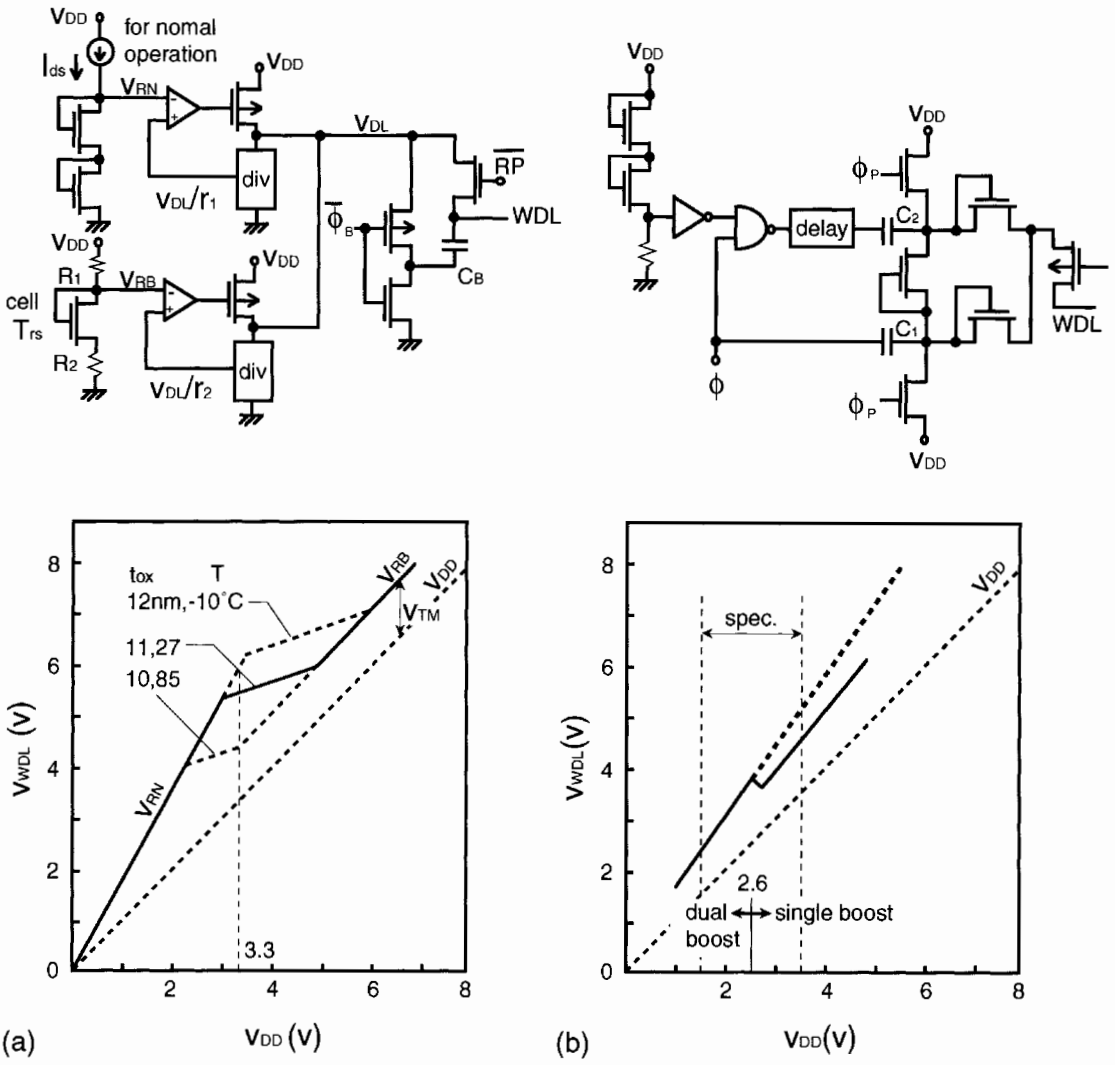


Fig. 3.51. Varied boost ration word drivers [3.4, 3.27, 3.28]. Switching of raised voltages (**a**) and boost number (**b**)

$$V_{\mathrm{RB}} = \frac{R_2}{R_1 + R_2}\left\{V_{\mathrm{DD}} + \frac{R_1 + R_2}{R_2} V_{\mathrm{TM}}\right\} ,$$
$$\therefore\ V_{\mathrm{WDL}} = BV_{\mathrm{DL}} = B(r_2 V_{\mathrm{RB}}) ,$$

where $V_{\mathrm{DD}} \gg V_{\mathrm{TM}}$, r_2 is a resistance division ratio, and B is a boost ratio. Hence,

$$V_{\mathrm{WDL}} = V_{\mathrm{DD}} + r_2 B\, V_{\mathrm{TM}} ,$$

for $Br_2R_2/(R_1 + R_2) = 1$. An additional voltage r_2BV_{TM} is thus V_{TM} for $r_2B = 1$, which minimizes the necessary boosted voltage. The boosted word voltage V_{WDL} can track variations of V_{TM} because a memory-cell transistor is used.

Figure 3.51b shows a voltage up-converter for a battery operation [3.28]. It features dual boosting in the low-V_{DD} region, but single boosting in the

high-V_{DD} region. Thus, a raised S/N ratio even in the low-V_{DD} region, and the suppression of an excessive stress voltage in the high-V_{DD} region are obtained. This is essential to ensure the wide operational range of $V_{DD} = 2.6 \pm 1$ V that is inherent in battery operations. A voltage up-converter [3.29] described in Chap. 5 is a variation of this concept, in which the boost ratio is almost continuously changed over a wide range of V_{DD} values.

The Raised dc Supply-Voltage Driver [3.29–3.32]. A quasi-static supply-voltage (V_{DH}) generator, discussed in Chap. 5, would not only minimize the stress voltage to devices, but also enable fast operation, as follows. Figure 3.52a shows a word driver using the dc power supply V_{DH}. One decoder is shared with four word lines through transistors controlled by the decoded signals Φ_{X0}–Φ_{X3}. At the beginning of activation, the non-selected three of the four signals, which were all at V_{DD}, go down to 0 V while the remaining selected one goes to V_{DD}. After that, the decoder output goes down to 0 V and the selected PMOS word transistor is thus turned on, with a change in the gate voltage from V_{DH} to 0 V, so that the word line is activated at V_{DH}. During this process, the remaining three word lines are latched at 0 V. This driver eliminates the need to drive a heavily capacitive RX line and thus eventually realizes a higher speed, although it needs a little additional time to ensure the sequence of the enabling decoder after enabling Φ_{Xi}. Figure 3.52b shows the other word driver. A decoder and a level shifter are both shared with four word drivers. The decoder function given by Φ_{Xi} in Fig. 3.52a is replaced by a decoded RX scheme, in which each RX driver comprises a level shifter and an inverter (see Fig. 3.54), similar to the word driver. The timing requirement between the application of RX and the enabling of the decoder is relaxed, permitting even an advanced application of RX. Without a boosting effect, which is available at the NMOS word transistor combined with a MOS-

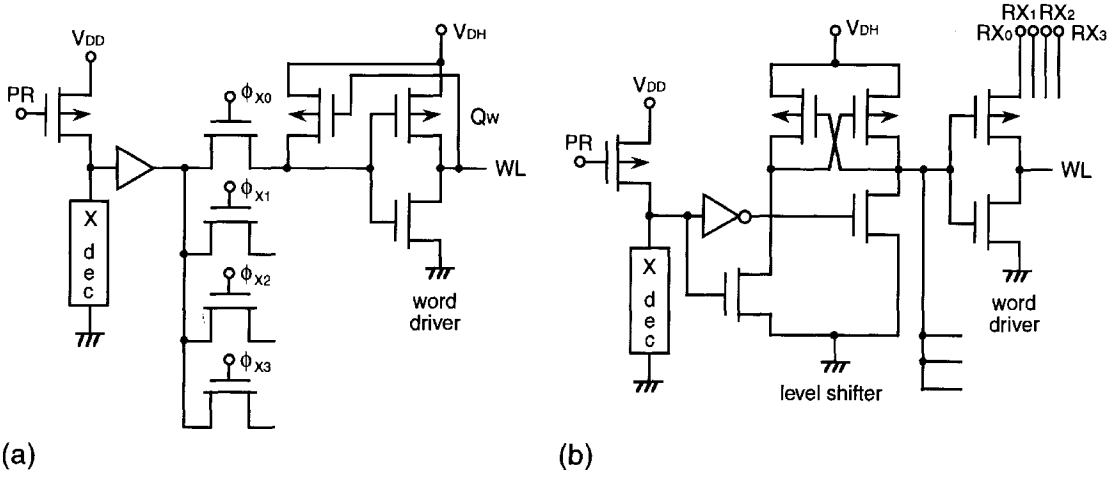


Fig. 3.52. A word driver using a raised dc supply voltage [3.4, 3.29–3.32]. (**a**) Static driver; (**b**) dynamic driver

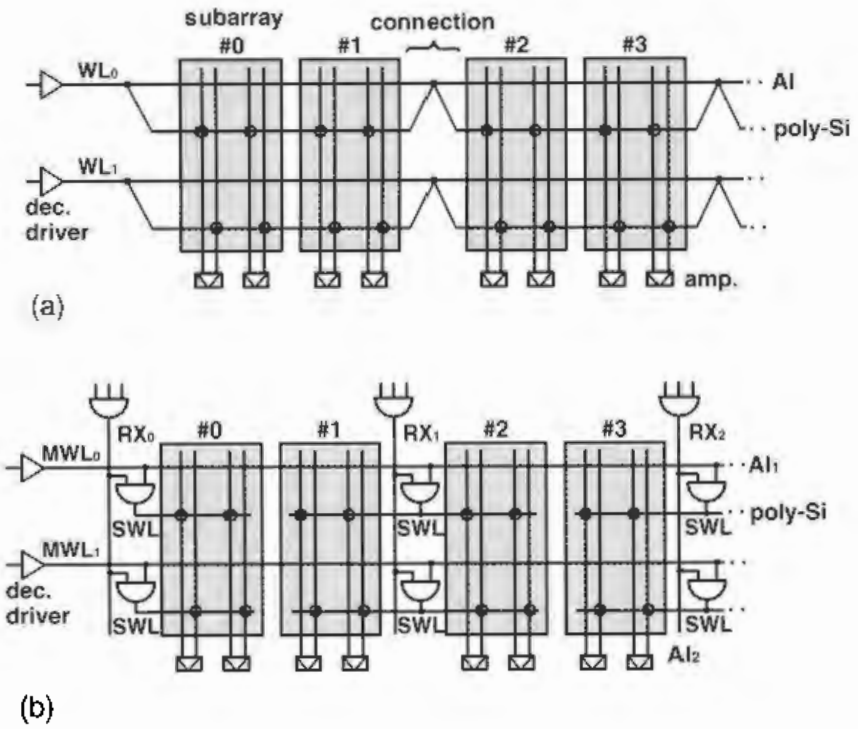


Fig. 3.53. The concept behing partial activation of a multidivided word line [3.18, 3.19, 3.33]. (**a**) No division; (**b**) multidivision

diode shown in Fig. 3.48, an increased word voltage is quickly outputted. This is due to the use of a V_{DH} level shifter.

The Reduction of Word-Line Delay. Even if word-driver speed is improved by using a PMOS transistor, the need to reduce a large word-line delay remains. Although the aluminum-strapped poly-Si or polycide word line, as shown in Fig. 3.53a, has been popular since the 1 Mb generation, a large RC word-line delay has been prominent in the 64 Mb generation and beyond. This is because of the ever-finer and longer aluminum line and the ever-increasing number of memory cells connected to a word line. Unfortunately, an increase in the number of word-line divisions causes an area penalty, with a resulting increased number of word drivers.

One solution is the hierarchical word-line structure [3.18, 3.19, 3.33], as conceptually shown in Fig. 3.53b. This is an example of the delay reduction scheme in Fig. 3.19d. One word line is divided into several by the small subword-line (SWL) drivers. All of the subword lines in a row are commonly controlled by a main word line (MWL). Therefore, they are simultaneously activated by selecting a main word line and all of the row select lines (RX_0, RX_1, etc.). The choice of aluminum lines for MWL and RX would improve the speed of the simple aluminum strapping in Fig. 3.53a, despite poly-Si or polycide subword lines. In simple strapping, a word line is heavily loaded with a huge number of memory cells in a row. The resulting heavy capacitance

develops a long delay even on a low-resistance aluminum line. In this case the total delay is approximately the sum of a large MWL delay and the subword-line delay. On the other hand, in the hierarchical structure a main word line is loaded by only quite a small number of AND logic gates, which is the same as the number of word-line divisions. Considering that the number of memory cells connected to one subword line ranges from 256 to 512, the loading for the main word line is quite light, which enables an extremely small delay compared with the subword-line delay. An RX line also enables a small delay because of the aluminum material. The delay could be further shortened if it is driven by a RX driver located at each data-line division. Eventually, the total delay of the hierarchical structure is almost confined to the subword-line delay. Thus, the delay is less than that for simple strapping.

Figure 3.54 shows an actual hierarchical structure applied to a 256 Mb chip [3.18, 3.19, 3.33]. It relaxes the MWL pitch to one-quarter so that aluminum wiring is enabled even on the top surface of the substrate, and it allows SWL drivers to be placed alternately to meet the tight word-line layout pitch. Eight-row word lines, each of which is divided into a number of subword lines (SWL_{00}, $\overline{SWL_{10}}$, ...), are controlled by a set of complimentary main word lines (MWL, $\overline{MWL}$), which are driven by a row decoder and a set of word drivers. One set of four SWL drivers is located at each end of each subarray. One of four SWL drivers is activated by one of the decoded RX lines from a RX driver. The full use of NMOS transistors realizes a small

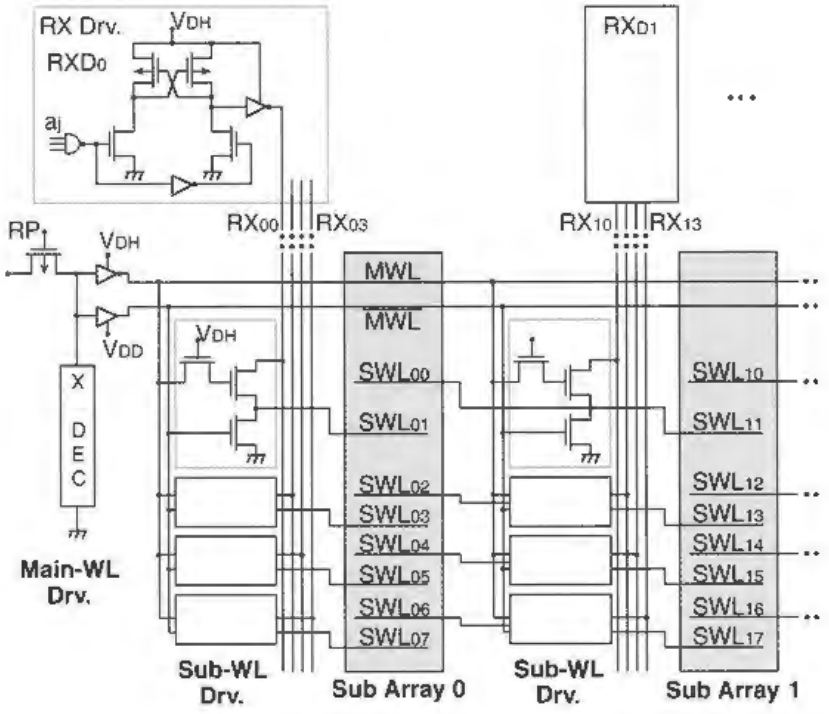


Fig. 3.54. Hierarchical word-line architecture [3.18, 3.19]

SWL driver that drives 512 pairs of data lines in total. Tungsten polycide, first (lower) level aluminum, and second (upper) level aluminum are used for the SWLs, MWLs, and RX lines, respectively.

In this structure, even partial activation of a word line is possible by selecting a main word line and only one of a number of subword lines in a row. Note that the number of cells connected to one SWL in this physical array corresponds to m in the logical array shown in Fig. 3.15. The resultant architecture does not meet the requirements of a traditional address multiplexing scheme, since it increases the number of row address signals or introduces a speed penalty involved in additional selection. However, as far as power reduction is concerned, it has great potential. For example, C_{DT} could be reduced down to about 100 pF at 256 Mb, as shown in Fig. 3.37, assuming a C_D of 200 fF, a ΔV_D of 1 V, and a reduction in the number of activated data lines by a factor of 32, as in [3.19]. Consequently, the architecture almost halves the chip power of partial activation of the data line.

3.6.4 The Sensing Circuit

Voltage margin and speed of DRAM chips are closely related to the sensing operation. However, successful sensing is never realized without a deep understanding of memory-cell structures, circuit configurations, data-line driving schemes, and noise-generation mechanisms in the memory array. In this section, first, basic circuits for sensing and amplification are explained in terms of differential sensing, data-line arrangements, and data-line precharging. Next, sense circuits and high-speed sensing schemes for modern CMOS DRAMs, such as a sense circuit for divided data lines, sense-current distributed arrays, and a direct sensing scheme, are discussed. Other interesting circuits for a higher speed or low-voltage operation are omitted: these include which are a constant data-line charging current scheme [3.13, 3.34, 3.35] with a current mirror, an instantaneously raised gate-voltage scheme with a sense amplifier [3.36, 3.37], and a well-driving scheme of a sense amplifier [3.38].

Basic Circuits for Sensing and Amplification.

Differential Sensing and Amplification [3.4]. A cross-coupled differential amplifier connected to a pair of data lines can amplify a small signal voltage on one data line up to V_{DD}, with reference to a voltage on the other data line. Thus, the read information is discriminated by the polarity of the read signal for the reference voltage (V_r), as shown in Fig. 3.55, while any common-mode voltage coupled to both data lines is canceled. The discrimination stability is closely related to how to arrange the data lines and how to generate the reference voltage.

Data-Line Arrangements [3.4]. Figure 3.56 shows two kinds of data-line arrangement, the open data-line arrangement and the folded data-line arrangement, and examples of corresponding memory-cell structures. The folded data-line arrangement has been unique in terms of low noise and low-power operation, as discussed in Chap. 4.

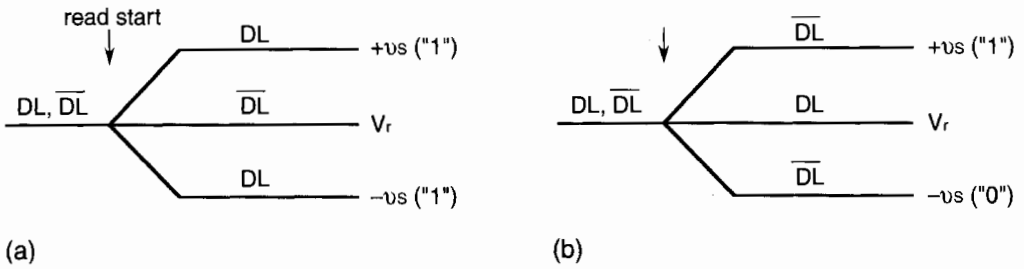


Fig. 3.55. A signal voltage on a pair of data lines [3.4]. (**a**) Signal on data line DL; (**b**) signal on data line $\overline{DL}$

Reference Voltage Generations [3.4]. Figure 3.57 shows reference-voltage generation for the folded data-line arrangement. A simple, small cross-coupled amplifier is used, because it meets a tight data-line pitch and allows the amplified signal voltage to be utilized as a rewrite voltage of the cell. Reference-voltage generation depends on data-line precharging schemes; that is, V_{DD} precharging and half-V_{DD} precharging. V_{DD} precharging, in which the data-line voltage after precharging is V_{DD}, necessitates a special reference-voltage generator (a v_{REF} generator) to generate an intermediate voltage (v_{REF}) on one data line, which is set between the binary information voltages on the other data line. Otherwise, when a memory cell storing a high voltage (V_{DD}) is read, no signal voltage component is developed on the data line because there is no voltage difference between the cell storage node and the data line. Hence, no differential voltage is developed between the pair of data-lines, and the information fails to be discriminated. In any case, the signal voltage is read out to be positive or negative for the data-line reference voltage ($V_{DD} - v_{REF}$), enabling successful sensing. The amplified signal voltage, V_{DD} or 0 V, is utilized for the subsequent rewrite operation. Half-V_{DD} precharging, in which the data-line voltage after precharging is $V_{DD}/2$, does not necessarily need a v_{REF} generator. The reference voltage is $V_{DD}/2$ since a unit of binary information appears as a plus or a minus for the $V_{DD}/2$ level. In this precharging, however, an active restoring circuit is indispensable for raising the data-line high voltage (i.e. $V_{DD}/2$) up to V_{DD} after amplification for cell restoring (rewrite). Half-V_{DD} precharging, which favors a CMOS amplifier, has been widely used, although V_{DD} precharging was popular in the NMOS era up to the 256 Kb generation. Half-V_{DD} precharging is superior in terms of noise, power, and voltage margin if a CMOS amplifier is added, as discussed in Chaps. 4 and 7.

V_{DD} Data-Line Precharge [3.4]. Figure 3.58 shows an NMOS V_{DD} precharging circuit. The reference voltage v_{REF} is generated by a dummy cell which is similar in structure to a memory cell and is activated simultaneously with the memory cell. The capacitor in the dummy cell has half the memory-cell capacitance C_S, and is always precharged to 0 V during the precharge

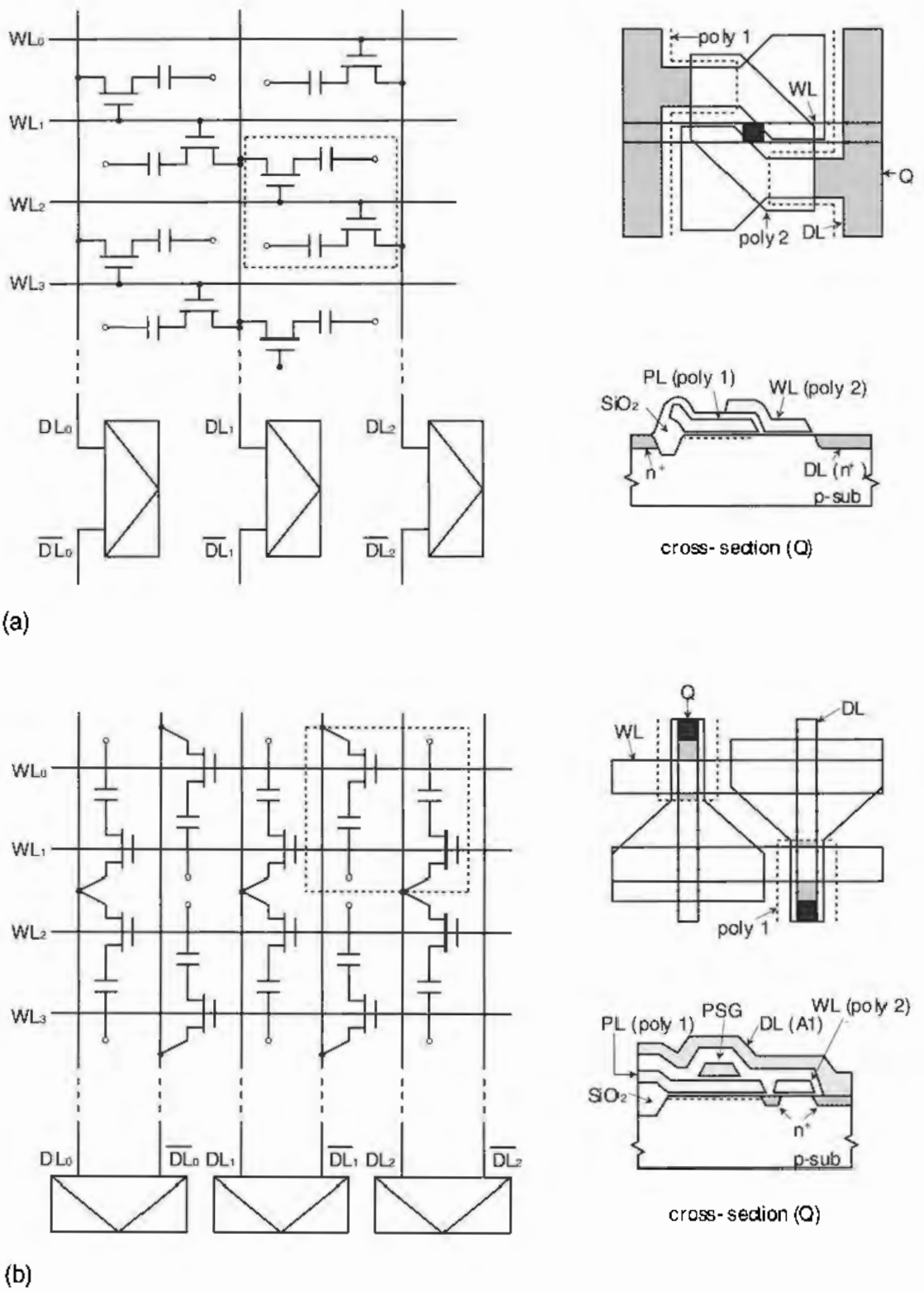


Fig. 3.56. Data-line arrangements and their structures, exemplified by planar capacitances [3.4]. (**a**) Open data-line arrangement; (**b**) folded data-line arrangement

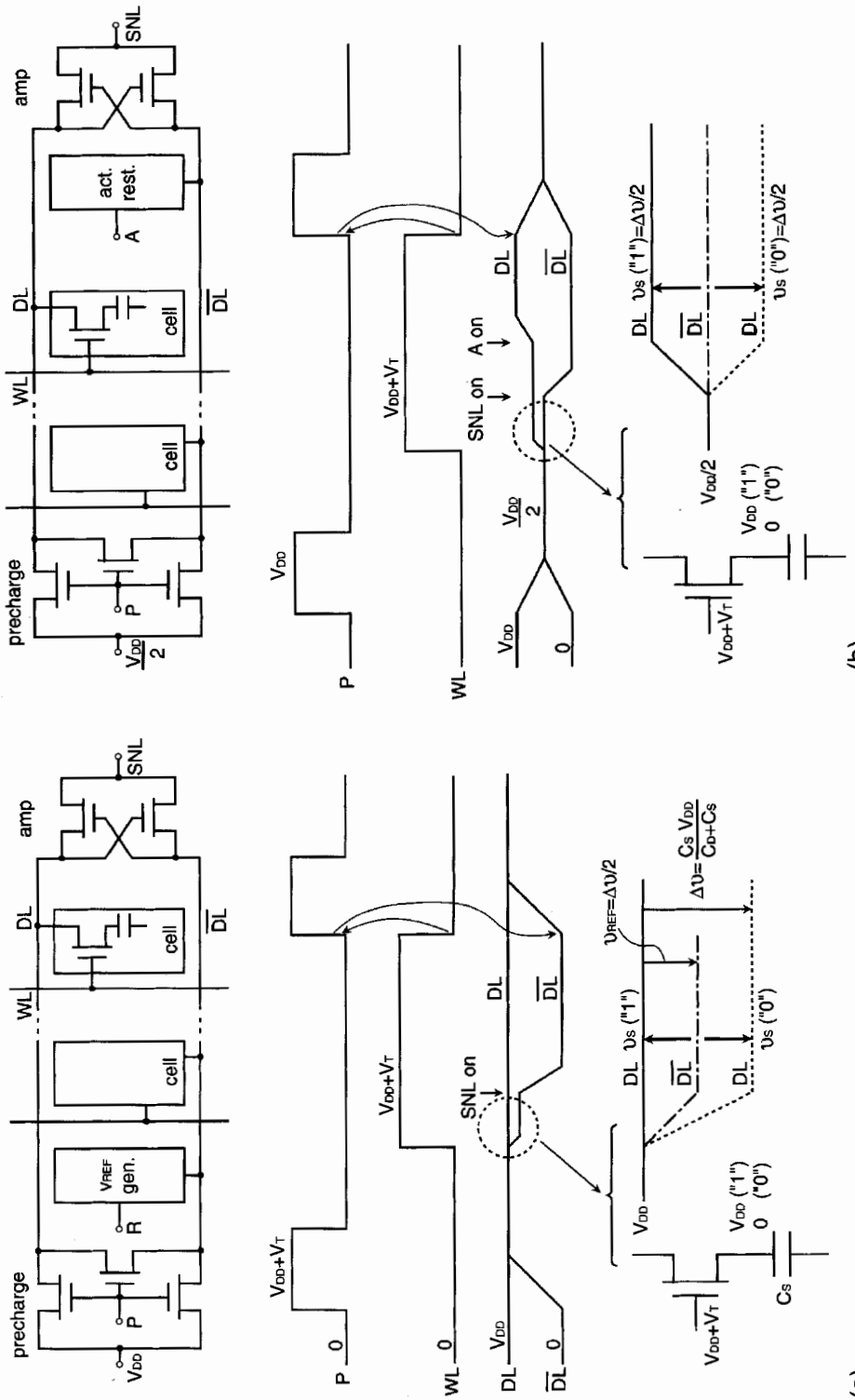


Fig. 3.57. Reference-voltage generation for the folded data-line arrangement [3.4]. (**a**) V_{DD} precharging; (**b**) $V_{DD}/2$ precharging

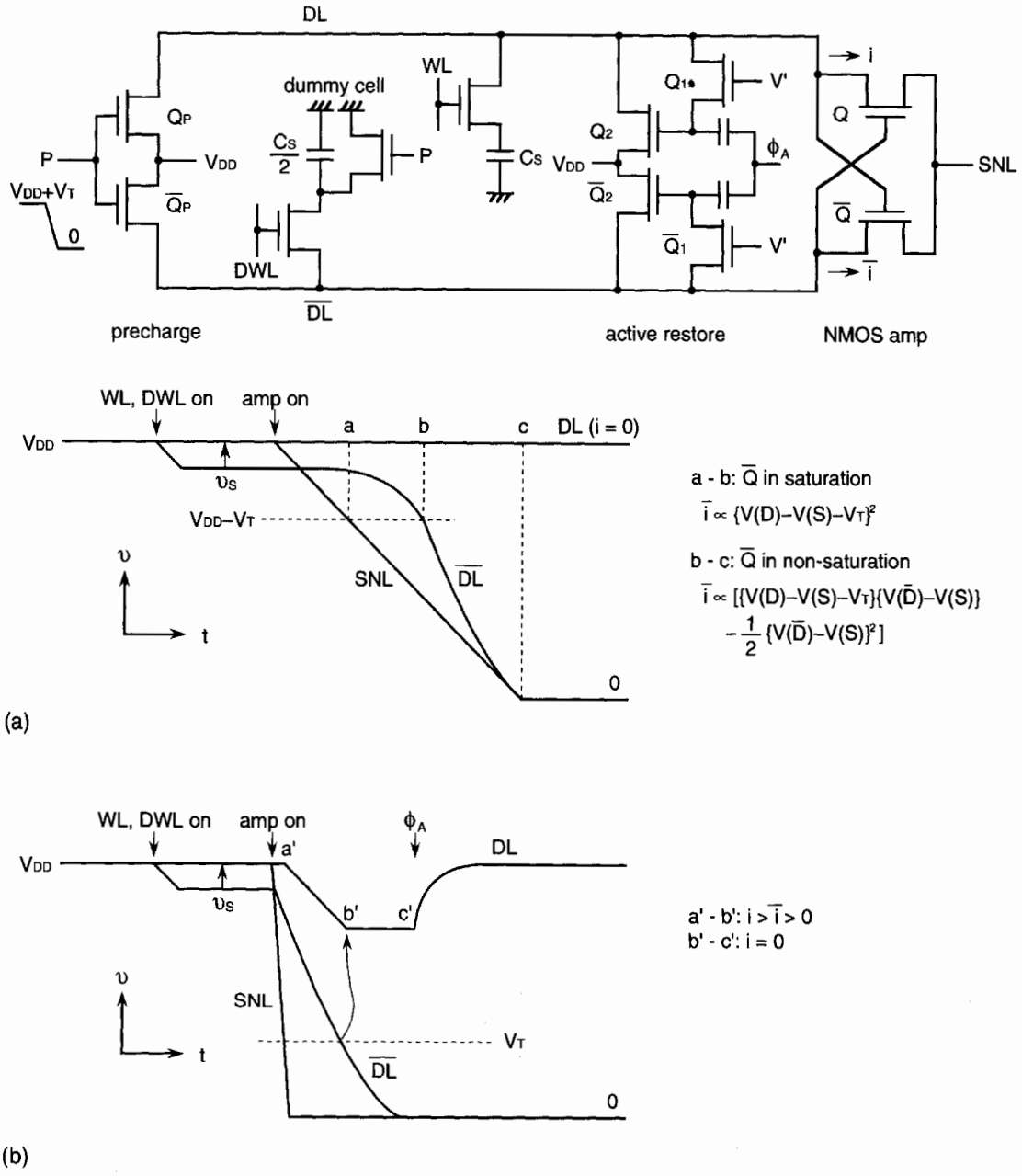


Fig. 3.58. The amplification mechanism for V_{DD} precharging [3.4]. (**a**) Slow amplifier activation; (**b**) fast amplifier activation

period. Therefore, v_{REF} is set at an intermediate level when the dummy cell is activated. One dummy cell is common to all of the memory cells connected to one data line.

The following is a detailed explanation of what happens when a memory cell storing a high-voltage level (V_{DD}) is read. The precharge operation is completed by turning off the precharge pulse applied to P and precharging a pair of data lines to a floating V_{DD}. After that, the simultaneous activation

of a selected word line (WL) and a selected dummy word line (DWL) enables the outputting of no voltage change on one data line (DL) but a dummy-cell signal v_S on the other data line ($\overline{DL}$). Next, the cross-coupled sense amplifier is activated by applying a falling pulse to the common source SNL. Note that an active restoring circuit is not needed if the fall time of the SNL pulse is rather larger than the discharging time of the data line by Q or $\overline{Q}$, as in Fig. 3.58a. Otherwise, it is needed because the high level is degraded, as in Fig. 3.58b. When the SNL pulse falls at an slow enough rate and reaches the point at which the DL–SNL voltage difference is equal to the $\overline{Q}$ threshold voltage V_T, $\overline{Q}$ starts to conduct at first, since the $\overline{Q}$ gate voltage is higher by v_S than the Q gate voltage. Thus, the resulting current $\bar{i}$ starts to discharge $\overline{DL}$. Consequently, Q is turned off more deeply, with its gate–source voltage more negatively biased, because $\overline{DL}$ falls faster than SNL. In other words, Q eventually stays switched off, since the gate voltage drops before it would conduct. Hence, $\overline{Q}$ continues to conduct, with staying Q off, until $\overline{DL}$ is discharged to 0 V, passing the $\overline{Q}$ saturation state (a-b in the figure) and then the non-saturation state (b-c) when the DL–$\overline{DL}$ voltage difference is larger than V_T. Finally, DL is still at V_{DD}, implying no need for an active restoring circuit. Obviously, either of the two data lines can be discharged, depending on the polarity of the DL–$\overline{DL}$ voltage difference, but independently of a common-mode noise coupled to both data lines. Thus, the amplifier surely works as a differential amplifier.

For a sufficiently fast SNL activation, as in the usual applications, both Q and $\overline{Q}$ are turned on and both DL and $\overline{DL}$ are thus discharged, with a sharper slope for the $\overline{DL}$ waveform. When $\overline{DL}$ reaches V_T, Q becomes switched off, and $\overline{DL}$ finally discharges to 0 V. As a result, unlike the previous example, the high level after amplification is degraded, which is enhanced by faster SNL activation. Thus, an active restoring circuit, as shown in the figure, is needed to restore the degraded level to a full V_{DD}. The active restoring circuit needs two transistors and one MOS capacitor per one data line. Just after amplification has finished, the data-line voltages stored temporarily at the Q_2 and $\overline{Q_2}$ gates through the isolation transistors Q_1 and $\overline{Q_1}$ are conditionally boosted by applying Φ_A to the capacitors. For example, if the resultant Q_2 gate voltage is high, the gate voltage is sufficiently boosted by the MOS capacitor for Q_2 to quickly drive DL to V_{DD}. The diode connection of Q_1 isolates the Q_2 gate from DL. However, if the $\overline{Q_2}$ gate voltage is low enough, the MOS capacitor no longer works as a capacitor. In addition, the capacitor can see a large data-line capacitance, through $\overline{Q_1}$ being in conduction. Thus, the $\overline{Q_2}$ gate voltage is never boosted, leaving $\overline{Q_2}$ switched off. Note that an early Φ_A application increases the power dissipation due to a ratio operation by Q_2 and Q.

The Half-V_{DD} Data-Line Precharge [3.4]. Figure 3.59 shows half-V_{DD} precharge using a CMOS cross-coupled amplifier that works as a sense amplifier and an active restoring circuit. Fast SNL activation is shown in the figure.

The CMOS amplifier is regarded as a parallel connection of an NMOS cross-coupled amplifier and a PMOS cross-coupled amplifier. Thus, the NMOS amplifier tends to strongly discharge the data line to a lower voltage, while the PMOS amplifier tends to strongly charge up the data line to a higher voltage. Consequently, a small differential signal voltage is finally amplified to a large differential voltage of V_{DD}. Note that when the PMOS amplifier is activated, $\overline{DL}$ is instantaneously charged up by $\overline{Q_P}$. However, the charging up is suppressed by the ever-rising DL voltage, and thus the increasing $\overline{Q_N}$ transconductance. In practice, both amplifiers are activated, with quite a short time interval at high speed, despite the increased dc current that penetrates both amplifiers. In principle, activation of the PMOS amplifier could precede that of the NMOS amplifier unless the noise generated in an NMOS memory-cell array (see Chap. 4) and the offset voltage of the PMOS amplifier are risks. A dummy cell is usually unnecessary.

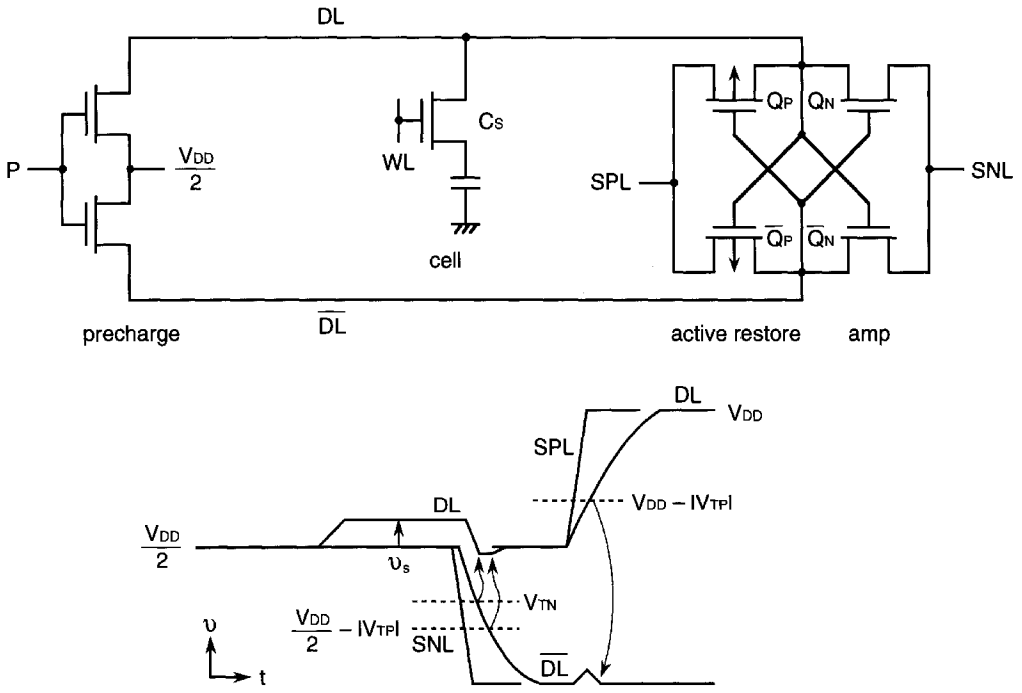


Fig. 3.59. The amplification mechanism for $V_{DD}/2$ precharging [3.4]. V_{TN} and V_{TP} are the threshold voltages of the NMOSFET and the PMOSFET, respectively

The Sense Circuit for Divided Data Lines. Figure 3.60 shows detailed circuits [3.8] of the shared sense amplifier, shared I/O, and shared Y-decoder (Fig. 3.35) for modern CMOS DRAMs. Two of the divided subdata lines ($DL_1/\overline{DL_1}$ and $DL_2/\overline{DL_2}$) are precharged to half-V_{DD} by a shared precharge circuit, with both isolation signals ISO1 and ISO2 kept high. Then, either of the two is selected by ISO1 and ISO2, although the drive circuit for ISO2 is omitted here because it has the same circuit configuration. For example,

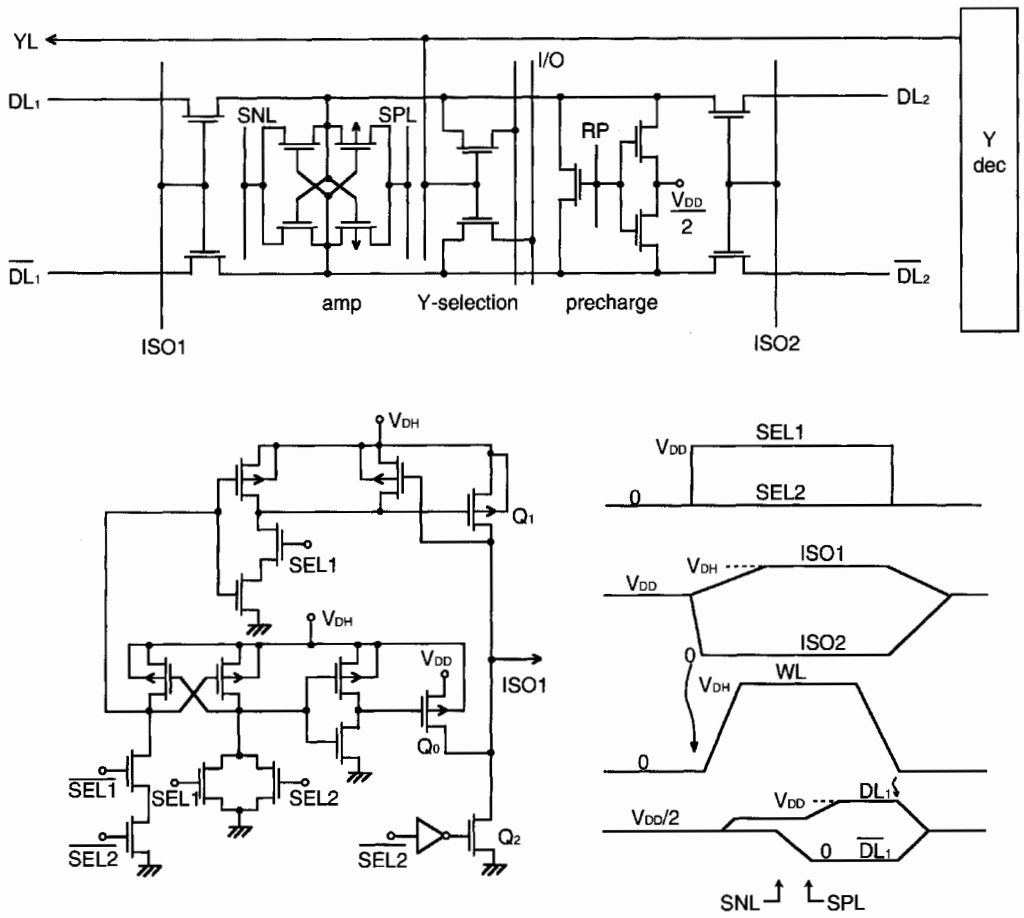


Fig. 3.60. The configuration of a shared sense amplifier, a shared I/O, and a shared Y-decoder [3.4, 3.8]

when a memory cell on the left is activated, the cell signal is inputted to the CMOS sense amplifier with ISO1 kept high but with ISO2 turned off. To restore a full V_{DD} to the cell, ISO1 must be boosted to a V_{DH} higher than $V_{DD} + V_T$ by the time amplification has completed.

The Sense-Current Distributed Array. Since the drive lines of sense amplifiers are common, the sensing speed can be degraded [3.34]. This will be explained by using the jth subarray (Fig. 3.61), which results from dividing each data line in the array shown in Fig. 3.10. As soon as the NMOS sense amplifiers are activated by turning on Q_{SD} so that the common drive (or source) line SNL is driven from $V_{DD}/2$ to 0 V, the discharging currents of a number of data lines rush into the parasitic distributed resistor of an aluminum SNL line. The resultant increased source voltage Δv of the amplifier transistor degrades the amplifier driving capability, preventing high-speed discharging, especially at the end of SNL where Δv is maximized. Δv becomes larger with a larger resistance, caused by a larger memory capacity and higher density. One solution would be to reduce the resistance itself or the number of data

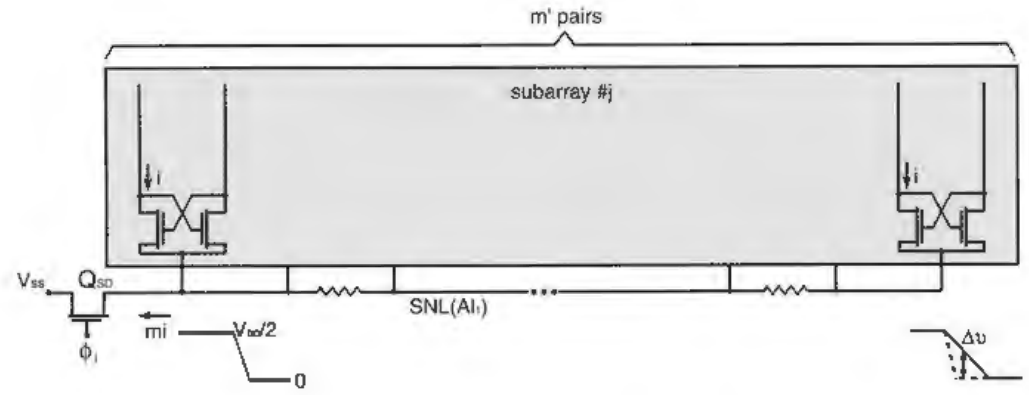


Fig. 3.61. Current accumulation and a deformed voltage waveform on a sense amplifier drive line [3.4]

lines connected to an SNL. However, reduction of the resistance through widening of the SNL line is not possible without an area penalty, since the number of SNL lines is as many as half the number of data-line divisions. An additional data-line capacitance is also entailed. Thus, a reduction of the number of data lines through multidivision of an SNL is a practical solution. In fact, the reduction shortens an access time of about 30 ns by 5 ns for 16 and 64 Mb chips [3.26, 3.39–3.41].

Figure 3.62a realizes fast sensing through distribution of Q_{SD} drivers, a shorter SNL, and an orthogonal layout of the V_{SS} line to SNL, and features meshed V_{SS} lines [3.39]. Each word line is shunted with first-level aluminum (Al_1), as shown in Fig. 3.40b. Each SNL (Al_1) is also divided by utilizing the word-line shunted area, and both ends of the resultant sub-SNL are driven by two Q_{SD}'s (for example, Q_{SD0} and Q_{SD1}). Moreover, to halve the number of current sources driven by the two drivers, the sense amplifiers are arranged alternately. The V_{SS} lines (Al_1) of the SNL drivers are shunted by second-level aluminum (Al_2) lines running over subarrays every eight paired data-lines. The resultant subarray comprises a 32 Kb array of 256 word lines and 128 paired data lines. The discharging current flowing into SNL is distributed through the paths of the V_{SS} line (Al_1) and the orthogonal V_{SS} line (Al_2). Thus, fast sensing is realized. Figure 3.62b shows a V_{SS} line (Al_2) orthogonal to SNL (Al_1) [3.26]. Figure 3.62c shows the V_{SS} line (Al_2) running over a subarray every four paired data lines [3.40, 3.41].

Direct Sensing. DRAM-cell sensing is inherently slow because of a succession of two slow operations; the amplification of a small cell-signal voltage up to a full V_{DD} on the capacitive data line, and transmission of the amplified signal on the I/O line with a parasitic capacitance that is about ten times larger than the data-line capacitance. Direct sensing [3.30, 3.42–3.44] solves this problem, despite an area penalty. The circuit features separated I/O lines; that is, read out paired lines (RO, $\overline{\text{RO}}$) and write-in paired lines (WI, $\overline{\text{WI}}$), as shown in Fig. 3.63. It allows the small cell signal to be directly out-

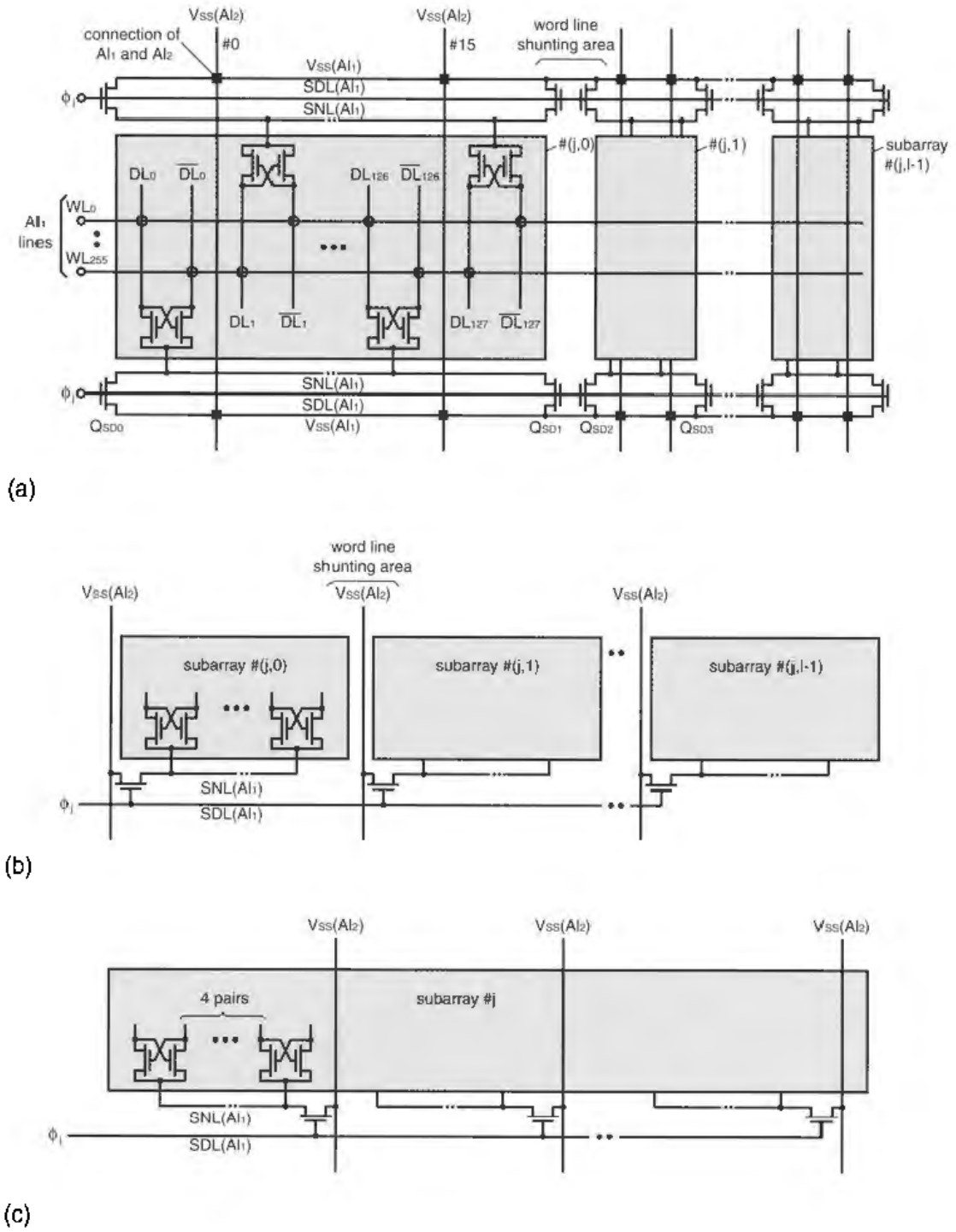


Fig. 3.62. Sense amplifier current distributions with multilevel metal lines [3.4, 3.26, 3.39–3.41]. (**a**) Meshed V_{SS} lines; (**b**) V_{SS} lines orthogonal to amplifier drive line (SNL); (**c**) V_{SS} lines running over subarrays

putted to the RO lines through a buffer without waiting for amplification. In conventional sensing, the timing margins required between a word pulse, the CMOS amplifier activation pulses, and the Y control pulse that connects the data line to the I/O line are the origin of slow sensing. A sense amplifier must be activated after the cell signal is fully developed, which requires a timing margin t_{WS} between the word pulse and the sense-amplifier activation pulses. Furthermore, YL must be activated after the cell signal is amplified, which requires another timing margin t_{SY}. The necessary t_{SY} is closely related to the amplification speed on the data line, and the capacitances and imbalances of the paired data lines and I/O lines. When YL is activated, the low level of the signal waveform in the process of being amplified is instantaneously raised due to charge-sharing between the data line and the heavily capacitive I/O line, which has remained at a high precharge voltage. Thus, the differential voltage (V_{smin}) of the paired data lines is degraded. Obviously, an earlier YL activation causes a smaller V_{smin}. If V_{smin} is larger than the sensitivity of the sense amplifier discussed in Chap. 4, the degraded V_{smin} will be increased again by continuing activation of the sense amplifier. If it is smaller, the succeeding operation will fail. Thus, there is a minimum t_{SY} for successful sensing.

In direct sensing, a small differential cell-signal voltage of about 200 mV is converted into a differential signal current of about 50 μA, superposed on a common-mode current of 1 mA, by a read circuit controlled by a read-selection signal YR. The resulting signal current developed on the RO line is discriminated directly by means of current or voltage sensing. Since the read circuit isolates the data lines from the RO lines, YR can be activated even before word-line activation, allowing both t_{WS} and t_{SY} to be substantially eliminated for sensing. For an actual 1 Mb design [3.44], a t_{WS} of more than 5 ns was needed, and is the same for both sensing schemes. t_{SY} was more than 4.5 ns, assuming imbalances of ±10% in ΔC_D, of ±10% in the transconductance, and of ±30 mV for V_T between the cross-coupled transistors in the sense amplifier. Thus, in conventional sensing, the sum of t_{WS} and t_{SY} was at least 9.5 ns. On the other hand, in direct sensing, the time needed for word-line activation to current sensing on the RO lines was only 1.5 ns. Thus, the access time was shortened by 8 ns, about 30% of the access time. It has been also reported that the chip area increased by 3% for a 64 Mb design due to the need for separate read and write circuits [3.45].

3.6.5 The Common I/O-Line Relevant Circuit

To improve the inherently slow speed of the common I/O line, I/O-line configurations to reduce the parasitic capacitance and resistance, main amplifiers to detect a small voltage swing at a high speed, and a small-signal I/O transmission have been proposed. In this section, high-speed schemes – mainly for the common I/O-line configuration – are discussed.

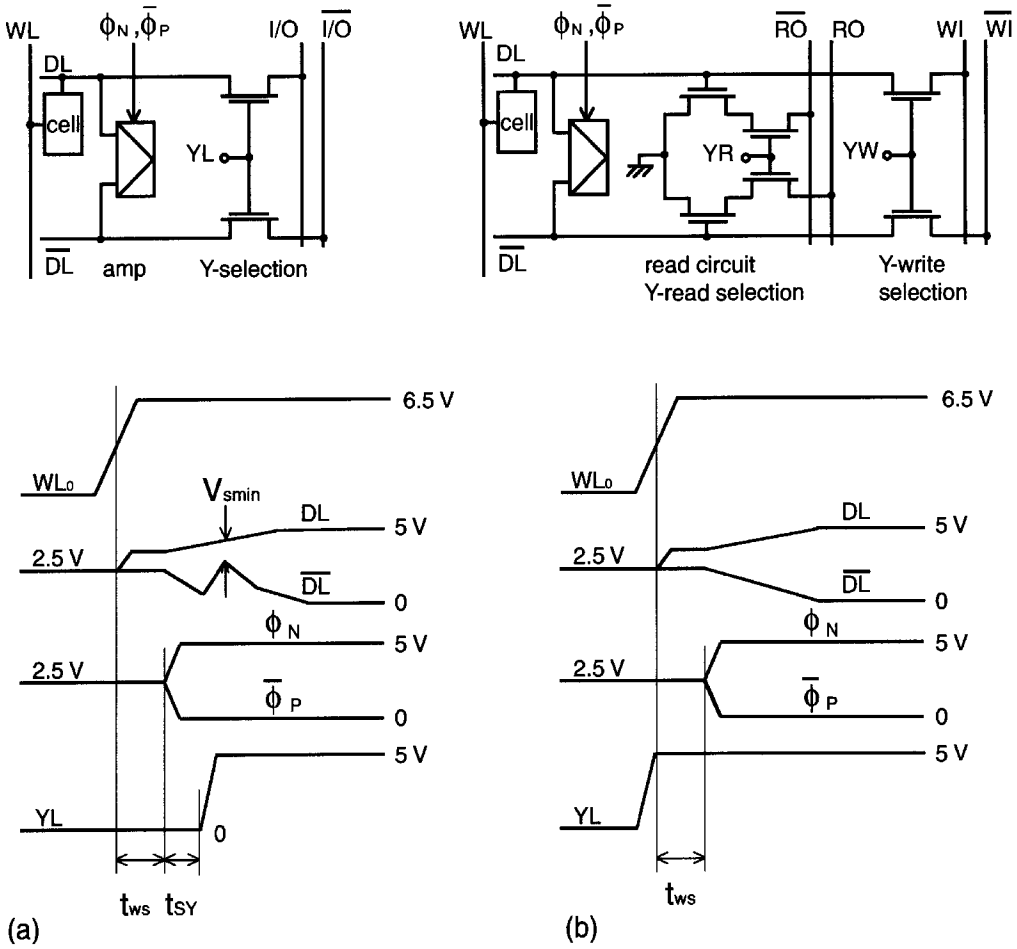


Fig. 3.63. Direct sensing compared with conventional sensing [3.4, 3.30, 3.42–3.44]. (**a**) Conventional sensing for common I/O; (**b**) direct sensing for separated I/O

I/O-Line Configurations. The high-speed I/O-line configurations [3.46] shown in Fig. 3.64 have been used since two-level metal layers became available. Figure 3.64a is simplest and is composed of two I/Os: I/O_1 and I/O_2. In general, I/O_1 is heavily capacitive, because it connects many MOS switches; while I/O_2 is not so capacitive, because there are fewer switches if the second-level aluminum layer is used. However, the delay time between sense-amplifier SA and the main amplifier MA is very large, because both I/O capacitances must be driven by only one sense amplifier in the selected subarray (for example, the $\#k - 1$ subarray). Figure 3.64b realizes a higher speed with a reduced load capacitance of the sense amplifier through the help of an additional main amplifier MA_1 to drive the I/O_2 line. Amplifier MA_1 plays the role of a repeater in Fig. 3.19e. Figure 3.64c features a hierarchical I/O-line configuration composed of multidivided I/O_1 lines, a number of I/O_2 lines orthogonal to the I/O_1 lines, and an I/O_3 line. The resultant sub-I/O_1 line is connected to an I/O_2 line at each division. The I/O_2 line, as well as YL, run orthogonally over subarrays with the second-level aluminum layer

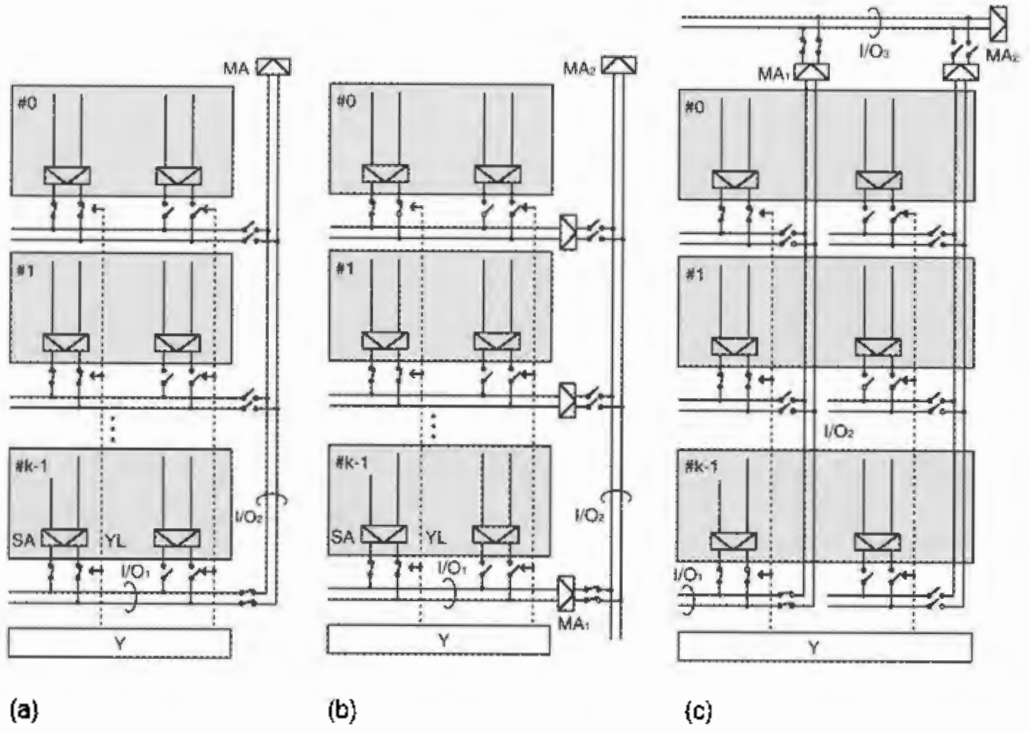


Fig. 3.64. Various I/O configurations [3.4, 3.46]. Y, column decoders and drivers. (a) Two kinds of I/Os; (b) two kinds of I/Os with MA; (c) three kinds of I/Os

(Al_2). Note that each word line is shunted with the first-level aluminum layer (Al_1). Obviously, the sub-I/O capacitance can be reduced without any area penalty, although the number of I/O_2 lines increases with the number of I/O_1 divisions. Thus, a sense amplifier can directly drive both sub-I/O_1 and I/O_2 lines at a high speed. Since MA_1 drives the lightly capacitive I/O_3 line quickly, the total speed is improved. The following are comparisons between the above three configurations, exemplified by a 1 Mb array of 2048 word lines at a 3 µm pitch and 512 pairs of data lines at a 7 µm pitch [3.46]. For configuration (a), the I/O_1 and I/O_2 capacitances are 2.3 pF and 0.9 pF, respectively, while for configuration (c) they are 0.6 pF and 0.9 pF for four I/O_1 divisions. Thus, the total capacitance that a sense amplifier must drive is 3.5 pF for (a), 2.6 pF for (b), and 1.8 pF for (c) as a result of adding a data-line capacitance of 0.3 pF. Therefore, configuration (c) achieves the fastest speed. Here, the chip area of (c) is smaller than that of (b) because the number of MA_1's is eight for (b) and four for (c).

The hierarchical I/O configuration in Fig. 3.64c has become increasingly important with increasing memory capacity, because it provides high-speed and beneficial functions without an area penalty. There has been a proposal for a 64 Mb chip [3.41], in which an I/O_2 line (Al_2) of 9.5 mm length runs along the word-line shunted region. Moreover, a parallel architecture for a 64 Mb chip [3.47], to increase the throughput or to shorten the testing time,

has been proposed. In the architecture, one YL simultaneously selects four pairs for each of the data lines, I/O_1 lines, and I/O_2 lines. If the space caused by a reduced number of YLs (Al_2) is utilized, four pairs of I/O_2 lines can easily be laid out with the second-level aluminum. An amplifier added at each sub-I/O_1 line [3.48, 3.49] further improves the speed.

High-Sensitivity Amplifier. Main amplifiers (MA_1 and MA_2) are categorized into voltage sense amplifiers, which are widely used for the common I/O-line configuration, and current sense amplifiers, which are suitable for the separated I/O-line configuration.

Figure 3.65 shows a typical voltage sense amplifier consisting of a two-stage current-mirror CMOS amplifier [3.21, 3.50], and relevant circuits. They are controlled by pulses ($\overline{EQ}_1$, $\overline{EQ}_2$, $\overline{EQ}_3$) generated from the ATD pulse $\overline{EQ}$ (3.43). It quickly amplifies a small differential voltage on the I/O lines that comes from a ratio operation of transistors in the sense amplifier, Y-switch, and load (Q_L) on the I/O lines. Such a main amplifier, however, is a static circuit, and will always consume a dc current. Thus, the current must be cut off, excluding the period necessary to detect the signal and send the

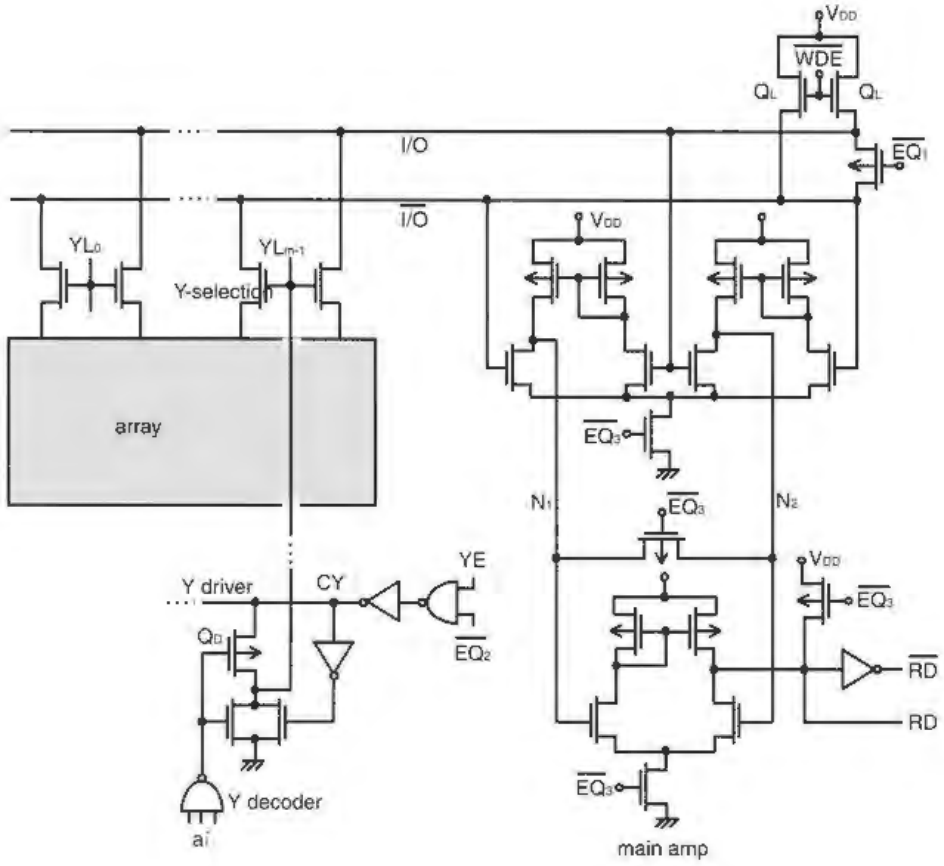


Fig. 3.65. Main amplifier and relevant circuits [3.4, 3.21, 3.50]. YE, acknowledgement signal of completion of cell-signal amplification

detected signal to the succeeding stage. This control is carried out by $\overline{\mathrm{EQ}}_3$. For example, in the static column mode, as soon as the column addresses are switched, the I/O lines are equalized by $\overline{\mathrm{EQ}}_1$, and the output of the selected column decoder becomes low. The $\overline{\mathrm{EQ}}_2$ concurrently becomes high, so that the column driver and Y-switch (YL_{m-1}) are turned on. Every time the column addresses are changed, a small ratio signal is outputted from the corresponding data lines to the I/O lines under the control of these $\overline{\mathrm{EQ}}$ pulses, followed by successful amplification at low power.

Figure 3.66 shows a current-detecting main amplifier. Figure 3.66a is a negative feedback amplifier [3.45], similar to an on-chip voltage down-converter (see Fig. 5.45). A small voltage swing developed on the I/O lines as a result of a current flow is eventually suppressed by turning on the PMOS feedback transistor, with a resultant large voltage swing at the gate, with referrence to the reference voltage V_{REF}. Thus, a small differential current on the I/O lines develops a large differential voltage between the two PMOS gates. For a 1.5 V 64 Mb design with an I/O capacitance of 7.6 pF [3.45], the time from sense-amplifier activation to outputting of a 200 mV differential voltage at the PMOS gates was shortened by 20 ns, compared with that of a voltage main amplifier. Figure 3.66b shows a current-mirror amplifier [3.47]. Each input is composed of a PMOS current mirror, followed by an NMOS current mirror. In this circuit, the currents at the NMOS current mirrors are given by

$$m_1 i - m_1 m_2 i' = m_1(1-m_2)i_0 + m_1(1+m_2)\Delta i \qquad \text{at } A\,,$$
$$m_1 i' - m_1 m_2 i = m_1(1-m_2)i_0 - m_1(1+m_2)\Delta i \qquad \text{at } A'\,,$$

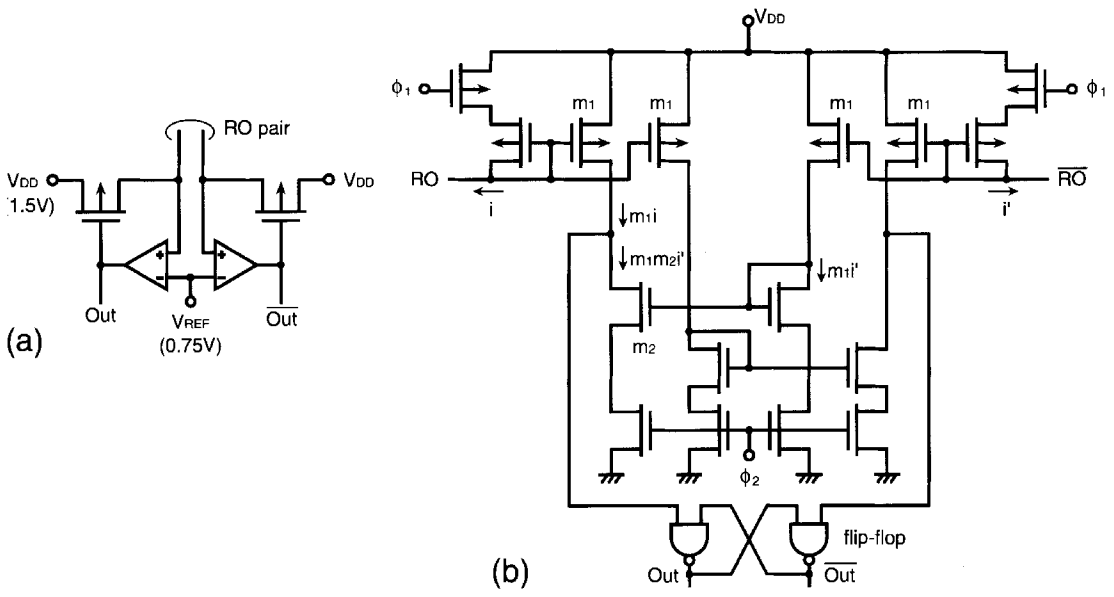


Fig. 3.66. Current-detecting main amplifiers [3.45, 3.47]. (**a**) Negative feedback; (**b**) current mirror

where i and i' are currents at RO and $\overline{\text{RO}}$, m_1 and m_2 are mirror ratios at the PMOS and NMOS current mirrors, respectively, and differential currents of $i = i_0 + \Delta i$ and $i' = i_0 - \Delta i$ are assumed. Thus, one NMOS current mirror charges up one input capacitance of the flip-flop, while the other discharges the other input capacitance. If the RO and $\overline{\text{RO}}$ lines are kept at $V_{\text{DD}} - V_{\text{T}}$ (i.e. V_{T} for PMOS) during the inactive period, the main amplifier does not consume any current, because $i = i' = 0$. The amplifier is cut off by Φ_1 : H and Φ_2 : L during the write operation.

Multistage Small-Signal Transmission. The signal amplified by the main amplifier – for example, MA_1 in Fig. 3.64b – transmits on the I/O_2 line and reaches another main amplifier MA_2. Since MA_2 is usually placed close to the data output buffer at the edge of the chip, a line delay due to the resulting long I/O_2 line needs to be reduced. A solution is to have a small signal transmission. Figure 3.67 shows an example for a 64 Mb chip [3.26]. A small signal of 0.2 V is transmitted on the I/O lines that are as long as over 20 mm in total via three current-mirror amplifiers, MA_1 MA_3, and reaches the data output buffer, where it is converted into a large voltage swing. The scheme cuts the delay by 4 ns, enabling an access time of 33 ns. The amplifier is just a repeater in Fig. 3.19e.

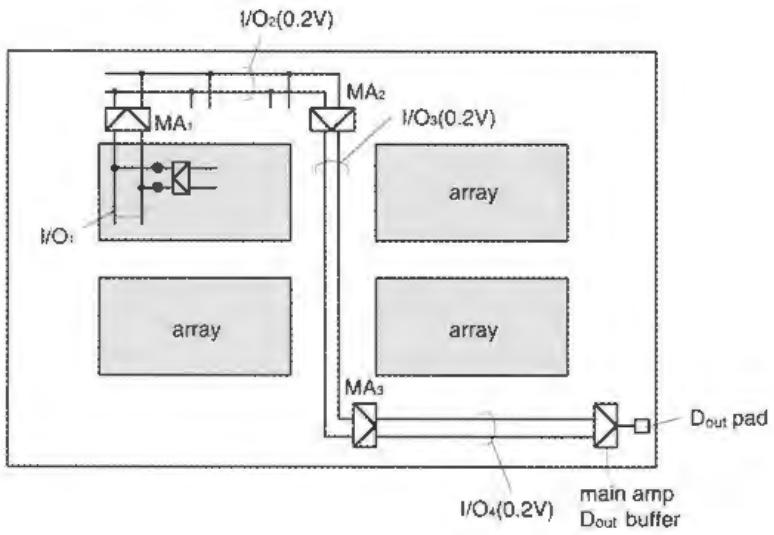


Fig. 3.67. Multistage small signal transmission [3.26]

3.6.6 The Data-Output Buffer

There is an ever-increasing need for high-speed and low-noise data-output (D_{out}) buffers, in accordance with trends in wider I/O-bit configurations and lower-voltage operations. As the number of the buffers, each of which must

quickly drive a large off-chip capacitance of 100 pF in TTL catalog specifications, is increased from 4 to 8, to 16 and then to 32, the total spike current flowing from or to power-supply lines is also increased. The output waveforms would be oscillatory due to parasitic inductances developed between package pins and interconnections to the buffers. Moreover, the spike current could trigger a CMOS latch-up if a PMOS output transistor was used in the buffer. Thus, an NMOS transistor is usually used. At lower V_{DD} , however, even the NMOS degrades speed. Note that while the chip is not selected with $\overline{RAS}$: H the buffer must keep a high impedance (Hi-Z), to avoid interference with another selected chip, which is Wired-ORed with many non-selected chips.

Figure 3.68 shows a D_{out} buffer, using capacitor boosting [3.30]. First, a small transistor, Q_1 or Q_3, depending on the read data RD and $\overline{RD}$, drives the output in order to suppress a large noise caused by a large voltage-changing rate of the D_{out} waveform in the initial stage of D_{out} driving. Next, a large transistor, Q_2 or Q_4 quickly drives the output as soon as D_{out} reaches a certain level. To quickly charge up the output, despite the use of the NMOS Q_2, the Q_2 gate is raised to higher than V_{DD}, using a capacitor. There is another example [3.8, 3.20] that uses an increased dc voltage V_{DH} instead of a capacitor.

Figure 3.69 shows a D_{out} buffer using a raised power-supply voltage V_{DH} [3.20]. Every time a data-strobe pulse TRS from ATD is turned off, a short pulse is generated at N_3 or N_4, depending on the read data RD′ and $\overline{RD'}$. Since the pulse width τ is short enough not to fully charge or discharge D_{out}, a tri-state circuit consisting of Q_3, Q_4, and a delay component can preset D_{out} to an intermediate floating voltage (i.e. Hi-Z). Then, the floating D_{out} is quickly charged up to V_{DD} or discharged to 0 V. An increased Q_1 or Q_2 gate voltage and the charging or discharging from an intermediate voltage are responsible for rapid low-noise operation.

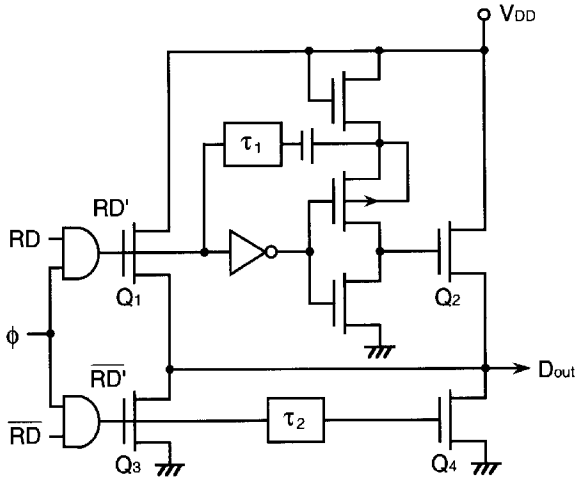


Fig. 3.68. A D_{out} buffer using capacitor boosting [3.30]

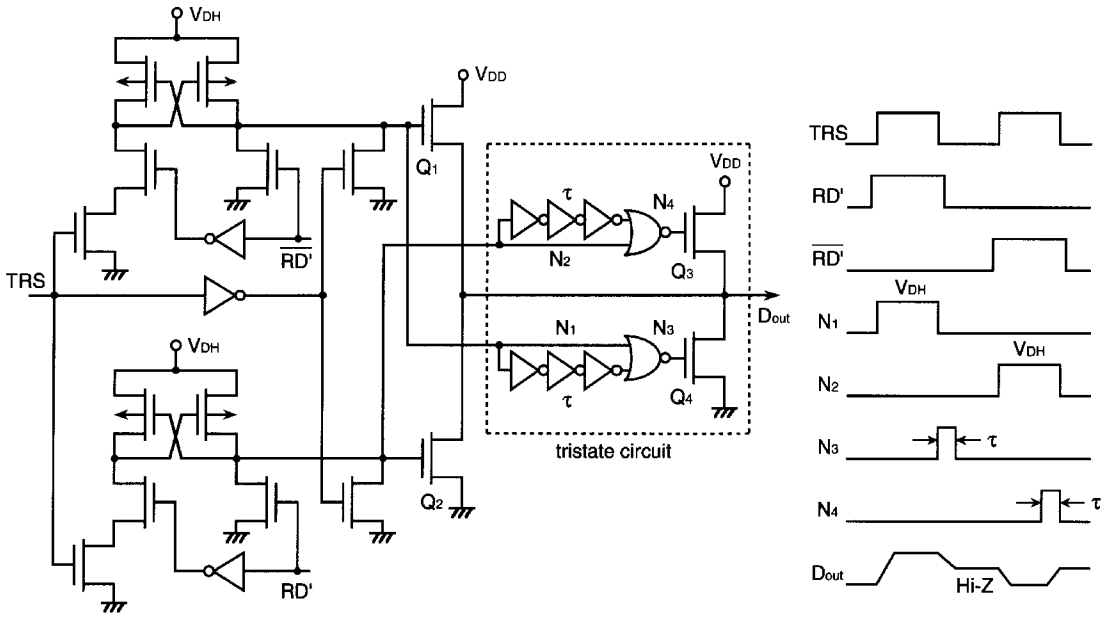


Fig. 3.69. A D_{out} buffer using a raised power supply [3.20]

It has been also reported [3.51] that the rise and the fall times of the D_{out} waveform can be kept constant, independently of the magnitude of the off-chip capacitance, by feedback from the D_{out} waveform.

3.7 Write and Relevant Circuits

In the write operation, differential voltage, corresponding to a input data D_{in} to the selected memory cell, is applied through the write enable ($\overline{WE}$) and data-input buffers, the I/O lines, and the data lines. The operation is carried out while all the memory cells on the selected word line are being read and then amplified. The data input D_{in} is converted to a differential data-input voltage and sent to the I/O lines under the control of D_{in} and the $\overline{WE}$ buffers, as shown in Fig. 3.70. Note that during the write operation, all of the read-relevant circuits, such as the main amplifiers and load MOS transistors on the I/O lines, are cut off, using an internal write-enable signal, to achieve low power. After that, the read data on the selected pair of data lines, which has been amplified to V_{DD}, is replaced by the data-input voltage on the I/O lines with the help of a sense amplifier. The write operation does not entail S/N ratio problems, unlike the read operation, because the relevant signal voltages are always large.

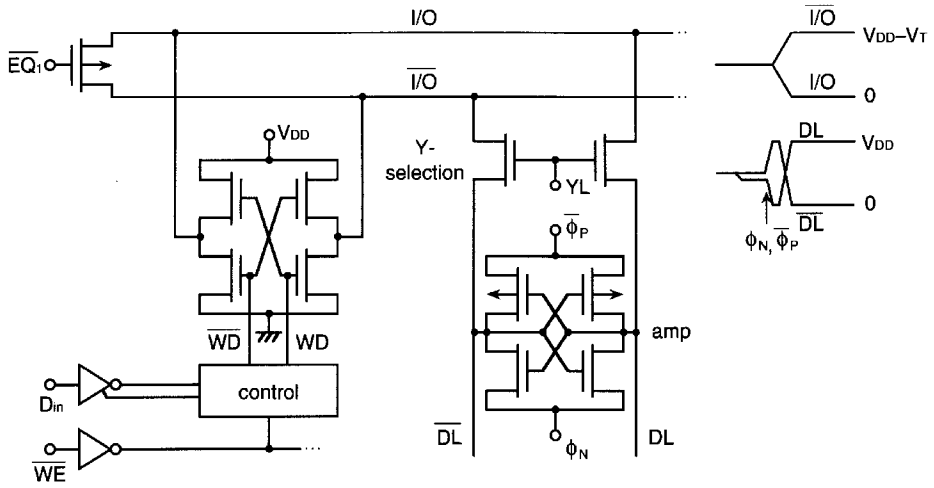


Fig. 3.70. Write-relevant circuits [3.4]

3.8 Refresh-Relevant Circuits

3.8.1 Refresh Schemes

When a refresh cycle starts, all of the memory cells along the selected row (word line) are simultaneously refreshed. By sequential selection of all rows, all of the memory cells on a chip continue to be refreshed. In the case in which a refresh row address for specifying a refresh word line is given by a refresh address counter located outside the chip, we call this a $\overline{\text{RAS}}$-only refresh. In the case in which a refresh row address is given by an on-chip refresh address counter, we call this automatic refresh [3.52]. Under the control of the refresh request signal applied to an additional package pin, a row address is switched to an internal refresh row address. After refreshing one row, the refresh address counter counts up to prepare for the next refresh. A $\overline{\text{CAS}}$-before-$\overline{\text{RAS}}$ (CBR) refresh [3.6] can refresh memory cells by changing the timing sequence between two external clocks $\overline{\text{RAS}}$ and $\overline{\text{CAS}}$ when a refresh cycle starts, instead of adding an external pin dedicated to the refresh. When $\overline{\text{CAS}}$ has been at a low level before $\overline{\text{RAS}}$ becomes low, the subsequent cycle enters into an automatic refresh mode, using an internal refresh address counter.

The above refresh modes need a refresh address to be synchronized with an external clock $\overline{\text{RAS}}$. Self-refresh [3.52, 3.53] generates all the refresh-relevant signals, such as the refresh request and row address signals, on a chip. Under timing conditions of the external signals shown in Fig. 3.8, refresh operations continue without any special external signals. This refresh mode is suitable for battery back-up use. Users normally use the CBR refresh mode in normal operation modes, and switch to self-refresh mode on the occasion of the battery back-up mode.

3.8.2 The Extension of Data-Retention Time in Active Mode

The data retention time of a chip is determined by the worst memory cell with the shortest retention time. Thus, it is essential to reduce the leakage current of the memory cell. Elimination of the worst cells by redundancy and potential profile designs, to release the stress voltage across the p–n junction in the cell, are effective methods, as discussed in Chap. 4. Moreover, reduction of the junction temperature T_j is crucial, since the leakage current is quite sensitive to T_j , as exemplified by about a three-order increase for an increase in T_j of 100 °C [3.54]. Hence, the low-power circuits described in Chaps. 7 and 8, as well as low-thermal-resistance packages, are indispensable.

3.8.3 Current Reduction Circuits in Data-Retention Mode

To reduce stand-by or data-retention power, the current consumed during the refresh operations is a major concern. The current is composed of dc, array, and peripheral currents. The major circuit blocks that consume dc current are the V_{BB} generator and the voltage down-converter. The V_{BB} generator current can be reduced only in data-retention mode, but must be kept sufficiently large in active mode. This is because the substrate current of the chip is drastically reduced in the extremely long cycle in data-retention mode, since the current is roughly proportional to the operating cycle time. This enables reduced drivability – that is, a reduced current – to the V_{BB} generator. An example of conserving current in the V_{BB} generator itself is shown in Fig. 5.25. This generator consists of two V_{BB} generation circuits: one circuit operates continuously at a low current, while the other operates at a high current only in an active cycle or when V_{BB} deviates from a specified value. Another alternative (Fig. 5.26) for saving V_{BB} current is to stop the oscillation of the V_{BB} generator while the DRAM is not in an active cycle. An active and stand-by current reduction in the voltage down-converter (VDC) is also a key to reducing the chip operating current. The use of two VDCs (Fig. 5.71), one for an active node and the other for stand-by, is an efficient way of reducing the current consumed.

An efficient approach to reducing the array and peripheral currents is to make the refresh interval, t_{REF}, longer, as in case ③ in Fig. 3.13. t_{REF} in a standard DRAM is determined so that a stored signal charge has a sufficient margin for charge loss due to leakage current, at the highest possible temperature. This indicates that t_{REF} can be longer at a lower temperature, because the leakage current is then drastically reduced. For example, when the stand-by mode starts, T_j is reduced from the maximum value ($\simeq$ 100 °C), developed at the fastest active cycle time, down to almost an ambient temperature T_a (usually 40 °C max.) of the system, because of quite a small standby-power. Thus, the refresh interval could be extended by about two orders. One method of controlling t_{REF} according to the chip

junction temperature can be achieved by a refresh timer that features a self-refresh control with an on-chip monitor, as shown in Fig. 3.71. A refresh pulse, Φ_{REF}, with a refresh period of t_{REF}, is automatically aligned with the cell leakage current by the monitor cell. Thus, each chip can have its own refresh interval, determined by the monitor. If all of the memory cells in a memory cell array were uniform, a memory cell could be a representative and thus could be used as a monitor cell. In actual practice, however, each memory cell can have different characteristics regarding the amplitude of the leakage current, temperature, and voltages. Thus, to monitor even characteristics of all the memory cells in subsequent volume production, quite a wide refresh-time margin is added to the shortest refresh time of the worst memory cell in a certain chip.

Figure 3.72 shows another refresh timer [3.28]. A set of 1 Kb memory cells, each of which is exactly same in structure as the actual memory cell, is used as a monitor so that the leakage current of the average memory cell can be monitored. The refresh request pulse Φ_{REF} is generated every row-refresh cycle, with 512 cycles at T_1 in total, followed by a pause period of T_2. When the cell-node voltage V_{N} decays down to V_{REF}, Φ_{E} is activated, to trigger a stage in the succeeding timer. After the completion of 512 refresh cycles, Φ_{E} is reset to a low level and Φ_{P} stops precharging the cell node, allowing V_{N} to decay according to the cell leakage current. After time T_2, Φ_{E} is again activated to start the next refresh cycle. Thus, each memory cell is refreshed at a refresh interval of $512T_1 + T_2$ ($\simeq T_2$).

Other generation schemes for the refresh request pulse have been proposed. The programming of pulses by counters that are activated by a ring oscillator [3.29] allows Φ_{REF} to be generated so as to meet the cell leakage current characteristics as closely as possible. Different Φ_{REF}'s tailored to the individual row and subarray [3.55] can effectively extend the refresh interval

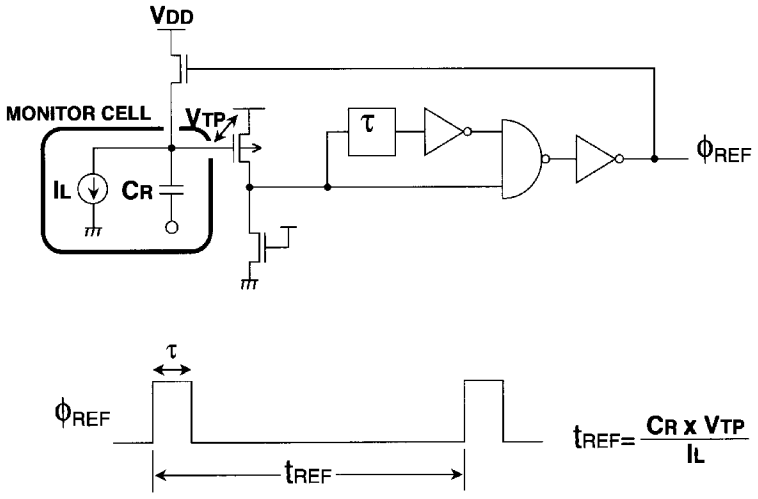


Fig. 3.71. A self-aligned refresh timer [3.54]

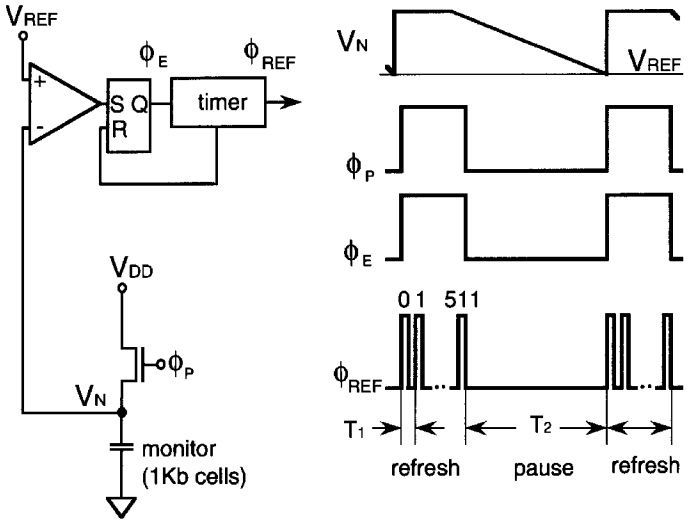


Fig. 3.72. A refresh timer [3.28]

of the chip. This scheme reduces the data-retention current of a 4 Mb DRAM down to 20% with an additional chip area of 4–5% , which is due to a fuse-relevant circuit.

There are different ways of reducing the retention current, by decreasing the peripheral current through array activation strategies. Quadrupling of the number of data lines activated in the data-retention mode, accompanied by reducing the number of refresh cycles by one-quarter instead, is effective in reducing the average current [3.56]. On the other hand, the reduction of the number of activated data lines to one-quarter in the data-retention mode [3.57] has been proposed. The resultings reduced peak current enables the battery life to be extended through flattened current waveforms. A charge recycling scheme [3.58], in which charges resulting from precharging of a certain subarray are used to precharge an adjacent subarray, halves the refresh current.

3.9 Redundancy Techniques

Redundancy techniques have been widely used as effective methods of enhancing the production yield and reducing cost per bit of DRAMs since the 64–256 Kb generations [3.6, 3.59–3.63, 3.82]. The techniques replace defective memory elements (usually word lines and/or data lines) by on-chip spare elements. Figure 3.73 shows the well-known redundancy technique [3.6, 3.63] applied to a DRAM without memory-array division. Redundant data-lines are omitted here for simplicity. The memory has L (here, $L = 4$) spare word-lines SW_0–SW_3 and the same number of address comparators AC_0–AC_3. If the addresses of the defective word lines (WL_0–WL_3) have been programmed into the address comparators during wafer testing, the address comparators

allow one of the spare word lines (SW_0–SW_3) to be activated whenever a set of input address signals (A_0–A_{n-1}) during actual operations matches one of the defective addresses. In modern DRAMs, as many as over one hundred defective elements could be replaced by spare elements in the early stage of production, with an additional chip area of less than 5%. The program elements are usually poly-Si fuses, which are blown by means of a laser beam or a pulsed current, although they do accept memory-cell capacitors [3.64]. Laser programming occupies a smaller chip area and does not normally affect circuit performance, but it does require special test equipment, and increased wafer handling and testing time. Also, the laser spot size and beam-positioning requirements will become more stringent with ever finer line widths. Electrical programming by a pulsed current is carried out using standard test equipment. Usually, a hole is cut in the passivation glass over such fuses to reduce the amount of programming current needed. The possibility of mobile-ion contamination of active circuit areas can be eliminated by using guard-ring structures surrounding the fuse area, or other techniques [3.80]. The area and performance penalties of electrical programming can be minimized by careful circuit design. Electrical programming is used when the number of fuses required is not large enough to offset the negative aspects of laser programming. In any event, laser programming has been widely accepted due to the smaller area and performance penalties, simplicity, assurance of cutting, and the ease of layout of fuses. Many programming methods have been proposed.

Figure 3.74 shows a typical example [3.59] of a spare decoder and driver to replace a defective word line. Laser programming is used. A programmable fuse is equipped for each output of the word drivers (each Q_D), while fuses are used in the spare row (word) NOR decoder. Before programming the spare decoder possesses fuses, so that sets of complementary addresses (i.e. a_i and $\overline{a_i}$, etc.) are input. When a defective word line is detected during wafer testing, the spare decoder is programmed so as to meet the address of the

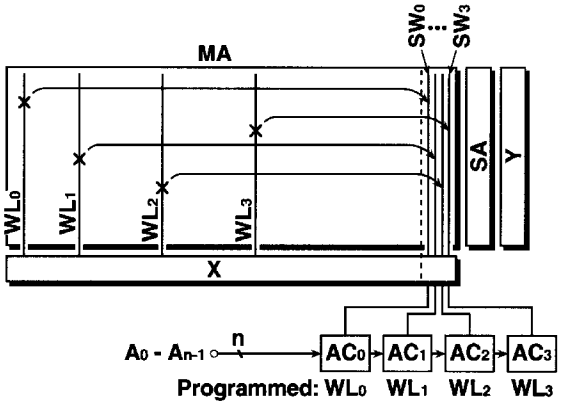


Fig. 3.73. A conventional redundancy technique applied to a DRAM [3.6, 3.62]. X, row decoder; Y, column decoder; SA, sense amplifier; AC, address comparator

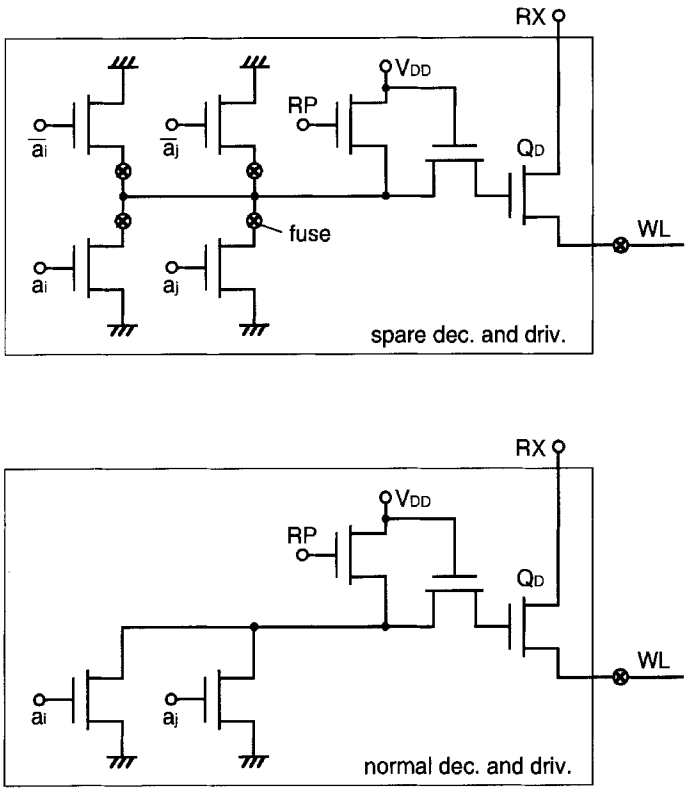


Fig. 3.74. A scheme for replacing a defective word line with a spare word line [3.59]

given defective word line. In addition, the defective word line is isolated from the word driver by cutting the fuse. Thus, even when the input address meets that of the defective word line, and thus the decoder and driver are activated during actual operation, no word pulse is outputted to the defective word line. Instead, the spare decoder, driver, and word line are activated. Obviously, no speed penalty is caused. In the case of no defective word line, all fuses in the spare word line are blown to avoid detrimental effects such as increased power dissipation. The drawbacks, however, are the need for a layout pitch of fuses as fine as that of the word lines, accurate positioning of the laser spot, and large fuse areas. Although a shared decoder scheme (for example, four word lines per decoder) allows the fuse layout pitch to be relaxed to some extent [3.65], the approach eventually becomes a limitation on a higher-density layout.

Figure 3.75 shows an example of disabling the normal decoders by using a disable pulse [3.66]. After a spare decoder that has been programmed in the same manner above is selected (i.e. a high voltage at the Q_D gate), a word activation pulse RX_1 is applied to the spare decoder, so that the resulting pulse Φ_{DA} disables each of the normal decoders by turning on Q_{DA}. After that, a delayed pulse RX_2 is applied so that only the spare word line is activated. Whenever an input address does not meet the programmed address at the spare decoder, the Q_D gate and Φ_{DA} are at low levels and thus only

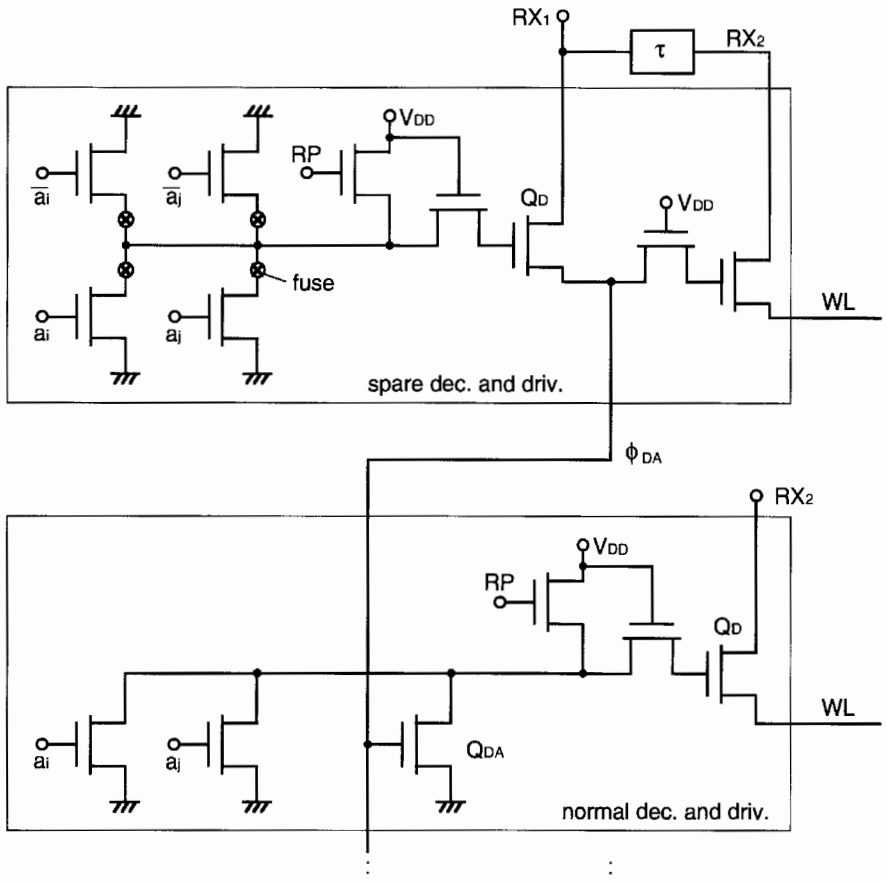


Fig. 3.75. The disabling of normal decoders and drivers by a disabling pulse, Φ_{DA}, which is generated from an address comparator that consists of a spare decoder and driver [3.66]

the normal decoder is activated. This scheme causes a delay, although only a small number of fuses are needed.

Figure 3.76a shows the circuit for two spare rows (word lines) of a 16 Kb NMOS SRAM [3.81]. When the spare rows are to be used, fuses are blown within the spare NOR decoders and each gate of the pulldown transistors (the Ts) is brought to ground. Then the programming elements come into play. Under the control of a fuse, either address true or address complement is transmitted through each programming element. Thus, by blowing the proper fuses, the addresses of faulty rows in the array are programmed into spare NOR decoders. Figure 3.76b gives the basic configuration of a programming element. V_{PP} is a special high-voltage supply used only during programming. It is brought on-chip by an extra pad, which is probed at wafer sort. Later, this pad is grounded by an on-chip transistor, so that no inadvertent programming can take place at the package level. $\overline{P}$ is a logic-level signal that determines which spare row is being programmed. Once a spare row decoder is enabled, it causes the normal element disable signal (NED) to rise as well. NED is inputted to one extra input of every normal decoder, as shown in Fig. 3.75.

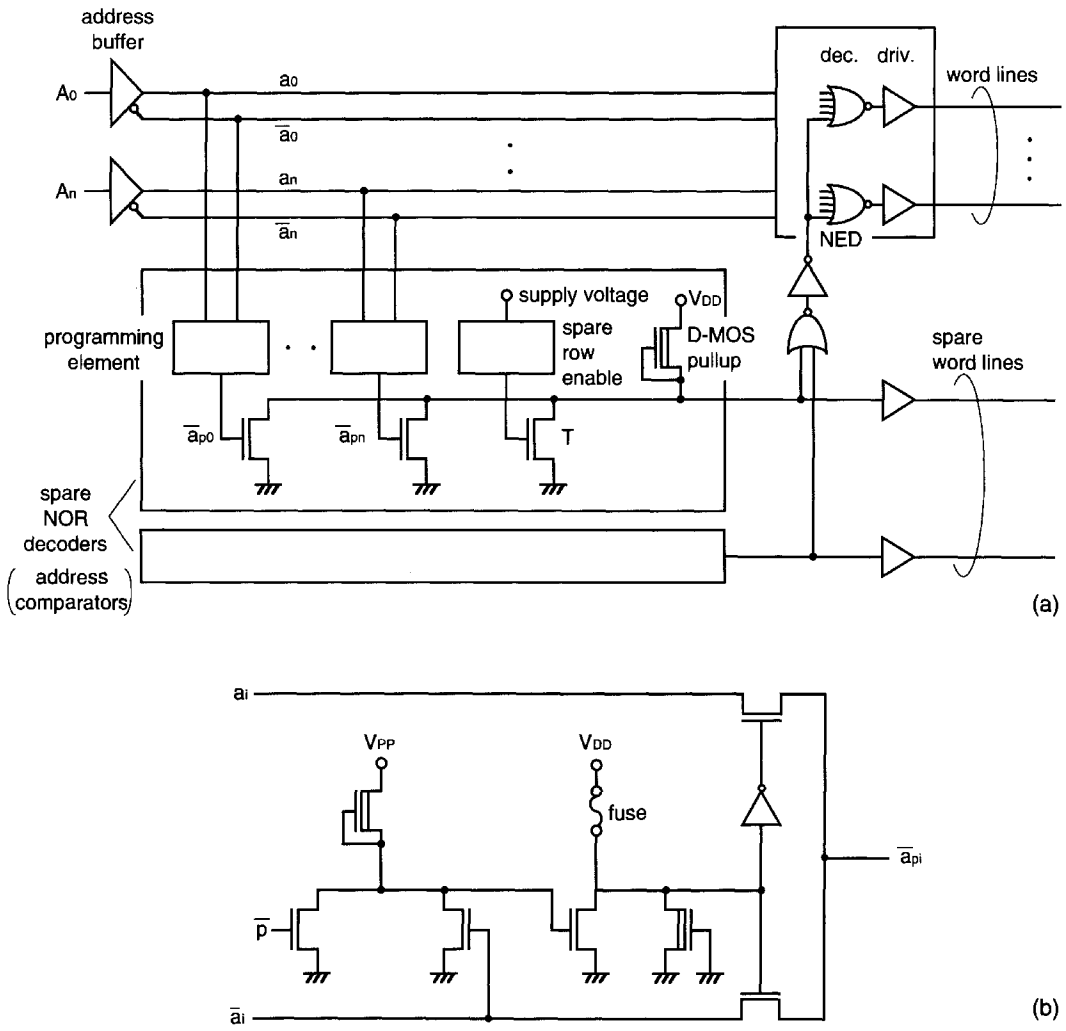


Fig. 3.76. A block diagram for spare rows (**a**) and circuitry for a programming element (**b**) [3.81]

Thus when a spare row decoder is selected, it automatically deselects not only the faulty row line, but also every other normal row of the array. An additional chip area of 6% and an access-time penalty of 7% have been reported.

Figure 3.77 shows a CMOS address comparator [3.82]. A defective row address is programmed by blowing fuses in programming elements. A fuse in a spare row-enable circuit is blown, so that the resulting NED disables the normal CMOS NAND decoders.

The confinement of the number of spare lines [3.67], which increases rapidly with memory capacity, and the replacement of defective subarrays [3.68] that have dc failures, are also indispensable. The error-checking and -correcting (ECC) techniques implemented in 16 Mb DRAMs [3.51, 3.69] avoided not only the above-described hard errors, but also soft errors that oc-

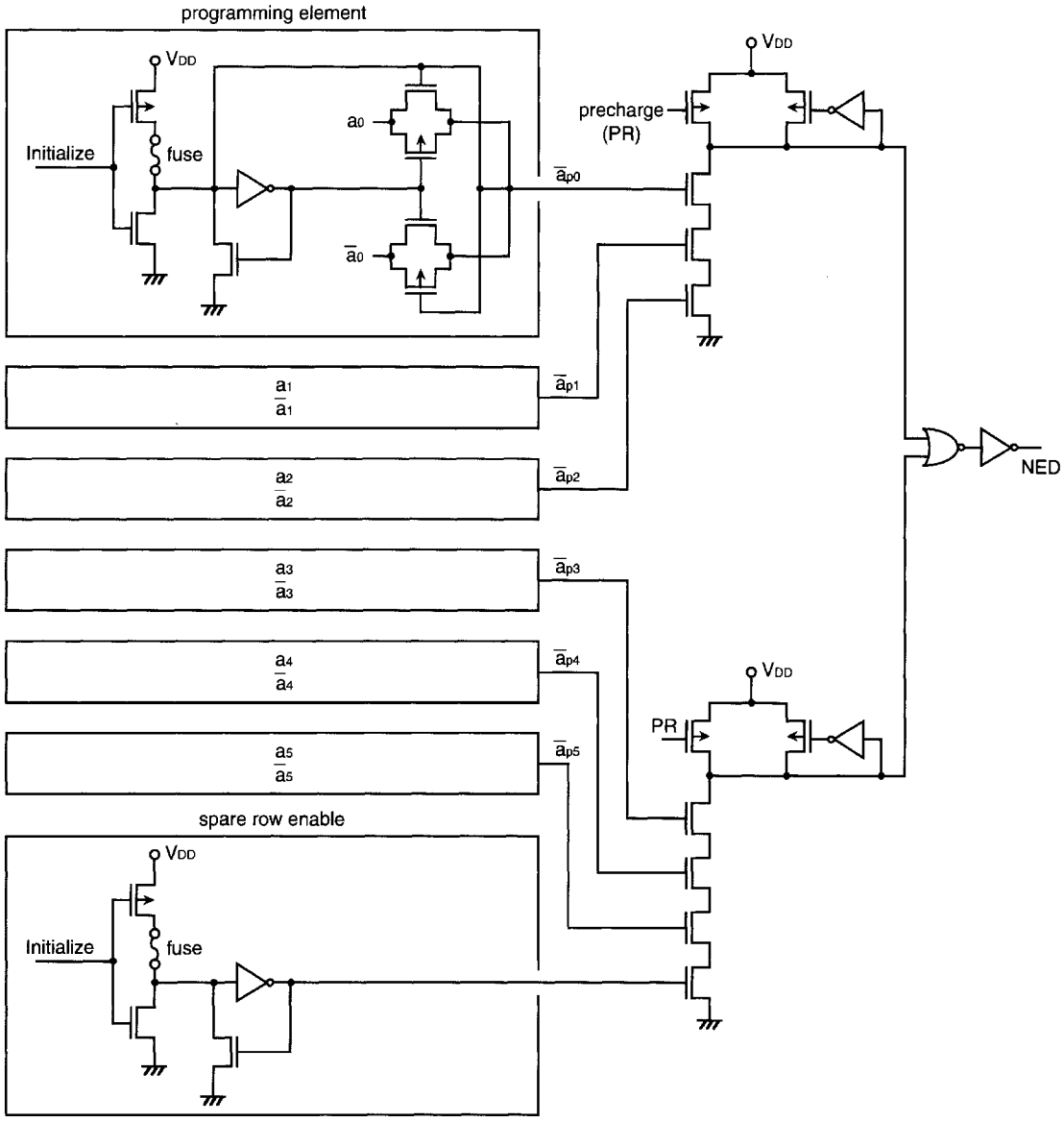


Fig. 3.77. An address comparator [3.82]. A fuse in a spare row enable circuit is blown when replacing a defective address

cur at random due to alpha-particle irradiation. The speed and area penalties caused by ECC are less than 5 ns and 11–12%, respectively.

In the following sections, redundancy techniques for high-density DRAMs will be described in more detail. The techniques are to solve the following two problems that have arisen with the increase in memory capacity: (1) the increase in memory-array division (as discussed earlier in this chapter) reduces the replacement flexibility between defective lines and spare lines; (2) the defects that cause dc-characteristics faults, especially excessive stand-by current faults, cannot be repaired using the conventional redundancy techniques.

3.9.1 Issues for Large-Memory-Capacity Chips [3.63]

As memory capacity has been increased, the following two problems have arisen [3.63]. One is the increase in memory-array division shown in Fig. 3.78. The number of subarrays doubled at each generation prior to 64 Mb. This is mainly due to data-line division for enhancement of the S/N ratio and the reduction in the charging/discharging current [3.70, 3.71]. The number of divisions even quadrupled each generation after the introduction of hierarchical word-line architecture [3.33, 3.72]. The boundaries between subarrays work as barriers to the replacement of defective elements, and reduces replacement flexibility, resulting in yield degradation. The other problem concerns the defects that cause dc-characteristics faults, especially excessive stand-by

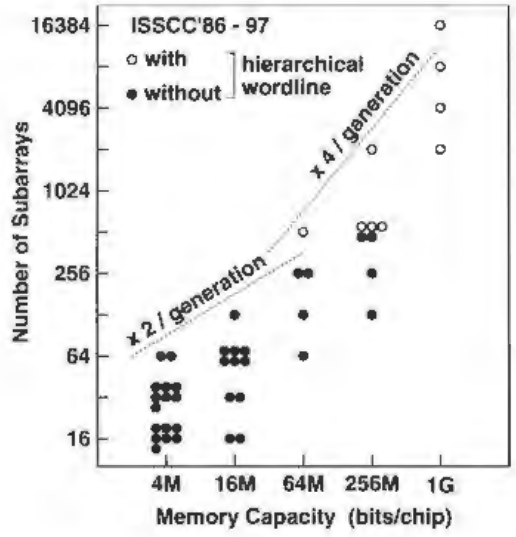


Fig. 3.78. Trends in memory-array division [3.63]

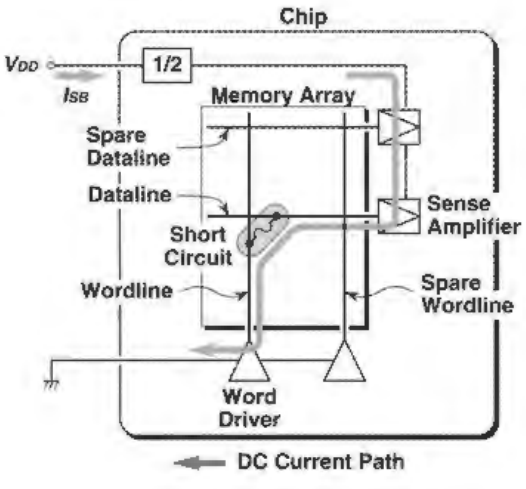


Fig. 3.79. An I_{SB} fault model [3.68]

current (I_{SB}) faults. An example of an I_{SB} fault is shown in Fig. 3.79 [3.68]. A short-circuit between a word line (electrically connected to the ground in the stand-by state) and a data line (connected to the data-line precharge voltage, $V_{DD}/2$) creates an illegal dc current path from $V_{DD}/2$ to ground. Replacing the word line (data line) by a spare word line (spare data line) inhibits the defective line from being accessed, but the current path still remains. Thus, the fault is not repaired by the conventional redundancy technique. A single (I_{SB}) fault can increase the stand-by chip current by as much as 100 mA, causing a yield loss to the I_{DD} stand-by specifications.

3.9.2 Intra-Subarray Replacement Redundancy

Let us now consider dividing the memory array into subarrays. Two approaches within the limits of intra-subarray replacement are shown in Figs. 3.80a, b. Here the memory array MA in Fig. 3.73 is divided into four subarrays, MA_0–MA_3, only one of which is selected.

In simultaneous replacement (Fig. 3.80a), the number of address comparators is equal to L, the number of spare word-lines in a subarray. Each address comparator compares only intra-subarray address signals (here, A_0–A_{n-3}), and the output is commonly supplied to all of the subarrays. The inter-subarray address signals (here, A_{n-2} and A_{n-1}) in turn select one of the four spare word lines. As many defective word lines can be repaired as are shown in Fig. 3.73, if L is the same as in Fig. 3.73. In this approach, four normal lines are simultaneously replaced by spare lines. That is, to replace one defective normal line, three other normal lines with the same intra-subarray address are also replaced, even if they are not defective. This causes the following problems. First, the usage efficiency of spare lines is lower, and the number of spare lines should be larger, which results in an increase in chip-area. Second, the probability of unsuccessful repair due to defects in the spare lines that have replaced normal lines is higher, which results in yield degradation.

In individual replacement (Fig. 3.80b), every spare line in every subarray has its own address comparator. The number of address comparators is therefore $L * M$, where M $(= 4)$ is the number of subarrays. Each address comparator compares both intra- and inter-subarray address signals. This approach has the following advantages over simultaneous replacement. First, a smaller L is statistically required (here, $L = 2$) to repair as many defects. This is because the probability of clustered defects in a particular subarray is small under random defect distribution. Second, since only one normal line at a time is replaced by a spare line, the probability of a defect in the spare line is lower. However, this approach has the disadvantage of lower usage efficiency of address comparators, resulting in an increase in the area of address comparators.

Figure 3.81 shows the flexible intra-subarray replacement scheme [3.67], which has been proposed to overcome the problems described above. The

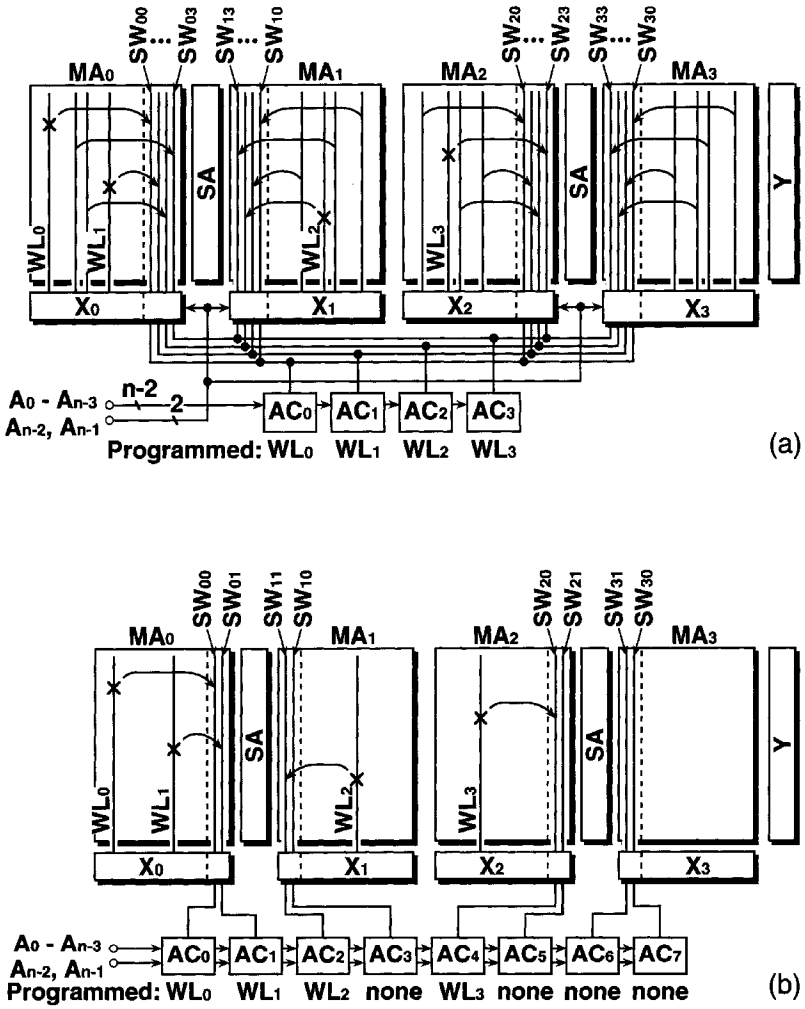


Fig. 3.80. Conventional intra-subarray replacement redundancy techniques applied to a DRAM with memory-array division [3.63]. (**a**) Simultaneous replacement; (**b**) individual replacement

spare lines and address comparators are not connected directly, but through the OR gates G_0 and G_1. Each address comparator compares both intra- and inter-subarray address signals. This connection provides a flexible relationship between spare lines and address comparators. In the architecture shown in Fig. 3.80, this relationship is fixed, so that a spare line can be activated only by a particular address comparator. However, in Fig. 3.81, a spare line can be activated by one of several address comparators. Another advantage of this architecture is that more flexible selection of the number of address comparators C, as well as the relationship $L \leq C \leq L*M/m$ stands, where m is the number of subarrays in which defective normal lines are simultaneously replaced by spare lines.

The calculated yields through the conventional (Fig. 3.80a) and flexible (Fig. 3.81) intra-subarray replacement redundancy techniques are shown in

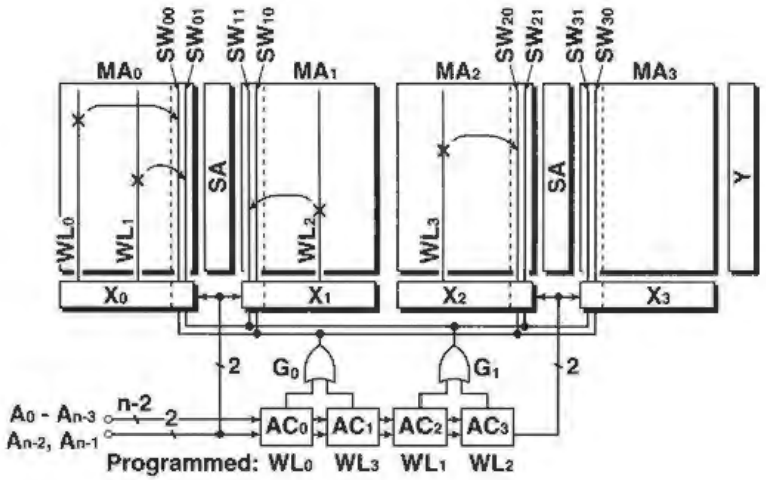


Fig. 3.81. A flexible intra-subarray replacement redundancy technique applied to a DRAM with memory-array division [3.63, 3.67]

Fig. 3.82. In a 4 Mb DRAM, the yield improvement factors through both techniques are almost the same. The advantage of the flexible technique becomes apparent in 64 Mb and 1 Gb DRAMs, especially for a large defect density; that is, in the early stages of production. For a 1 Gb DRAM, however, the yield is determined mainly by fatal defects, such as those that cause an excessive stand-by current.

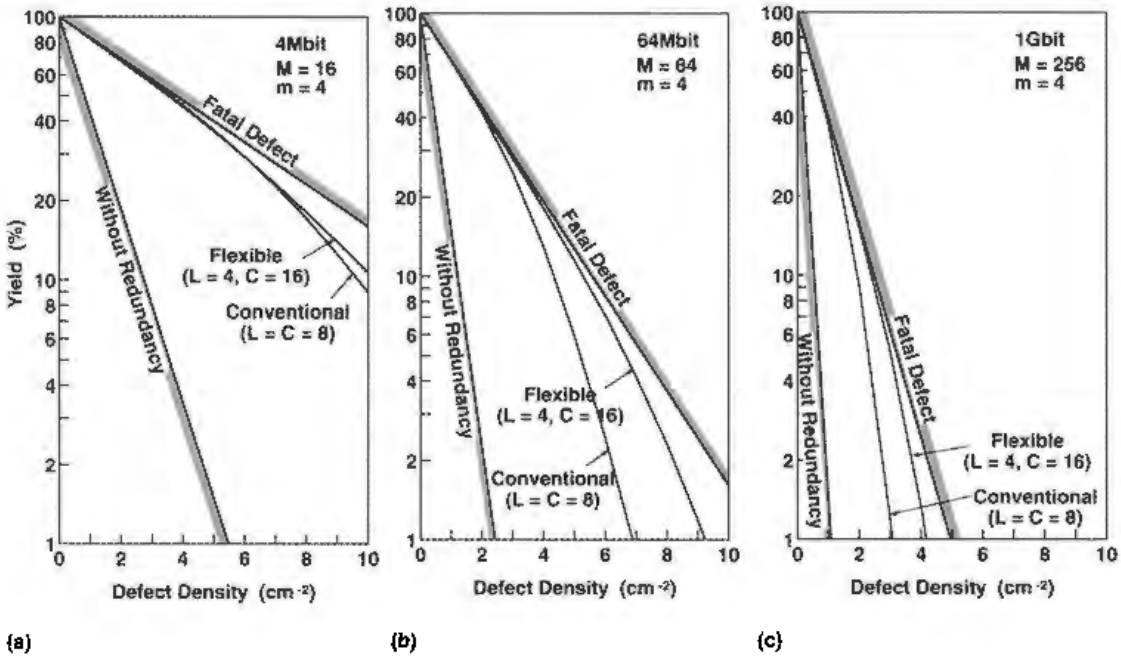


Fig. 3.82. The calculated DRAM yield with conventional and flexible intra-subarray replacement redundancy techniques [3.63]. (**a**) 4 Mb DRAM; (**b**) 64 Mb DRAM; (**c**) 1 Gb DRAM

When the flexible intra-subarray replacement is applied to data-line redundancy, the problem of a "global" defect (a defect over two or more subarrays) arises. A defect on a seuse amplifier or a column selection line (YL) in a DRAM using the multidivided data-line architecture [3.70, 3.71] causes two or more data lines to fail simultaneously, as shown in Fig. 3.83. Thus these types of defects are "global" and require more than one address comparator to be repaired. To solve this problem, the programming of "don't care" values into address comparators has been proposed [3.67].

The access-time penalty due to redundancy is the delay time required for address comparison. Figure 3.84 shows a technique to eliminate this delay time for a high-speed SRAM [3.73]. In this technique, a defective line in a subarray is replaced by a spare line in the adjacent subarray. The two subarrays are activated simultaneously, and one of the data from them is selected according to the result of the address comparison. This technique is difficult to apply to the word-line redundancy of a DRAM, because of the doubling of the data-line charging/discharging current. However, it can be applied to data-line redundancy [3.74]. Note that this technique is not an

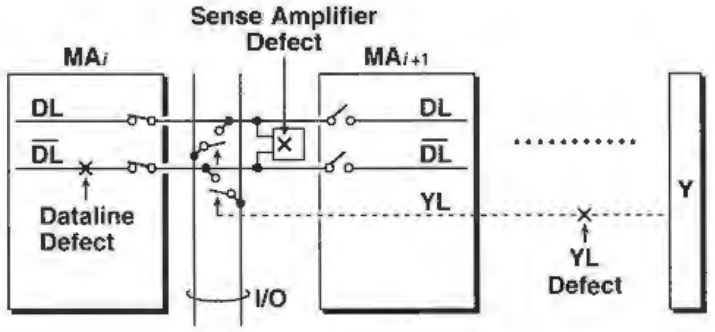


Fig. 3.83. Defect modes in a memory array using multidivided data-line architecture [3.63]

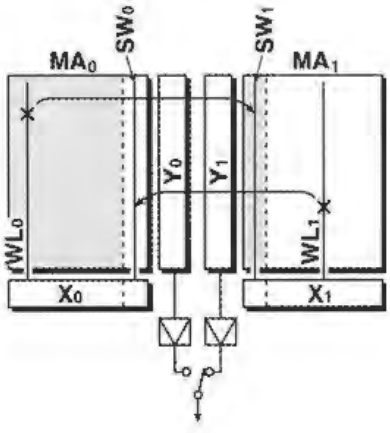


Fig. 3.84. An intra-subarray replacement redundancy technique with no access penalty [3.73] (simultaneous activation of normal and spare lines)

inter-subarray replacement. This will become clear if the hatched areas in Fig. 3.84 are assumed to be one subarray and the white areas are assumed to be another subarray.

3.9.3 Inter-Subarray Replacement Redundancy

With the further increase in memory-array division, the probability of clustered defects in a particular subarray is no longer negligible. In intra-subarray replacement, to repair clustered defects the number of spare lines in a subarray, L, must be larger or equal to the maximum number of defective lines in the subarray. This causes an increase in L and a chip-area penalty. To solve this problem, inter-subarray replacement redundancy techniques [3.75–3.77] have been proposed, which permit a defective line to be replaced by a spare line in any subarray. These are classified into two categories, as shown in Fig. 3.85.

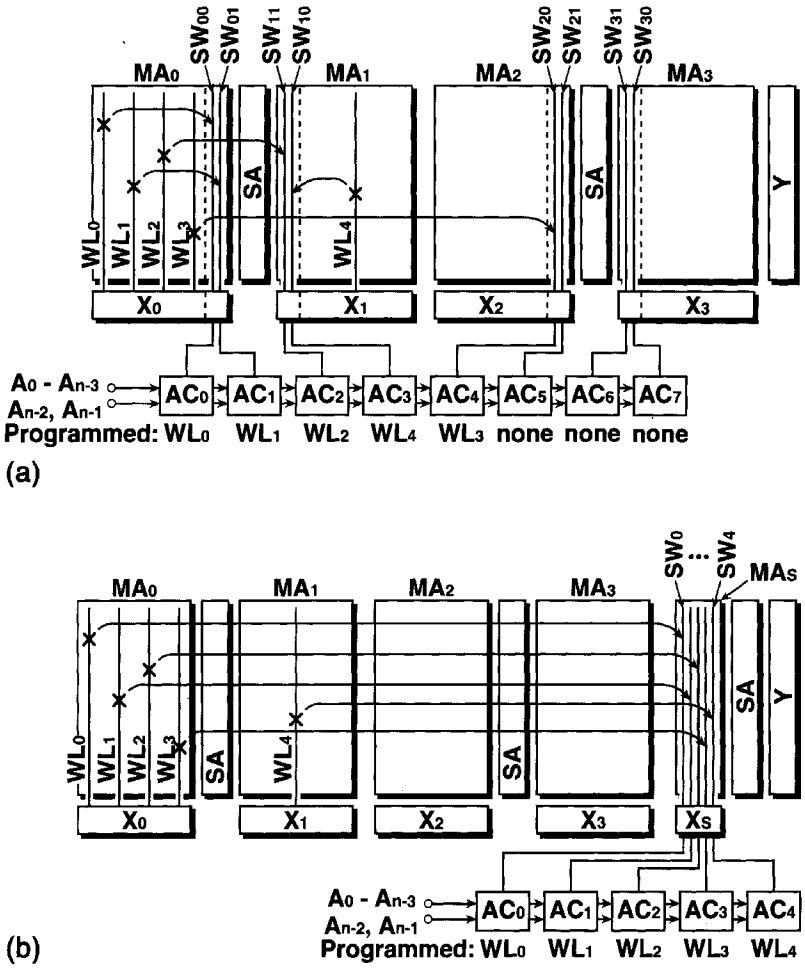


Fig. 3.85. Inter-subarray replacement redundancy techniques [3.63]. (**a**) Distributed spare lines; (**b**) concentrated spare lines

In the distributed-spare-line approach [3.75], shown in Fig. 3.85a, each subarray has its own spare lines, like intra-subarray replacement. However, each spare line can replace any defective normal line not only in the same subarray but also in another subarray. Therefore at most $L * M$ defects clustered in a particular subarray can be repaired, where M is the number of subarrays. In this example, four clustered defective normal word lines WL_0–WL_3 are replaced by the spare word lines in subarrays MA_0, MA_1 and MA_2. It is sufficient for successful repair that the number L is the average number of defective lines in a subarray and is smaller than that for intra-subarray replacement. The number of address comparators C is equal to $L * M$ in this case. However, this number can be reduced through the similar technique shown in Fig. 3.81.

In the concentrated-spare-line approach [3.76, 3.77] shown in Fig. 3.85b, each subarray has no spare lines. There is a spare subarray MA_S instead, composed of L' (here, $L' = 5$) spare lines. Each spare line can replace a defective normal line in any subarray. Therefore at most L' defects clustered in a subarray can be repaired. The number of address comparators C is equal to L'. This approach has an advantage of more flexible selection of L' $(= C)$ and better usage of address comparators compared to the distributed-spare-line approach. This is because the size of the spare subarray need not be the same as that of a normal subarray. The problem with this approach is that additional circuits (a decoder, a sense amplifier, and so on) for MA_S are needed. A solution of this problem using hierarchical data-line architecture is proposed in [3.76].

Figure 3.86 compares the probability of repair using intra- and inter-subarray replacement redundancy techniques [3.75, 3.77]. Here, for simplicity, defects that cause fatal faults and defects on spare lines are neglected. In the intra-subarray replacement, the probability of repair of a memory composed

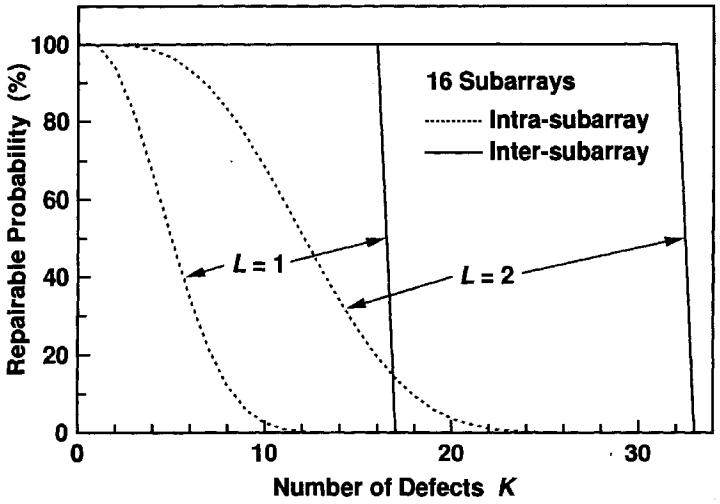


Fig. 3.86. A comparison between intra- and inter-subarray replacement [3.75, 3.77]

of M subarrays decreases with the increase in the number of defects, K, because the probability of excessive ($> L$) defects in a particular subarray increases. On the other hand, the probability of repair is constantly 100% as long as $K \leq L * M$ in inter-subarray replacement. When the number of subarrays is 16, the expectation of repairable defects through inter-subarray replacement is about three times that through intra-subarray replacement.

The access-time penalty of inter-subarray replacement is usually larger than that of intra-subarray replacement. This is because not only an activated line but also an activated subarray may be changed according to the results of address comparison.

3.9.4 The Repair of dc-Characteristics Faults

As described above, the yield of gigabit DRAMs will be mainly determined by defects that cause dc-characteristics faults, especially excessive stand-by current (I_{SB}) faults. Conventional line-replacement redundancy is not sufficient for DRAMs of 256 Mb or larger, due to I_{SB} faults [3.68]. Several redundancy techniques have been proposed to enable the repair of such faults [3.68, 3.77, 3.78].

Figure 3.87 shows two techniques to repair a short-circuit between a word line and a data line. Both techniques modify the data-line precharge circuit. The first approach (Fig. 3.87a) [3.77] limits the illegal dc current through the short-circuit to a small value ($\sim 15\,\mu A$/short circuit) using a current limiter. The I_{SB} of a memory chip with a relatively small number of short-circuits is thereby limited within the specification. The second approach (Fig. 3.87b) [3.78] cuts off the dc current path using a power switch controlled by a fuse. It has been reported that a test to locate short-circuit faults is possible using the switch.

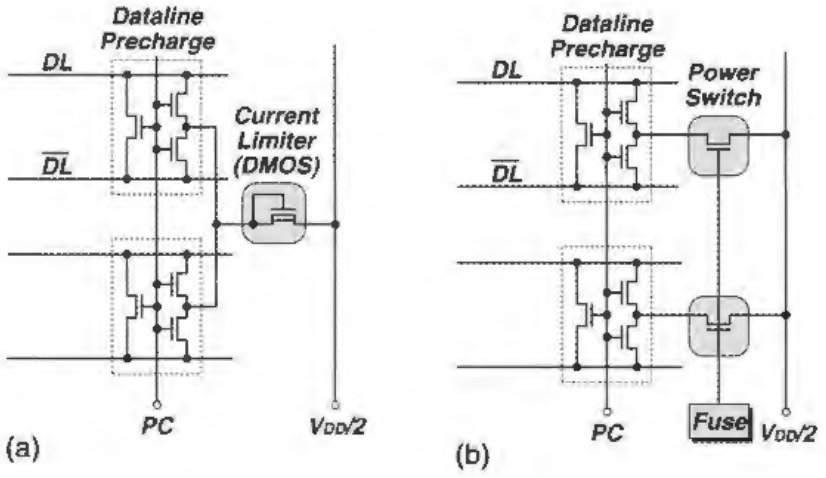


Fig. 3.87. Short-circuit defect repairing schemes [3.77, 3.78]. (**a**) Current limiter; (**b**) power switch

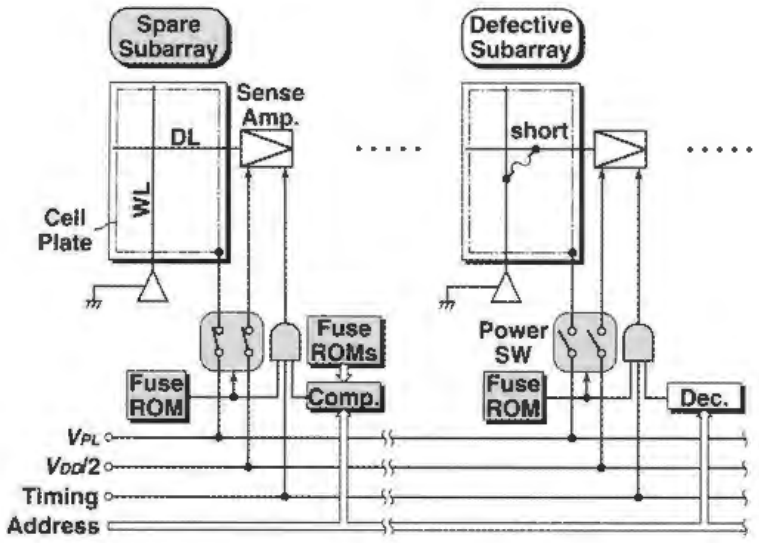


Fig. 3.88. A subarray replacement redundancy techniquew [3.68]

Figure 3.88 shows another technique [3.68] using spare subarrays. A defective subarray that includes an I_{SB} fault is replaced by an on-chip spare subarray. Each subarray has power switches for the data-line precharge voltage $V_{DD}/2$ and the memory-cell plate voltage V_{PL}, logic gates for timing signals, and a fuse to control them. The power switches of the defective subarray are turned off and those of the spare subarray are turned on. Thus the I_{SB} fault is repaired by the cutting of the dc current. The logic gates of the defective subarray are also turned off, to avoid unnecessary power dissipation in the subarray. This technique, combined with conventional line-replacement redundancy, doubles the yield of a 256 Mb DRAM. One advantage of this technique is that the word lines in an unused spare subarray are used as spare word lines of the concentrated-spare-line inter-subarray replacement redundancy described previously [3.79].

Since this redundancy technique requires spare subarrays, it is not suitable for a small-capacity memory with a small number of subarrays. However, the number of subarrays increases with every DRAM generation, as shown in Fig. 3.78. Thus, an area penalty will be allowable for DRAMs of 256 Mb and beyond [3.63]. It is interesting that memory-array division, which was the barrier to line-replacement redundancy, in turn supports subarray-replacement redundancy.

3.10 On-Chip Testing Circuits

The reduction of testing costs is a key to reducing the bit cost. In addition to the ever more expensive memory tester, the ever-longer testing time per chip, with the increasing memory capacity of the chip, makes the bit cost

higher despite parallel testing of a number of chips. Thus, on-chip testing schemes [3.65, 3.83] have been proposed since the 1 Mb generation. Although they cannot perform the perfect testing provided by an expensive tester, they can roughly screen bad chips. After that, the resulting good chips are tested rigidly by the expensive tester. Thus, the utility of the tester is improved.

One way to reduce testing time is parallel testing. There are two major approaches to parallel testing: (1) multibit testing [3.65, 3.66, 3.68, 3.83–3.85], and (2) line-mode testing [3.76, 3.86–3.89], as shown in Fig. 3.89. The multibit testing scheme was proposed during the 1 Mb DRAM generation. In this scheme, a number of memory blocks (q blocks in the example of Fig. 3.89a) are activated simultaneously. By incorporating a number of amplifiers and activating them at the same time, q bits of data can be written or read in parallel. The expected data patterns of the q bits can be modified, although the example shows only the "all 1" pattern. The testing time can be reduced

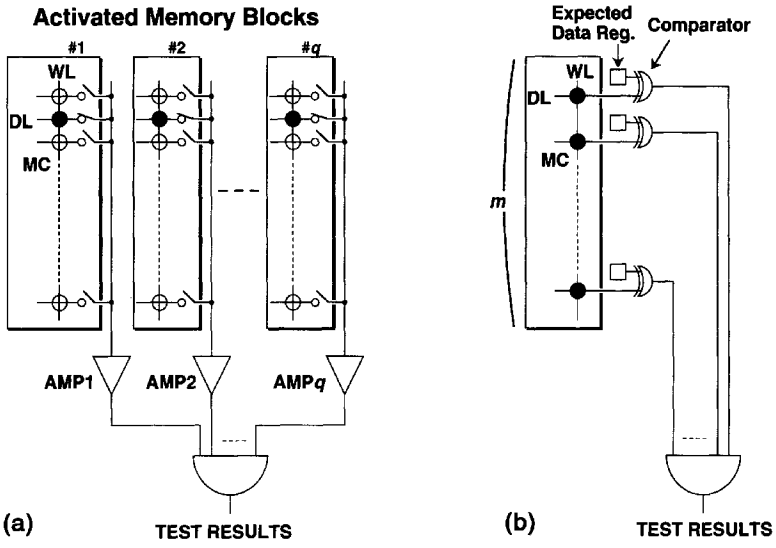


Fig. 3.89. Parallel testing schemes for DRAMs [3.90]. (**a**) Multibit testing; (**b**) line-mode testing

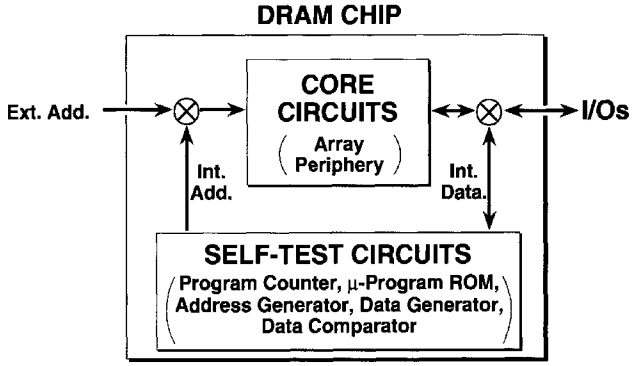


Fig. 3.90. The built-in self-test (BIST) function [3.37, 3.91]

to $1/q$ times that of a conventional read/write operation. A further reduction in the testing time can be achieved using line-mode testing (LMT), as shown in Fig. 3.89b. This concept is to test all memory cells on a selected signal line in one cycle. m bits of data are read and compared with the set of expected data. The testing time with this scheme is proportional to $\sqrt{M}$ for the marching test, and proportional to M for conventional testing. Thus, the testing time is cut drastically. A 16 Mb DRAM with a line-mode test function using multipurpose registers (MPR) to store the expected data has been reported, with a chip area increase of about 0.5%. A built-in self-test (BIST) function [3.37, 3.91], as shown in Fig. 3.90, is also expected to become important in the future. The self-test circuit includes a program counter (PC), a microprogram ROM, an address generator, a data generator, and a data comparator integrated on to the chip. This circuit generates simple test patterns, such as marching and checkerboard patterns.

4. High Signal-to-Noise Ratio DRAM Design and Technology

4.1 Introduction

One of the main contributions to DRAM advances is the one-transistor, one-capacitor (1-T) cell, as explained in Chaps. 1 and 3. The cell has been universally used for over 25 years, because it has the highest density. The drawbacks – no gain and the existence of leakage currents in the cell – have been overcome by successive developments in high signal-to-noise (S/N) ratio designs and technologies. Moreover, the multidivision of data lines by using multilevel metal wiring, explained in Chap. 3 has allowed not only a high-speed and low-power array, but also a high S/N array, while limiting the increase in chip area. Without high S/N designs and technologies, the kilobit and megabit eras would not have been developed at all. In the multigigabit era, however, there are many challenges to realizing high S/N cell design to cope with the ever-decreasing cell area and ultra-low-voltage operations. The development of higher-permittivity materials for capacitor dielectric films, while keeping the fabrication process as simple as possible, and the suppression of the random design-parameter variations of MOSFETs, which are prominent below 0.1 μm, are good examples. However, it will be more difficult than ever to accomplish these things, because of fabrication and physical limits.

This chapter addresses the S/N ratio issue of the 1-T cell. First, Section 4.2 reviews trends in high S/N designs and technologies for the 1-T cell. Then, Section 4.3 investigates in detail the data-line noise issue, which has been and will remain essential for the 1-T DRAM design.

4.2 Trends in High S/N Ratio Design

Memory-cell miaturization and low-voltage operation have to be accompanied by high S/N design [4.1–4.3], because they reduce the cell margin with a reduced signal charge. Before going into detail, the basic operation of a DRAM cell is explained, to clarify the importance of the high-S/N cell design.

Figure 4.1 shows a memory cell that has a capacitance, C_S, a pair of data lines or bit lines (each of which has a parasitic capacitance, C_D), a precharge

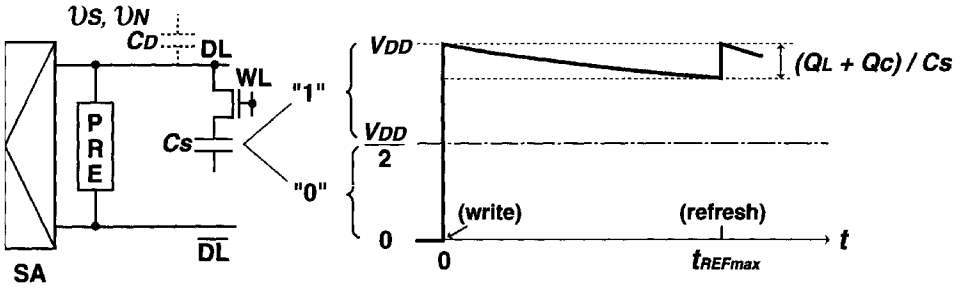


Fig. 4.1. The operation of a DRAM cell

circuit (PRE), and a CMOS latch-type sense amplifier (SA). The available signal voltage in the cell for "1" and "0" is $V_{DD}/2$ for the reference voltage of $V_{DD}/2$. When the cell is read just after a write operation, the signal voltage read out on the data line is a maximum, and is expressed by

$$v_{Smax} = \frac{C_S}{C_D + C_S}\frac{V_{DD}}{2} = \frac{Q_S}{C_D + C_S} \tag{4.1}$$

because of the maximum or fresh stored voltage, V_{DD}. Here full write and read operations are assumed, which are carried out by "word bootstrapping" (word-line voltage $V_{WL} \geq V_{DD} + V_T$) so that the V_T drop at the cell is eliminated. However, the initial high stored voltage, V_{DD}, decays during a data-retention period due to charge losses by leakage currents and α-particle irradiation [4.4]. The resulting degraded voltage must be restored by a refresh operation at the maximum refresh time, t_{REFmax},which is guaranteed in catalog specifications. This is done by reading the cell and restoring the resulting data by using the sense amplifier, so that the cell retains the data for at least t_{REFmax}. Note that the signal voltage just before the refresh operation is a minimum, and is expressed as

$$v_{Smin} = v_{Smax} - \frac{Q_L + Q_C}{C_D + C_S}. \tag{4.2}$$

For a successful refresh operation, v_{Smin} must be larger than the data-line noise, v_N, which is caused by capacitive coupling to the data line from other conductors, the electrical imbalance between a pair of data lines, and the amplifier offset voltage. Thus, $v_{Smin} > v_N$. The formula can be changed with the charge expression if $C_D \gg C_S$ as usual, as

$$Q_S > Q_L + Q_C + Q_N. \tag{4.3}$$

Here, Q_S is signal charge ($= C_S V_{DD}/2$), Q_L is a leakage charge that is the product of the cell leakage current, i_L , and t_{REFmax}, while Q_C is a soft-error critical charge, which is the maximum charge collected at the cell node by α-particle irradiation. Q_N is a data-line noise charge that is the product of v_N and C_D. To cope with cell miniaturization and V_{DD} reduction, v_{Smin} must

be increased. Consequently, it is essential to increase v_{Smax} by maintaining Q_{S} and reducing C_{D}.The reducing of Q_{L}, Q_{C}, and v_{N} is also crucial. In other words, the design and technology must be developed so that the charge relationship in (4.3) is always maintained through high-S/N techniques aimed at a larger Q_{S} and smaller effective noise-charges (Q_{L}, Q_{C}, and Q_{N}).

4.2.1 The Signal Charge

Figure 4.2 shows trends in the signal charge (Q_{S}) [4.16]. Due to various cell innovations, the actual signal-charge reduction from the 64 Kb to 64 Mb generations is less than 1/5, despite a cell area reduction of 1/100. A more detailed description with regard to Q_{S}-maintenance will be given in the following.

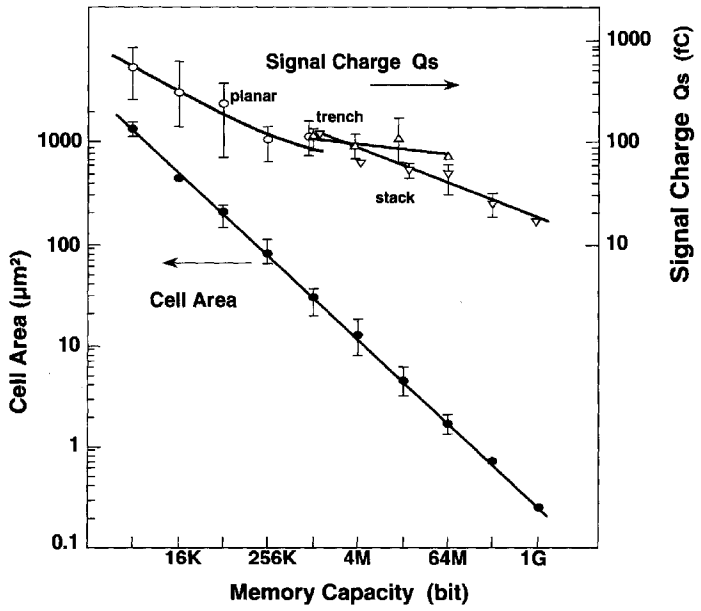


Fig. 4.2. Trends in Q_{S} and the DRAM cell area [4.3, 4.16]

As the memory capacity is quadrupled, maintaining a sufficiently high C_{S} has been indispensable in order to maintain a sufficiently high Q_{S}, despite reductions in the memory cell area and operating voltage [4.2] (Figs. 4.3 and 4.4). Here, C_{S} is defined as $C_{\mathrm{S}} = \varepsilon \cdot A_{\mathrm{S}}/t_{\mathrm{i}}$, where, ε, A_{S}, and t_{i} are the permittivity of the capacitor insulator, the surface area of the capacitor electrode, and the SiO_2 equivalent insulator thickness , respectively. In the kilobit era, from 1970 to 1985, decreasing t_{i} was the only way to cope with the decrease in cell area and maintain C_{S} as much as possible (Fig. 4.4). Note that word-line bootstrapping, explained previously, is a well-known circuit that increases Q_{S} while storing a full V_{DD} level. When approaching 1 Mb, the decrease in t_{i} reached the limit of electric breakdown field strength, making the following innovations necessary. Until the 64 Kb generation, the V_{DD}

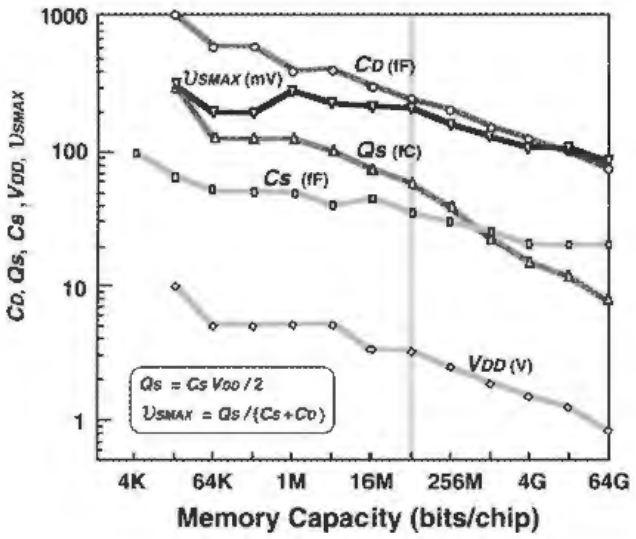


Fig. 4.3. Trends in sensing parameters [4.2]

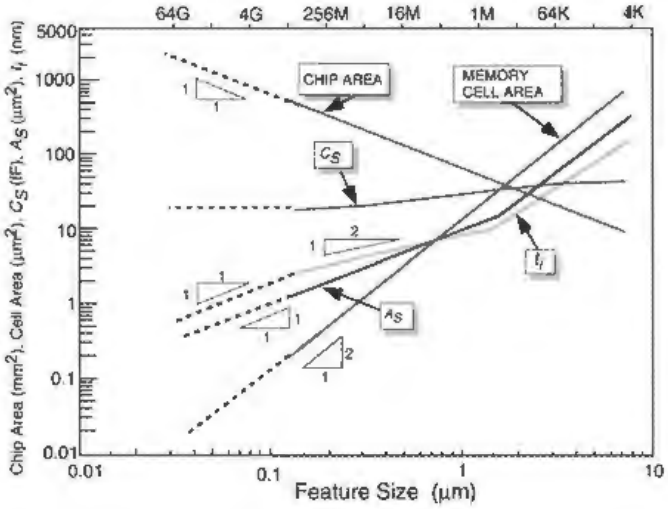


Fig. 4.4. Structural parameter trends in the DRAM cell [4.2]

plate, in which a gate–source MOS capacitor was utilized as a cell capacitor, was standard (Fig. 4.5). The capacitor-plate (gate) voltage is V_{DD}, while the storage-node (source) voltages are V_{DD} or 0 V, depending on the stored information. In the 256 Kb generation, however, a change was made to the V_{SS} (0 V) plate, since the V_{DD} plate suffered from a signal charge loss due to a V_{DD} bump at the plate, as explained later. The n^+ layer can form a capacitance despite the V_{SS} plate. The p^+ layer increases the capacitance and works as a barrier against soft error. Note that the maximum voltage across the gate insulator, for both the V_{DD} plate and the V_{SS} plate, is V_{DD}. For the 1 Mb generation and beyond, the half-V_{DD} plate [4.6] has become standard, since

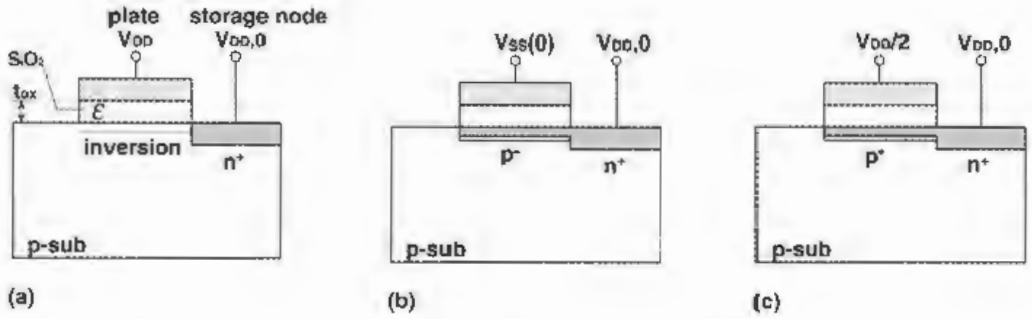


Fig. 4.5. Schemes for voltage supply to the cell capacitor plate [4.5]. (**a**) V_{DD} plate (4–64 Kb); (**b**) V_{SS} plate (256 Kb); (**c**) half-V_{DD} plate (1–256 Mb)

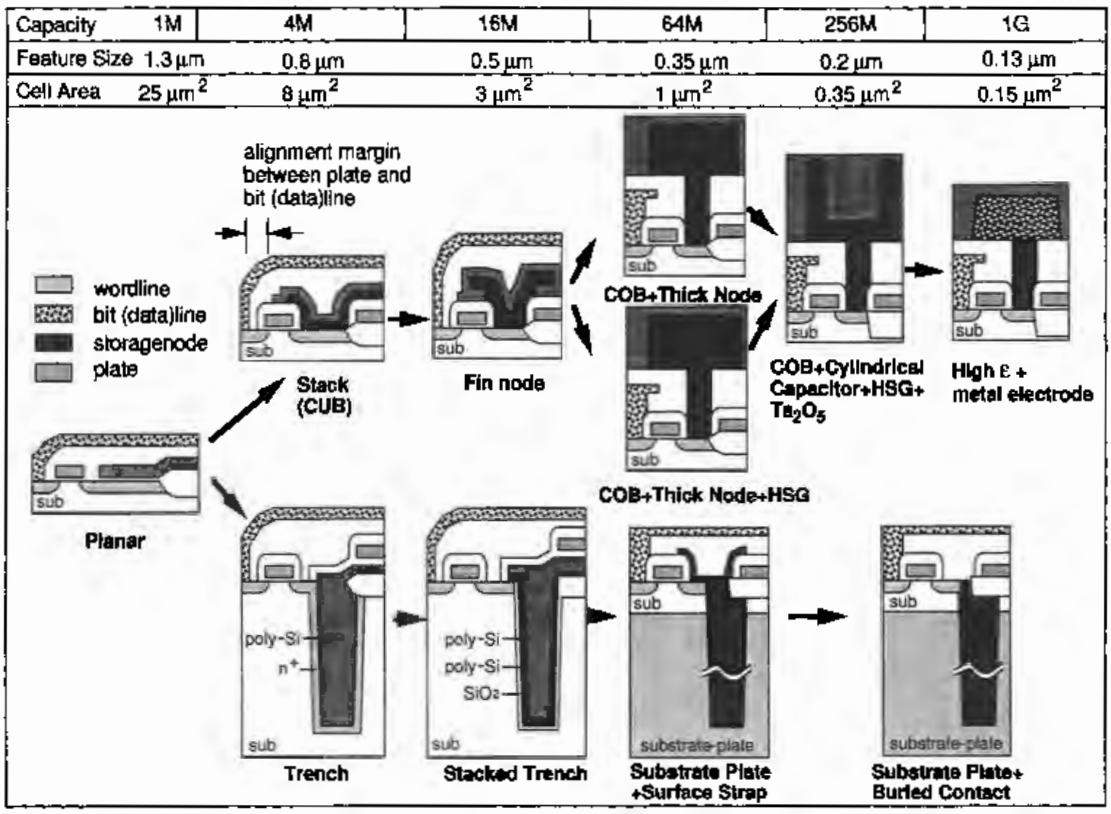


Fig. 4.6. Changes in DRAM memory cell [4.56, 4.64]

it halves the insulator thickness for the same electric field, enabling doubled values of C_S and Q_S.

To increase the surface area A_S , as shown in Fig. 4.6, the 4 Mb generation used three-dimensional cells such as stacked-capacitor cells (STC cells) and trench-capacitor cells (trench cells) [4.7–4.64] instead of the planar-type cells in which a capacitor dielectric film was formed on a silicon substrate. Figure 4.7 shows the detailed structures [4.5].

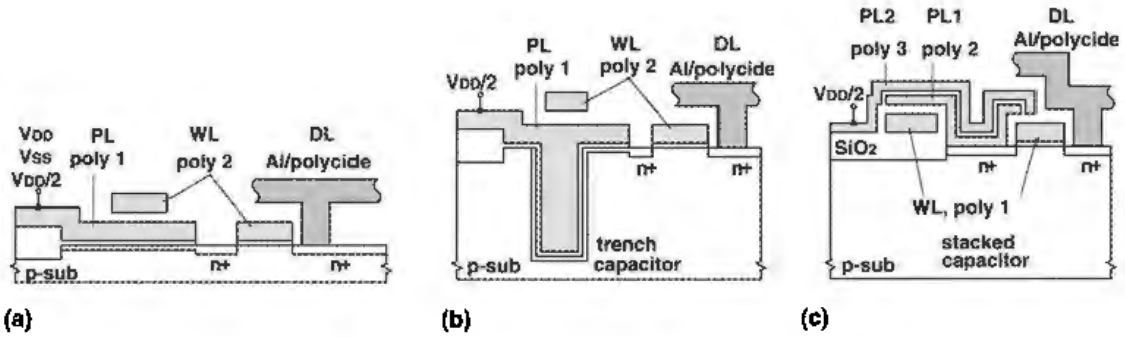


Fig. 4.7. Detailed structures of typical cells [4.5]. WL, word line; DL, data line; PL/PL1/PL2, capacitor plates (electrodes). (**a**) Planar cell; (**b**) trench cell; (**c**) stacked cell

The Stacked Capacitor Cell. Although capacitor formation on a poly-silicon surface was difficult, multilayered dielectric films consisting of silicon-dioxide and siliconnitride greatly increased the capacitance and reliability of STC cells [4.65]. For 16 Mb DRAMs, a storage-node electrode with a fin was formed to enable the desired storage capacitance, even thongh this process was a little complicated [4.66]. One drawback of the STC cell is a large difference in height between the memory cell array and the surrounding peripheral circuits, which makes the delineation of metal wiring difficult.

To keep the storage electrode height unchanged as much as possible, several innovative ideas were proposed for the STC cells used for the 64 Mb generation. These led to the capacitor-over-bit (or data) line (COB) cell and the hemispherical-grain (HSG) poly-silicon electrode [4.11, 4.12], combined with a thick node. The COB cell eliminated one of the lithographic alignment problems inherent in conventional STC cells that have a capacitor-under-bit-line structure (the CUB cell). Thus, the COB cell made enlargement of the storage capacitor easier. The HSG poly-silicon electrode also increased the capacitor area. The capacitance could be almost doubled compared with that of the conventional non-HSG electrode [4.67]. In addition, self-aligned contact between bit lines and word lines, and global planarization using chemical mechanical polishing, became essential methods for device miniaturization, starting with the 64 Mb generation. The employment of higher-ε films [4.13] is important, especially for STC cells. At around 64 Kb, Si_3N_4 film almost doubled ε, as shown in Fig. 4.8. Ta_2O_5 [4.67 -4.69], with another doubled ε, may be the next candidate at and from 256 Mb. Even with Ta_2O_5 film, the aspect ratio of the cylinder-type COB capacitor becomes more than 15 in 4 Gb DRAMs, as demonstrated in Fig. 4.9 [4.11, 4.17]. Thus, capacitor dielectric films with a higher permittivity are required. Some candidates are shown in Fig. 4.10. Note that the maximum storable charge of the material is proportional to the product of ε and the breakdown field strength. It is believed in R&D today that BST ($Bi_xSr_{1-x}TiO_3$) [4.70] may be the prime candidate, since it offers a higher ε at low voltage. Even with these materials,

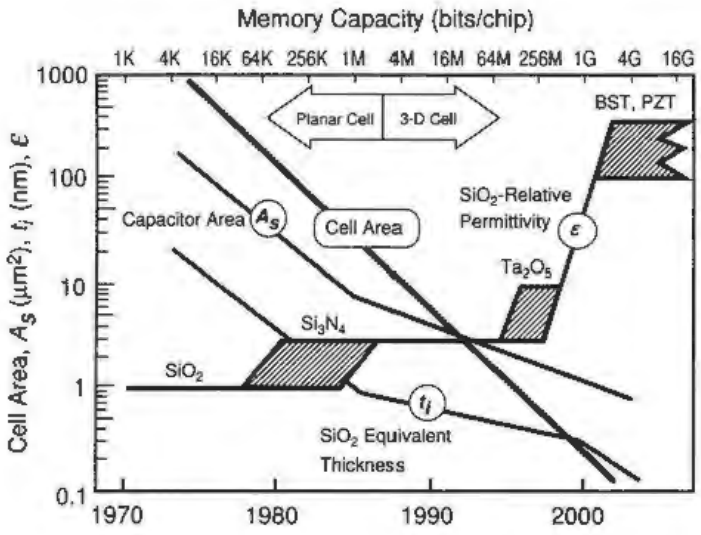


Fig. 4.8. Trends in the development of high-ε materials [4.2]

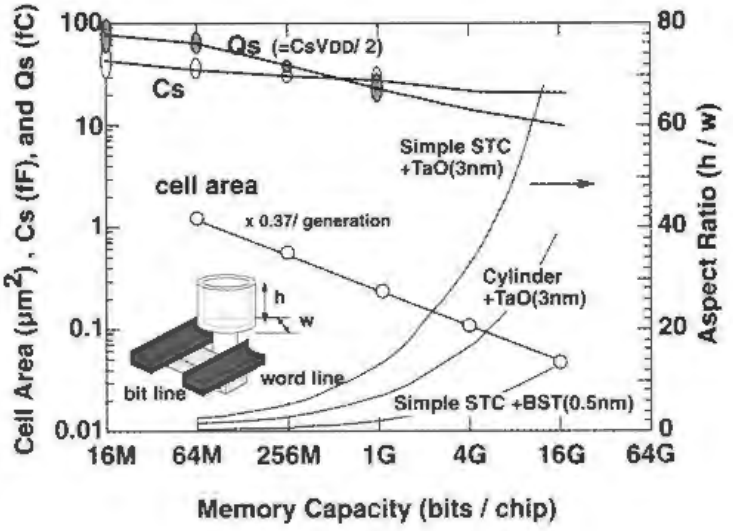


Fig. 4.9. Trends in the stacked-capacitor DRAM cell. Trends in storage capacitance are obtained by using data reported at IEDM and in [4.17]. A decrease in Q_S by a factor of 0.65 per generation results from the assumptions of a reduction in C_S by a factor of 0.93 and a reduction in V_{DD} by a factor of 0.7

further development of the 1-T cell DRAM may not be possible [4.2] when the insulator physical thickness (t_i') has reached about half of the feature size. At this point the film will fill up the gap in a storage node, preventing penetration of the capacitor plate into the gap. For example, a t_i of 0.2–0.4 nm is targeted today in high-ε films that have a physical thickness of 20–40 um. Thus, when the feature size reaches 40–80 nm, which is the case for 4 Gb to 64 Gb – the capacitor cannot be formed, as shown in Fig. 4.11. This implies the need for a physically thinner and higher-ε film. At present, however, ε decreases

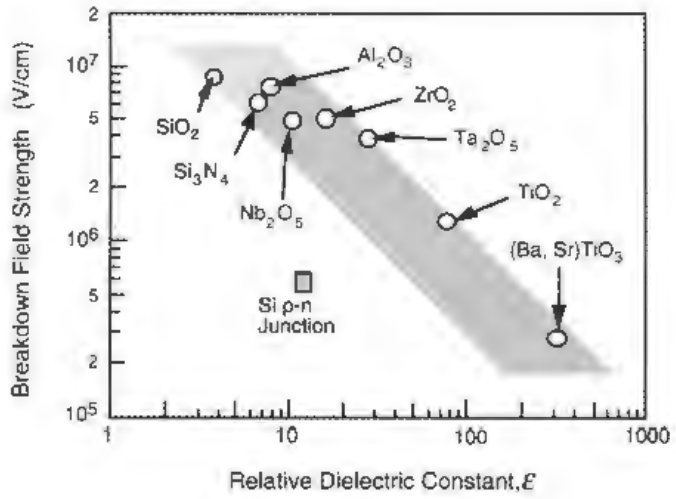


Fig. 4.10. High-ε dielectric materials [4.2]

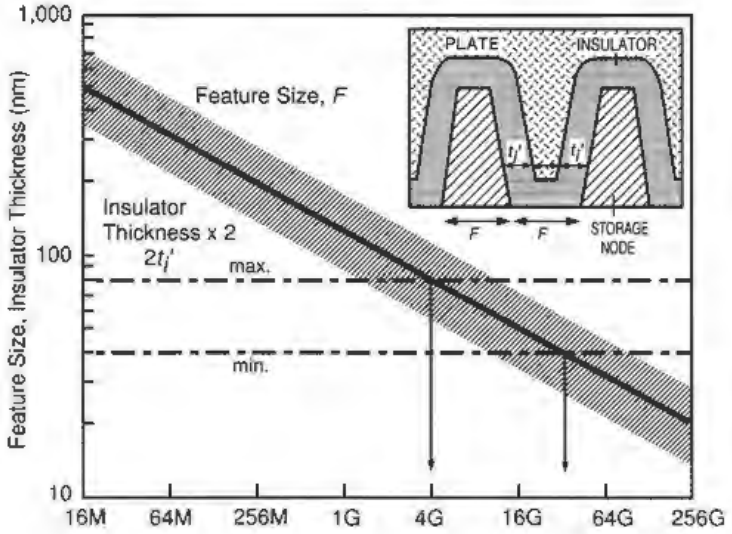


Fig. 4.11. The physical limitation of the insulator thickness, caused by the adjacent storage-node gap [4.2]

with decreasing physical thickness [4.14, 4.15], as shown in Fig. 4.12, and due to this decrease, the merit of a high-ε material is lost. One cause of the degradation in ε has to do with the lower-ε materials formed at the surface of the BST film during fabrication.

The Trench Capacitor Cell. For the trench cell, a sufficient storage capacitance has been obtained using a deep trench. However, a trench capacitor surrounded by an n^+-diffused region that is connected to a switching MOSFET has been found to be very vulnerable to α-particle irradiation. Thus, for the 16 Mb generation, a stacked trench cell in which a storage node of polysilicon was formed within a silicon dioxide cylinder [4.71] was used. However,

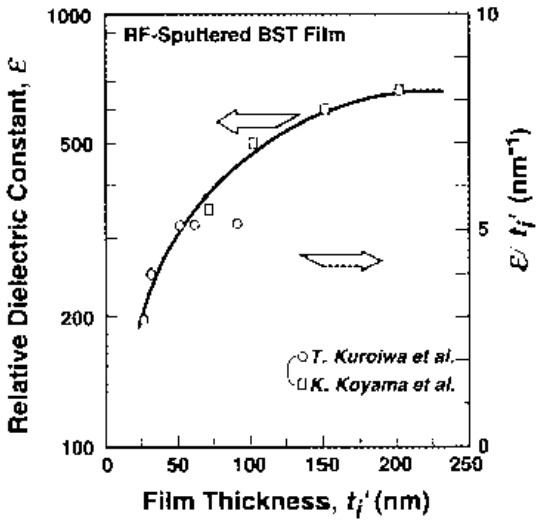


Fig. 4.12. Permittivity versus physical thickness for BST film [4.2]

a substrate-plate type of trench cell replaced the stacked trench cell in the 64 Mb generation, since making a storage node within a trench surrounded by an oxide film became very difficult. Also, innovative technologies were applied to the trench cell in the 64 Mb generation. A deep trench capacitor with a large aspect ratio, of more than 40 – which is a great challenge – was used, and chemical mechanical polishing was applied to achieve an entirely planar structure [4.63]. A buried contact that connects the storage node and the diffused region of a switching MOSFET in the substrate also improved scalability. The drawback of the trench cell, the need for a large-aspect-ratio trench, stems from the following two reasons: the capacitor hole is inevitably small, because it must be formed between MOSFETs; and the permittivity of the insulator, usually made of layered SiO_2 and Si_3N_4, is lower compared with that of the STC insulator. Note that higher-permittivity insulators are not accepted, since a high-temperature heat treatment during MOSFET formation degrades the insulator characteristics. An advantage of the trench cell, however, is that the substrate remains almost flat, even after the capacitor and isolation formation, enabling the accommodation of multilevel wiring with a tight pitch. In addition, since MOSFETs are fabricated after capacitor formation that needs a relatively higher-temperature heat treatment, the device performance of trench DRAMs is superior to that of STC DRAMs, especially in terms of short channel characteristics. Thus, the trench cell is considered to be more suitable for high-performance DRAM-embedded system LSIs. However, process harmonization or compromise is necessary to accommodate both the trench and the STC cell when implementing DRAM memory arrays, because DRAMs use several specific device structures to permit miniaturization, such as SAC (self-aligned contact), which are not always suitable for logic-device fabrication [4.72].

Both STC and trench cells are candidates for use in gigabit DRAMs. However, it is becoming very difficult to obtain sufficient storage capacitance as memory cells become smaller.

An empirical scaling technique [4.2] shown in Table 4.1 provides a summary. C_S should be maintained at around 20 fF, even in the multigigabit era, to keep the cell signal voltage large. Thus, to cope with cell-area reduction, t_i reduction must again be accelerated from $F^{1/2}$ to F. This is a strong driving force for the development of new high-ε materials. Otherwise, A_S-enhancement must take over the role of the insulator. However, no feasible techniques have been proposed yet.

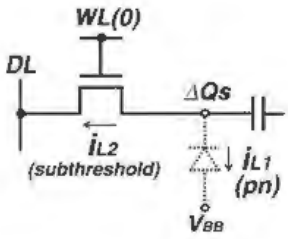
Table 4.1. The scaling law for DRAM-cell parameters [4.2][a]

Parameter	1Kb-1Mb	1Mb-1Gb	1Gb-1Tb
Cell area		$\propto F^2$	
Capacitor area, A_S	$\propto F^2$	$\propto F$	
Insulator thickness, t_i	$\propto F^2$	$\propto F^{1/2}$	$\propto F$
Storage capacitance, C_S	1	$\propto F^{1/2}$	1
Scaling technique	Area scaling	Three-dimensional	New material

[a] F, feature size; $C_S \propto A_S/t_i$.

4.2.2 Leakage Charge

There are two leakage currents [4.2] at the storage node of a non-selected cell: the p–n junction leakage current (i_{L1}) to the substrate, and the subthreshold current (i_{L2}) to the data line (DL), as shown in Fig. 4.13. Here, the worst case for each leakage current is shown. A cell that is not selected during the t_{REFmax} period must hold data even under the worst conditions. The worst conditions are established by the combination of the maximum junction temperature (T_{jmax}) and successive low-level disturbances from the corresponding data line. This is because T_{jmax} maximizes the above cell leakage currents. Here, T_{jmax} is attained by the maximum ambient temperature (usually 70 °C) and the minimum cycle time operation in the catalog specification. The disturbance further enhances the subthreshold current when the lowest DL voltage (0 V) is supplied for as long as possible. Thus, the worst conditions are eventually the successive low-level data-line disturbances due to successive operations of other cells on the data line at the minimum cycle time and maximum ambient temperature, as shown in Fig. 4.14. In the worst case, both leakage-current components (i_{L1} and i_{L2}) are maximized and contribute to the degradation of cell-retention characteristics. Note that if successive high-level (V_{DD}) data-line disturbances are applied, the i_{L2} component disappears and only the i_{L1} component is developed, enabling separation of the two



i_{L1} : pn- junction current
$< \Delta Qs / t_{REFmax}$

i_{L2} : subthreshold current
$< \Delta Qs / t_{REFmax}$
$\propto 10^{-V_T/S}$

Fig. 4.13. Two leakage-current components of non-selected cells [4.2]

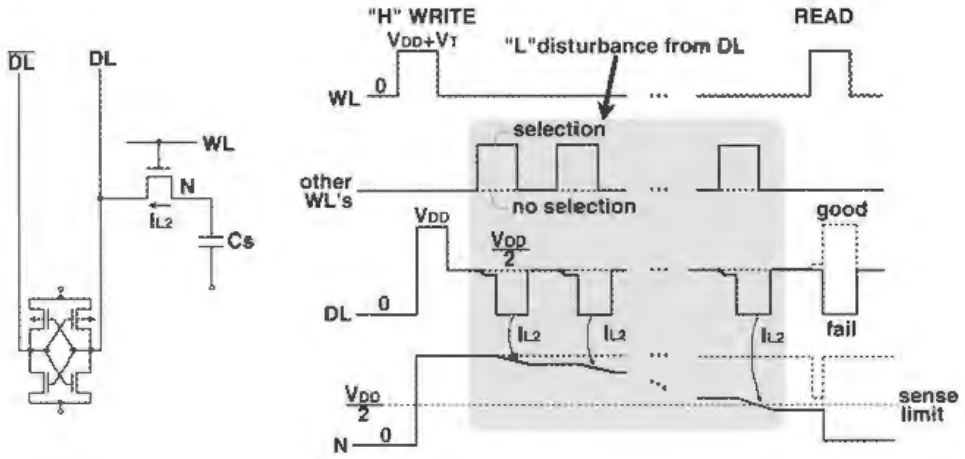


Fig. 4.14. The loss of stored data due to the data-line 'L' disturbances. $Q_C = 0$ and $Q_N = 0$ are assumed

components. The i_{L2} component is neglible for a high enough cell V_T. The acceptable leakage current must be evaluated for the worst conditions.

Reduction of i_{L1}. Figure 4.15 shows the acceptable i_{L1} for each generation [4.2]. A 10% degradation in Q_S due to i_{L1} during t_{REFmax} for the Q_S of each generation shown in the figure, and a sufficiently high cell V_T are assumed. Obviously, i_{L1} must be reduced in each successive generation, thus reaching less than 3.4 aA at 25 °C for 64 Gb. In the past, the current has been suppressed by adjusting the potential profile, so as to release the stress voltage at the junction [4.57–4.59]. In addition to reducing contamination and eliminating leaky cells by redundancy, reduction of the junction temperature (T_j) by using low-power circuits [4.16] and low-thermal-resistauce (θ_{ja}) packages has been effective, since i_{L1} is quite sensitive to temperature.

To relax the i_{L1} requirement, power suppression and a shortened t_{REFmax} specification – if users can accept it – are effective. Figure 4.16 shows the relationship between the power reduction and the cell-leakage charge at each generation [4.17]. Here, the leakage charge, ΔQ_L , is expressed as

$$\Delta Q_L = i_{L1} \cdot t_{REFmax},$$
$$i_{L1} \simeq A \exp(-E_a/kT_j),$$
$$T_j = T_a + \theta_{ja} P,$$

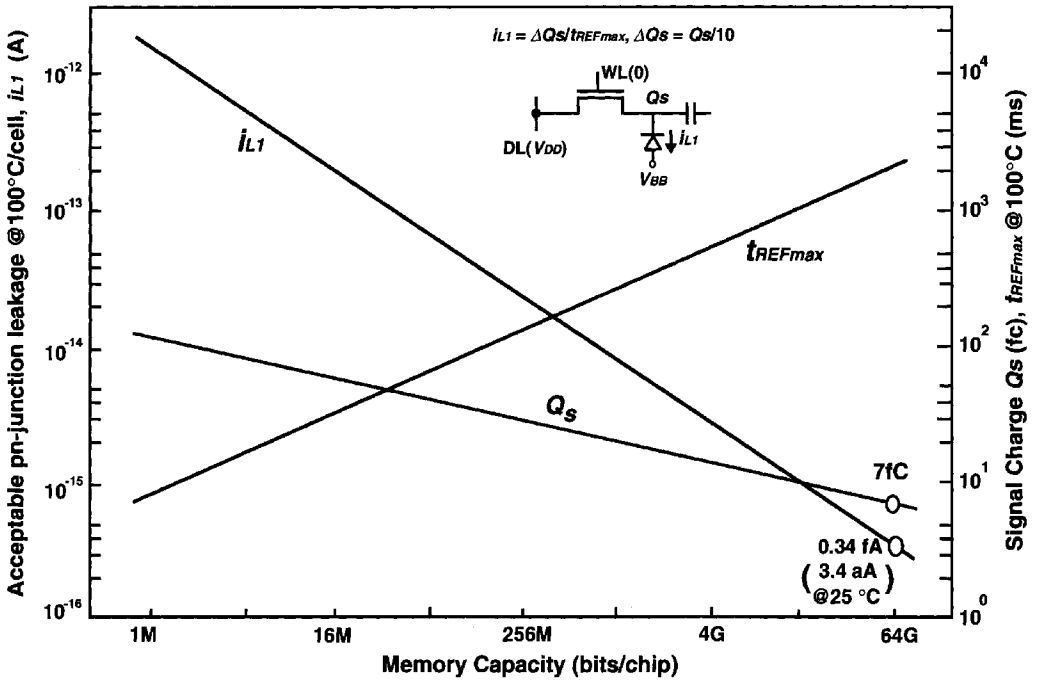


Fig. 4.15. An acceptable p–n junction leakage current for ensuring t_{REFmax} [4.2]

where A, E_{a}, T_{a}, and P are a factor that depends on the cell area, the activation energy, the ambient temperature, and the power dissipation of the chip, respectively. Thus, the ratio of the leakage charge to that of the preceding generation is given by

$$\Delta Q_L(4)/\Delta Q_L(1) \simeq \{t_{\mathrm{REFmax}}(4)/t_{\mathrm{REFmax}}(1)\} \cdot \{A(4)/A(1)\} \exp(E_{\mathrm{a}}\theta_{\mathrm{ja}}\Delta P/kT_{\mathrm{a}}^2),$$
$$\Delta P = P(4) - P(1),$$

where it is assumed that

$$T_{\mathrm{a}}(4) = T_{\mathrm{a}}(1) = T_{\mathrm{a}} \quad \text{and} \quad \theta_{\mathrm{ja}}(4) = \theta_{\mathrm{ja}}(1) = \theta_{\mathrm{ja}},$$

and where E_{a}, $A(4)/A(1)$, θ_{ja}, and T_{a} are assumed to be 0.63 eV, 0.63, 100 °C/W, and 343 K, respectively. Note that A is determined of the peripheral component by the p–n junction leakage current of the cell storage node, rather than that of the area component [4.18]. Here, we assume that the same amount of leakage charge is accepted at each generation, with $\Delta Q_{\mathrm{L}}(4)/\Delta Q_{\mathrm{L}}(1) = 1$. The traditional approach of doubling t_{REFmax} in each generation (Fig. 3.16) requires a power reduction of as much as 40 mW in each successive generation, as shown by point A in the figure. In the conventional design, however, the ever-stronger need to achieve a faster cycle time in each generation and a multi-data-bit I/O configuration lead to an increase in chip power. Thus, this approach would burden the chip design with a power reduction requirement that cannot be achieved, unless a generous margin of cell-retention time can be gained through the ultracleaning

process. If the factor by which t_{REFmax} increases in each generation is relaxed to only $\sqrt{2}$, the stringent need to reduce power reduction is relaxed to point B. This would allow an increase of as much as 15 mW. The degradation of the refresh-busy rate (γ) by $\sqrt{2}$ for each generation with a doubling of n may be acceptable, considering that the busy rate of less than 5% for a normal RAS cycle of 100 ns is still negligibly small. But the fixed t_{REFmax} approach suffers from an unacceptably large γ of doubled degradation for each generation, even though it allows power dissipation to increase by as much as 70 mW (point C). Thus, instead of the traditional approach, the approach involving the $\sqrt{2}\ t_{\mathrm{REFmax}}$ increase may be preferable in the gigabit era in terms of power and refresh-busy rate. Note that in actual design an acceptable ΔQ_{L} decreases with increasing memory capacity, implying that $\Delta Q_{\mathrm{L}}(4)/\Delta Q_{\mathrm{L}}(1) < 1$. As described earlier, this results from reduction of the effective cell-signal voltage caused by the ever-decreasing Q_{S} and the increasing noise at each generation, in spite of a decreased C_{D}. Thus, further power reduction is needed, as exemplified by points A′, B′, and C′ in the figure, for $\Delta Q_{\mathrm{L}}(4)/\Delta Q_{\mathrm{L}}(1) \approx 0.7$.

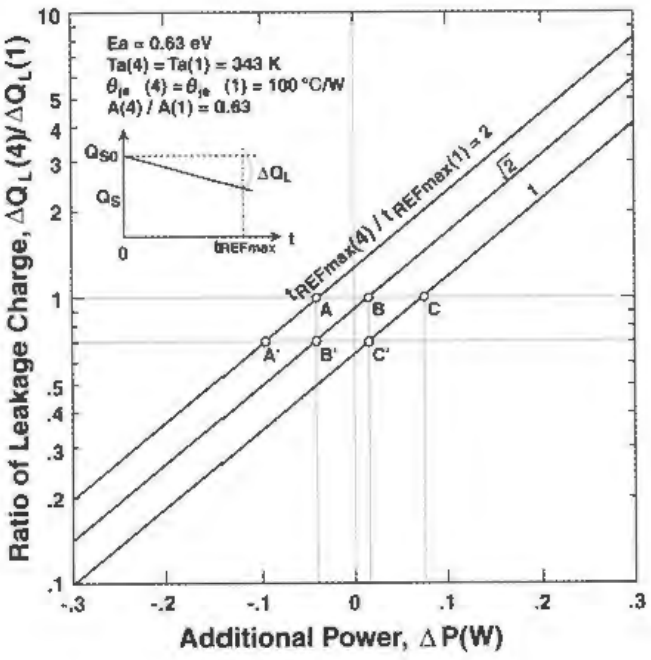


Fig. 4.16. The relationship between power reduction and the cell leakage current charge in each successive generation [4.17]. Q_{S0} corresponds to Q_{S} in Fig. 4.9

Reduction of i_{L2}. Figure 4.17 shows the minimum cell V_{T} at 25 °C [4.2] that corresponds to the acceptable i_{L2} for guaranteeing t_{REFmax} at a junction temperature of 100 °C. Obviously, to ensure an ever-longer t_{REFmax}, V_{T} has to gradually increase as memory capacity increases, reaching 1.1 V at 64 Gb.

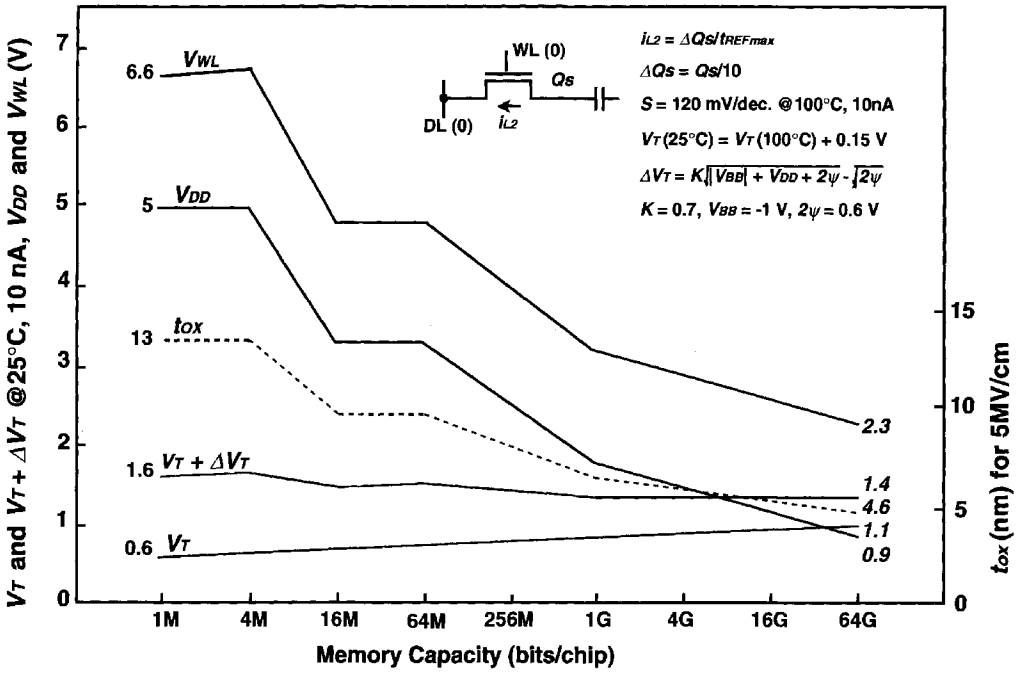


Fig. 4.17. An acceptable V_T of a cell FET, and a boosted word-line voltage for ensuring t_{REFmax} and a full write operation [4.2]. t_{REFmax} and Q_S are shown in Fig. 4.15

For 64 Gb, the boosted word-line voltage (V_{WL}), which is the sum of V_{DD}, V_T, and ΔV_T, is thus as high as 2.3 V. Here, ΔV_T is the increase in V_T for the source–follower mode of the cell FET, which is given by (2.10). To achieve the value of V_{WL} a gate oxide as thick as 4.6 nm under a stress voltage of 5 MV/cm is needed. Note that the gate oxide thickness of the peripheral-circuit MOSFETs is expected to be less than 3 nm for the 64 Gb generation. Thus, two gate-oxide thicknesses will be needed. Here, the boost ratio (V_{WL}/V_{DD}) is at least as large as 2.5, and requires an on-chip low-power voltage up-converter with a high conversion efficiency. The above leakage current suppression could be achieved by optimizing traditional device and circuit designs. It would also be helpful if the factor by which t_{REFmax} increased at each generation was relaxed to only $2^{1/2}$, as discussed above.

4.2.3 The Soft-Error Critical Charge

There have been two kinds of soft errors (i.e. non-destructive failures): α-particle induced soft errors [4.4, 4.19], and cosmic-ray neutron-induced soft errors [4.60]. As for α-particle-induced soft errors, it is fortunate that a smaller Q_C is obtained by increasing the cell density, since Q_C decreases with a reduction in the diagonal length of the depletion region of the memory cell, as shown in Fig. 4.18 [4.20]. Furthermore, fewer charge collection structures as the volume of the depletion region is reduced, such as stacked

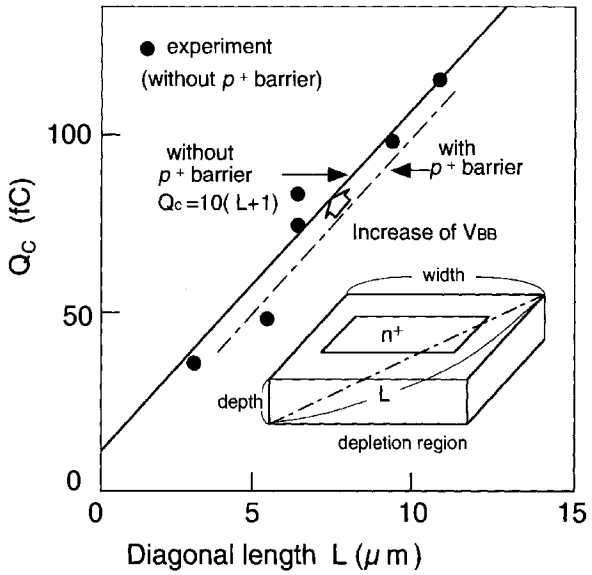


Fig. 4.18. The critical charge Q_C versus the diagonal length L of the depletion region in the memory cell [4.20]

capacitor cells and p^+ barriers, are efficient in obtaining a smaller value of Q_C. Chip coating and purification of materials are both indispensable in reducing the number of α-particles from the package and materials in a chip. On the other hand, cosmic rays hitting the atmosphere generate neutrons. These neutrons, with a small probability, interact with silicon nuclei. The resulting events, while each few in number, have a large probability of causing an error. This is because neutrons can generate about ten times as many charges (electron–hole pairs) as α-particles, as shown in Fig. 4.19. These charges are generated along the tracks of various kinds of ions that are produced by reactions between the neutrons and the silicon nuclei. The resultant electrons are collected at the cell storage node, causing a soft error. These neutrons are the main contributors to the soft-error phenomena in recent DRAMs. The soft-error rate per chip, however, remains constant even when the DRAM

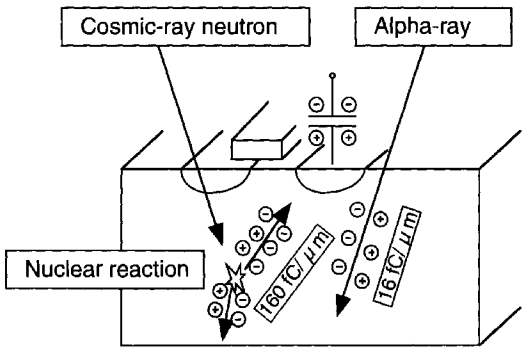


Fig. 4.19. Soft error phenomena induced by α-particles and neutrons [4.61]

memory capacity increases [4.61]. In addition, the triple-well structure of the memory-cell array improves resistivity against soft errors [4.62]. Thus, it might be possible to reduce Q_C even in the multigigabit era.

4.2.4 The Data-Line Noise Charge

A reduction in Q_N has been achieved by reductions in C_D and v_N [4.17], through a multidivided folded data-line structure combined with a shared I/O (Fig. 3.31). However, v_N is not necessarily scaled down, and requres a higher signal voltage on the data line, v_S. The extrinsic offset voltage (the V_T mismatch of the pair of FETs caused by short/narrow channel effects) of the sense amplifier is a good example. The offset voltage [4.2] generated by the deviation of numerous sense amplifiers that results from the multidivided data-line structure increases with increasing memory capacity. Note that there are one million amplifiers in a 1 Gb chip (Fig. 3.39). The offset voltage could be managed by using the traditional approach of the enlargement of the sense amplifier FETs, despite the area penalty. However, an additional (intrinsic) offset voltage resulting from random fluctuations of the MOSFET parameters [4.2], as discussed later, emerges as a serious concern. Data-line noise will be discussed in detail in the next section, because it has been, and will continue to be, essential for DRAM design.

4.3 Data-Line Noise Reduction

4.3.1 Noise Sources and Their Reduction

This section first summarizes the noises [4.1] coupled to the data line during the read operation of the 1-T cell, and describes the well-known noise cancellation method, which is paired configuration of the data lines. Next, seven kinds of noise [4.1] that still reside despite the configuration are clarified. Then, each of the noises is discussed in detail, in terms of variations of the paired configurations and the precharging method for the data line. Finally, it is concluded that a combination of the folded data-line arrangement, half-V_{DD} precharging, and simultaneous activation of the NMOS and PMOS sense amplifiers minimizes the noises.

Noise Sources. Figure 4.20 shows the noise sources of the 1-T cell. Noises are coupled to the data line during the read operation through various parasitic capacitances between the data line and cell nodes. For example, the application of a voltage to the word line causes a noise on the data line through C_{DW}. Moreover, any voltage fluctuation at the electrode, the adjacent data line, and the substrate also causes significant noise on the data line through C_{DP}, C_{DD}, and C_{DS}, if the fluctuation occurs in the floating period just before the sensing operation. Note that the components of the data-line capacitance

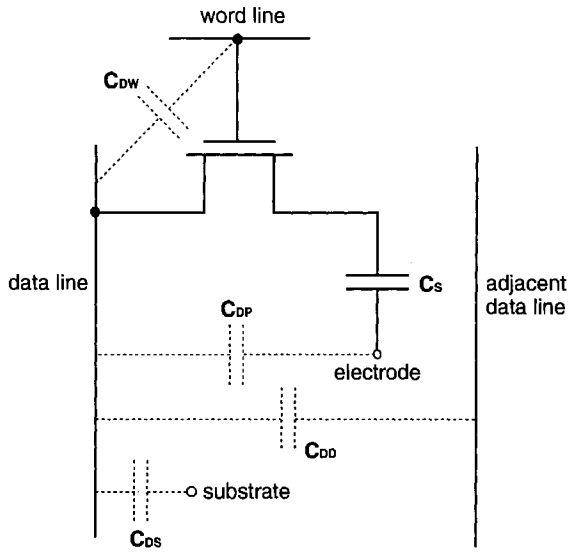


Fig. 4.20. The noise sources of the 1-T cell [4.1]

(C_D) for a 1.3 μm 1 Mb chip using a folded data-line arrangement [4.21] are 32–54% for C_{DW}, 10–15% for C_{DP}, 10–25% for C_{DD}, and 35–40% for C_{DS}, although this depends slightly on the data-line material (i.e. Al or polycide). The C_{DD} component rapidly increases with miniaturization of the cell [4.22–4.24].

Noise Cancellation. To cancel the noise, paired configuration of the data lines combined with differential sensing [4.1, 4.25] is most effective. The noise coupled to each of a pair of data lines from a noise source is rejected as a common-mode noise by a differential amplifier, if $\Delta C = \Delta C'$ and $C_D = C'_D$, as shown in Fig. 4.21a. Thus, only the signal voltage read out on either of a pair of data lines is amplified. In actual design, however, the condition of $\Delta C = \Delta C'$ is not necessarily established, even if $C_D = C'_D$ is satisfied. This causes a differential noise that affects the signal. In this case, an additional noise source that is physically equivalent can cancel the noise by the common or differential driving of two noise sources, as shown in Fig. 4.21b and c.

Residual Noises. Figure 4.22 shows various noises and their generation timings in read, and rewrite or write operations for a half-V_{DD} precharging scheme [4.1]. Our prime concern is with the floating period of the data line, from completion of data-line precharging to completion of amplification, because in this period there is a small signal that is susceptible to noise, and which easily couples various noises on the data-line due to the floating state. The following seven kinds of differential noise are generated due to the above incomplete cancellation, even if a folded data-line arrangement – that is, a paired data-line configuration to minimize noise – is adopted. Note that six major noise sources are shown as ① to ⑥ in the figure.

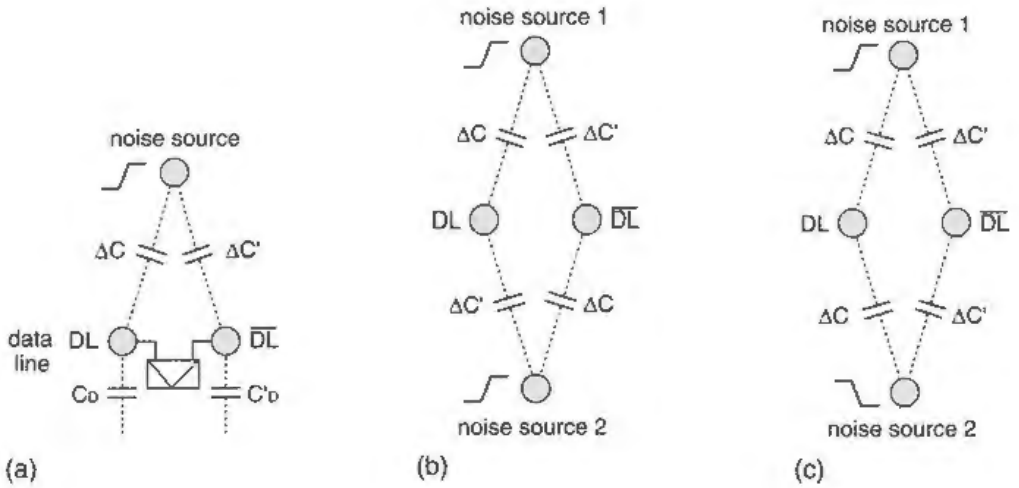


Fig. 4.21. Noise cancellation by a paired arrangement of data lines [4.1]. (**a**) Paired arrangement; (**b**) common drive; (**c**) differential drive

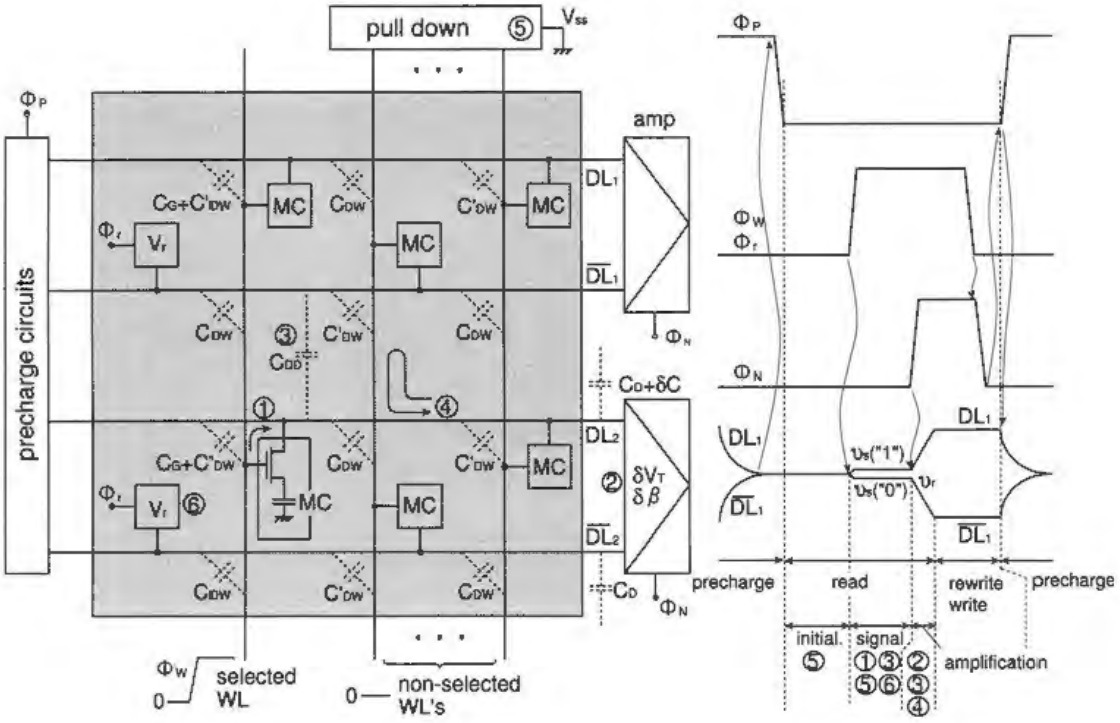


Fig. 4.22. Various types of noise and their generation timings [4.1]

- *Word-Line Drive Noise (source ①).* The activation of a word-line (WL) pulse of 5–7 V develops a differential noise superposed on a read signal on the data line (DL), because the capacitances at two crossing points between the word line and a pair of data lines are different. One is the sum of the MOSFET gate capacitance (C_G) and the WL–DL cross-capacitance

(C'_{DW}), the other is only the WL–DL cross capacitance (C_{DW}), which may differ from C'_{DW} due to the memory-cell layout. This type of noise generation corresponds to the case $\Delta C \neq C'$ in Fig. 4.21a.

- *Data-Line and Amplifier Imbalances (source ②).* A capacitive imbalance between a pair of data lines causes a differential noise when WL is activated, even if, as in Fig. 4.21a, $\Delta C = \Delta C'$. The imbalance also generates a noise during amplification. Furthermore, any imbalance regarding V_T and the conductance between paired transistors in a differential amplifier works as a noise.
- *Data-Line Interference Noise (source ③).* Interference occurs not only between small signals on the data lines, but also between large signals that are increasing during amplification. The resultant noise increases rapidly as the data-line pitch becomes smaller because of an increased value of C_{DD}. The noise depends on the read data pattern along the word line.
- *Non-Selected Word-Line Noise (source ④).* Numerous WL–DL crossing points are noise sources during amplification. Large voltage swings on the data lines during amplification couple a voltage on each of the non-selected word lines through WL–DL cross-capacitances. The resulting voltages on all of the non-selected word lines in turn couple another voltage (noise, in some cases) to each pair of data lines through all of the cross-capacitances along the data line. If the pull-down circuit at the end of each non-selected word line is ideal, all of the non-selected word lines are grounded and thus no noise is generated. Unfortunately, however, the circuit has a certain impedance.
- *Power-Supply Voltage Bounces (source ⑤).* Any voltage bounce at the V_{SS} (ground) line of the pull-down circuit causes voltage fluctuations on all of the non-selected word lines. Thus, the fluctuations couple noises on the floating data lines through all of the cross-capacitances.
- *Reference-Voltage Variation (source ⑥).* The reference voltage from the generator (V_r) must be set at the intermediate level between the "1" and "0" read signals for successful discrimination. Otherwise, the deviation works as a noise.
- *Other Noises.* Typical examples are a V_{DD} bump noise, incomplete data-line precharging, and a noise from or to the common I/O line.

4.3.2 Word-Line Drive Noise

The magnitude of WL drive noise [4.1] depends on the data-line arrangements [4.26, 4.27] – the open data-line and the folded data-line arrangements – as shown in Fig. 4.23. Here, it is assumed that a cell connected to the data line (DL) is selected. Where there is no dummy cell, the open data line never cancels the noise, despite the paired data-line configuration. A coupling voltage through $C_G + C'_{DW}$ is developed only on the data line (DL), causing a differential noise that is superposed on a read signal voltage. The folded data line, however, generates the differential noise, corresponding to

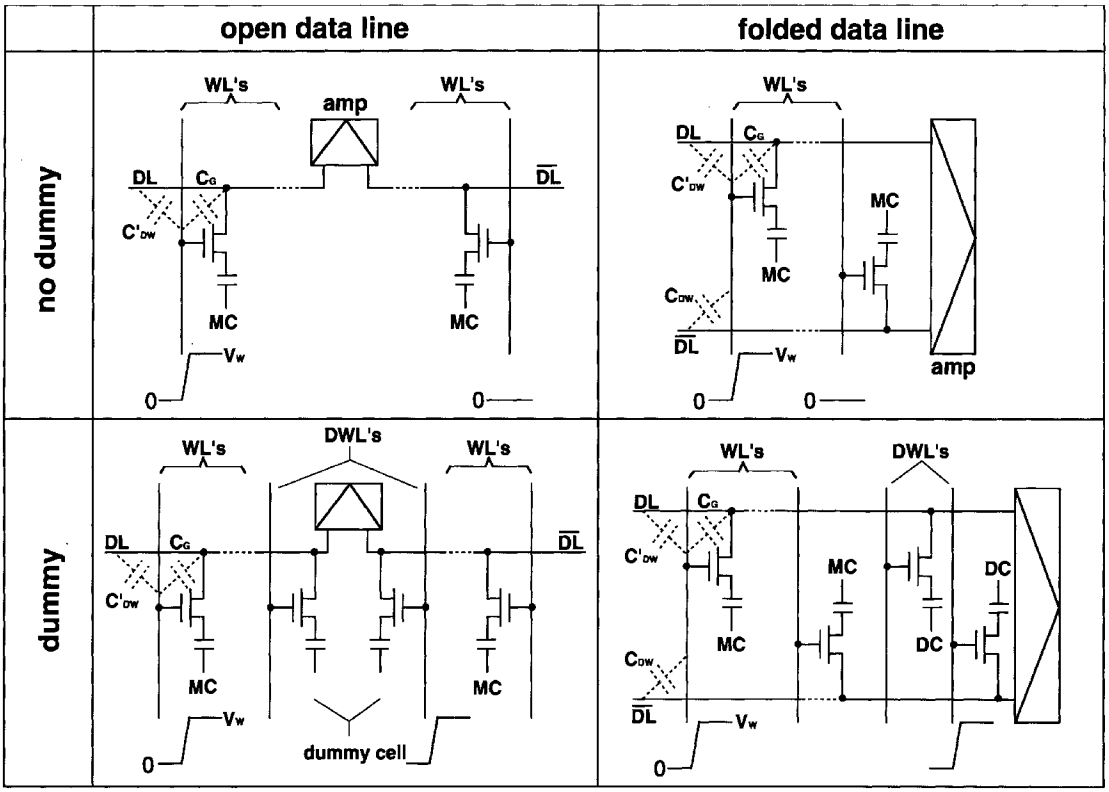


Fig. 4.23. Word-line drive noise and its reduction [4.1]

$C_G + C'_{DW} - C_{DW}$, reduced by the C_{DW} component. The noise is completely cancelled by adopting the dummy cell scheme shown in the figure. The scheme utilizes the dummy word line and the dummy cell, both of which have the same electrical characteristics as those of the actual word line and cell. In principle, one dummy cell is added to one data line: thus,there are two dummy cells for each pair of data lines. Two dummy word lines are needed to select either of two dummy cells. For example, when a word pulse is applied to read a memory cell on the data line (DL), the same pulse is simultaneously applied to a dummy word line belonging to the other data line ($\overline{DL}$). Thus, in principle, the common mode voltages coupled to each pair of data lines become the same, so that they are completely cancelled by the differential amplifier. It is fortunate if the dummy cell enables adjustment of the reference voltage of the data line and equalization of both capacitances of a pair of data lines during amplification, as discussed later.

Figure 4.24 shows experimental data [4.1] for a 1.3 μm 1 Mb cell, to estimate the degree of noise cancellation. A poly-2 word line shown in Fig. 3.56b is strapped with a metal (Al) line, and a half-V_{DD} ($V_{DD} = 4$ V fixed) precharge is applied. Here, consider the case when the cell storing a high voltage is read with an increasing word voltage (V_W), as shown in the figure. Note that a high voltage is $V_W - V_T$ (i.e. the cell-FETs V_T) for $V_W < V_{DD} + V_T$ ($\simeq 4 + 1.5 \simeq 5.5$ V), while it is V_{DD} for $V_W > V_{DD} + V_T$. The coupling voltage

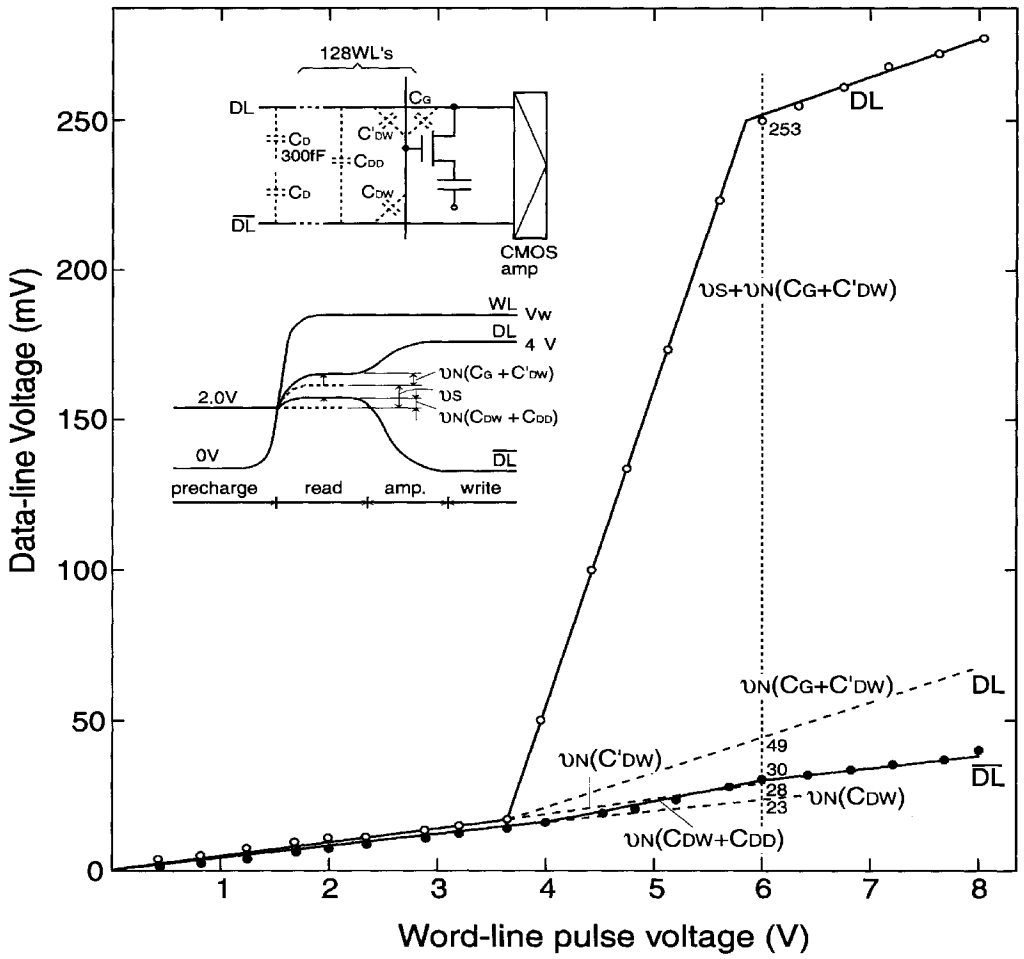


Fig. 4.24. The S/N characteristics of the 1-T cell [4.1]. WL pitch, 3.4 μm; DL-pair pitch, 10.6 μm

to the cell-connected data line (DL) gradually increases when V_W increases from 0 V, because the coupling capacitance almost equals the small WL–DL wiring capacitance (C'_{DW}). However, it starts to increase rapidly when V_W exceeds the sum of $V_{DD}/2$ (i.e. the precharged data-line voltage of 2 V) and V_T ($\simeq$ 1.5 V), and then increases in proportion to V_W. The following two components are responsible for the rapid increase. One is an increase in the coupling capacitance to $C_G + C'_{DW}$, caused by the MOS capacitance characteristics discussed in Chap. 2. The other is the appearance of a signal voltage (v_S) on the data line whose amplitude is proportional to the high stored voltage ($V_W - V_T$) described above. However, the coupling voltage tends to saturate beyond $V_W = 6$ V, because the increase in v_S finally stops due to saturation of the stored voltage (i.e. V_{DD}, because $V_W > V_{DD} + V_T$). Therefore, only $C_G + C'_{DW}$ contributes to the slope beyond 6 V. If the same slope line is drawn at $V_W = 3.5$ V, it implies noise coupled only from $C_G + C'_{DW}$. On the other hand, a voltage through C_{DW} couples to another data line (DL) until V_W

reaches 3.5 V. Above 3.5 V, however, v_S on the data line (DL) couples to DL through C_{DD}, slightly raising the DL voltage. Hence, the C_{DD} noise is the difference between the slope below 3.5 V and the slope above 3.5 V. Above 6 V, the slope is again the same as below 3.5 V because of the saturation.

The noise components at $V_W = 6$ V, $v_N(C_G + C'_{DW})$, $v_N(C'_{DW})$, $v_N(C_{DW} + C_{DD})$, and $v_N(C_{DW})$, are estimated with these data as 49 mV, 28 mV, 30 mV, and 23 mV, respectively. Thus, a $v_N(C_G)$ of 21 mV and a $v_N(C_{DD})$ of 7 mV are obtained. Moreover, since $v_N(C'_{DW})$ and $v_N(C_{DW})$ are proportional to V_W (6 V) and C_D is 300 fF, C'_{DW}/C_D, C_{DW}/C_D, C'_{DW}, and C_{DW} are 0.47%, 0.38%, 1.41 fF, and 1.14 fF, respectively. The reason why C'_{DW} is 24% larger than C_{DW} comes from additional Al–poly-2 and gate-diffusion overlapped capacitances at the Al-diffused contact of the data line. Since the difference in V_W between 6 V and 3.5 V contributes to $v_N(C_G)$, C_G/C_D and C_G are 0.88% and 2.63 fF, respectively.

The following is a noise estimation [4.1] for the data-line arrangements and precharging methods using the 1.3 μm 1 Mb data above. The voltages are changed to practical values of $V_{DD} = 5$ V, $V_W = 7.5$ V, and $V_T = 1.5$ V.

Data-Line Arrangements. A noise comparison between the open data-line and the folded data-line is made, assuming the use of a half-V_{DD} precharge and the same C_D, and no dummy word line. For the open diffused data line shown in Fig. 3.56a, C'_{DW} is assumed to be half that of the folded data line, causing a C'_{DW}/C_D of 0.24%. This is because the open data-line structure is simple, while a folded Al data line is sandwiched between the lower poly-2 word lines and the upper Al-2 word lines. Moreover, C_{DD} is negligible for the open data line because of the doubled pitch of the data line buried in the substrate. For the open data line, the total noise is 49 mV, which is the sum of $7.5\,\text{V} \times C'_{DW}/C_D$ and $(V_W - V_T - V_{DD}/2) \times C_G/C_D$. On the other hand, for the folded data line it is 38 mV which, is sum of $7.5\,\text{V} \times (C'_{DW} - C_{DW})/C_D$ and $(V_W - V_T - V_{DD}/2) \times C_G/C_D$. Thus, the noise of the folded data line is smaller by about 20%. These noises are not negligible when compared with the signal voltage of 100–200 mV.

Precharging Methods. For the folded data-line arrangement and V_{DD} precharge, the total noise is as small as 16 mV, which is the sum of $7.5\,\text{V} \times (C'_{DW} - C_{DW})/C_D$ and $(V_W - V_T - V_{DD}) \times C_G/C_D$. Therefore, the V_{DD} precharge almost halves the noise (38 mV) of the half-V_{DD} precharge.

In summary, the open data-line arrangement and half-V_{DD} precharge generate more noise. The noise tends to increase with cell miniaturization and an increase in memory capacity because of the reduced value of C_D, as shown in Chap. 3, an increase in the fringe effect, causing a larger C'_{DW} and C_{DW}, as shown in Chap. 2, and an increased C_{DD}, although it is lowered by the reduction of V_W. Thus, the ever-increasing noise poses a serious concern in the design of a higher-density DRAM, although the noise level is still low at 1.3 μm generation. In principle, a dummy cell scheme cancels the word-line

noise, independently of the data-line arrangements and precharging methods, despite an area penalty.

4.3.3 Data-Line and Sense-Amplifier Imbalances

Investigation of the noise-generation mechanism in the well-known sensing system, which is a pair of data lines connected to a differential amplifier, is indispensable in determining the necessary signal voltage of the 1-T cell. The noise derives from imbalances not only between a pair of data lines, but also between a pair of transistors at the input of the amplifier. The need for tight layouts for data lines and amplifiers is subject to the local imbalances. These imbalances are quite large when one considers their statistical deviations, which are caused by the large number of data lines and amplifiers that are distributed throughout the chip, occupying almost half of the chip area. In this section, the noise-generation mechanisms [4.27] for the sensing system are first clarified. Next, based on the results, the low-noise and high-sensitivity circuits proposed so far are investigated.

Figure 4.25 shows an NMOS amplifier, which amplifies a signal (v_S) on the data line (DL), connected to a pair of data lines. The imbalances for the data-line capacitance (C), threshold (V_T), and conductance (β) are assumed to be δC, δV_T, and $\delta\beta$, respectively. The degree of influence of the imbalances on the signal depends on the polarity of signal. Since a data line with a smaller capacitance is discharged more easily, the voltage margin of the memory cell on the data line (DL) is widened for a negative v_S for a positive δC, while it is narrowed for a positive v_S. Thus, δC works as a gain for $v_S < 0$ while it works as a loss (noise) for $v_S > 0$. The same situation occurs for $\delta\beta$. In addition, δV_T is an effective noise, because it directly affects the necessary v_S. The imbalances, δV_T and $\delta\beta$, are regarded as built-in noises of the amplifier.

If the voltage, $V_3(t)$, at the common source of Q_1 and Q_2 is decreased with the slope of K from the initial value of V_P for $v_S > 0$, Q_2 first starts to turn on at the time (t_1) when the difference between the Q_2 gate voltage, $V_1(t)$, and $V_3(t)$ reaches V_{T2}. The succeeding $V_1(t)$ and $V_2(t)$ waveforms depend on the relationship between K and the V_2 discharging speed. If β_2/C_2 is large enough, or K is small enough, both $V_2(t)$ and $V_3(t)$ decrease at the same speed, while turning Q_1 off. In actual design, however, $V_2(t)$ discharges more slowly than $V_3(t)$ because K must be quite large for high-speed sensing, and β_2/C_2 must be small due to layout limitations. Therefore, the difference in magnitude between $V_2(t)$ and $V_3(t)$ becomes larger with time, and finally Q_1 is also turned on at the time, t_2, when the difference reaches V_{T1}. After t_2, both Q_1 and Q_2 stay turned on until t_3, when $V_2(t)$ decreases to V_{T1} and thus Q_1 is turned off. Thus, $V_1(t)$ stays at a constant voltage, V_H, after t_3. $V_2(t)$ decreases more rapidly hereafter, as shown in the figure, because the Q_2 operation changes from the saturation mode to the non-saturation mode. The following analysis, however, assumes a linear change of $V_2(t)$ from t_2 to t_3, since a short period during small-signal amplification is considered.

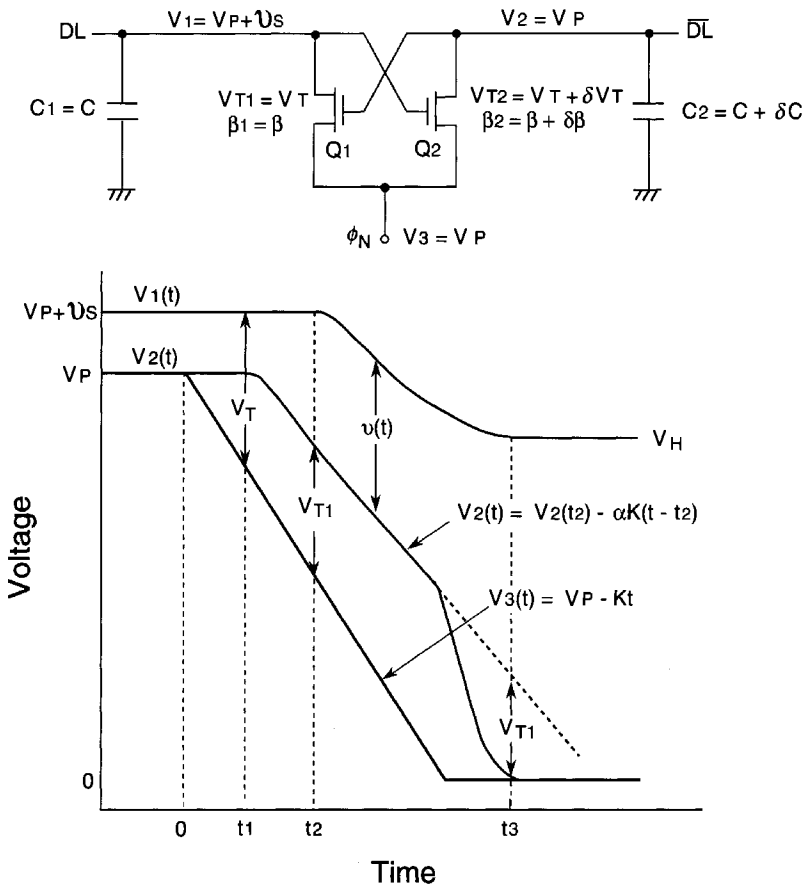


Fig. 4.25. A noise-analysis model for an NMOS flip-flop differential amplifier connected to a pair of data lines [4.1]

1. $0 \leq t \leq t_1$

$$V_1(t) = V_P + v_S,$$
$$V_2(t) = V_P,$$
$$V_3(t) = V_P - Kt,$$
$$t_1 = (-v_S + V_{T2})/K.$$

2. $t_1 \leq t \leq t_2$: Q_1 is turned off, and Q_2 is in saturation. Thus,

$$V_1(t) = V_P + v_S,$$
$$-C_2\,\mathrm{d}V_2(t)/\mathrm{d}t = \beta_2\{V_1(t) - V_3(t) - V_{T2}\}^2/2.$$

$V_2(t)$ is obtained by using the initial conditions of $t = t_1$ and $V_2(t) = V_P$ as

$$V_2(t) = V_P - \frac{\beta_2}{6C_2K}\{v_S - V_{T2} + Kt\}^3. \tag{4.4}$$

Since Q_1 starts to be turned on at $t = t_2$,

$$V_2(t_2) - V_3(t_2) = V_2(t_2) - V_P + Kt_2 = V_{T1};$$
$$\therefore -\frac{\beta_2}{6C_2K}\{v_S - V_{T2} + Kt_2\}^3 + Kt_2 = V_{T1}.$$

3. $t_2 \leq t \leq t_3$: Both Q_1 and Q_2 are in saturation. Assuming that $dV_2(t)/dt \simeq -\alpha K$ $(0 < \alpha < 1)$,

$$\begin{aligned} dV_1(t)/dt &= -i_1/C_1, \\ dV_2(t)/dt &= -i_2/C_2, \\ i_1 &= \beta_1\{V_2(t) - V_3(t) - V_{T1}\}^2/2, \\ i_2 &= \beta_2\{V_2(t) + v(t) - V_3(t) - V_{T2}\}^2/2, \\ v(t) &= V_1(t) - V_2(t); \\ \therefore dv(t)/dt &= -(\beta_1/2C_1)\{-v(t) + (2C_2\alpha K/\beta_2)^{1/2} \\ &\qquad + V_{T2} - V_{T1}\}^2 + \alpha K. \end{aligned} \tag{4.5}$$

Thus,

$$\begin{aligned} v(t) = {} & (2C_1\alpha K/\beta_1)^{1/2}[1 + A\exp\{-(2\beta_1\alpha K/C_1)^{1/2}(t - t_2)\}]/ \\ & [1 - A\exp\{-(2\beta_1\alpha K/C_1)^{1/2}(t - t_2)\}] + (2C_2\alpha K/\beta_2)^{1/2} \\ & +(V_{T2} - V_{T1}), \\ A = {} & \{v_e - (2C_1\alpha K/\beta_1)^{1/2}\}/\{v_e + (2C_1\alpha K/\beta_1)^{1/2}\}, \\ v_e = {} & v_S + (\beta_2/6C_2K)(v_S - V_{T2} + Kt_2)^3 - (2C_2\alpha K/\beta_2)^{1/2} \\ & -(V_{T2} - V_{T1}). \end{aligned}$$

In addition, since $V_1(t)$ stops discharging when $V_2(t) = V_{T1}$,

$$t_3 = t_2 + \{V_2(t_2) - V_{T1}\}/\alpha K.$$

4. $t \geq t_3$: Since Q_1 stays turned off, the final level, V_H, of $V_1(t)$ is expressed as follows:

$$V_H = V_1(t_3) = v(t_3) + V_2(t_3) = v(t_3) + V_{T1}.$$

Now, based on the above formulations, the noise can be analyzed as follows. The condition $dv(t)/dt > 0$ in the period $(t_2 \leq t \leq t_3)$ means the process of amplification, because the voltage difference between a pair of data lines becomes large with time, while $dv(t)/dt = 0$ means no amplification. Thus, if the noise is defined as the value of v_N that satisfies $dv(t)/dt = 0$, and $v(t) - v_N = 0$,

$$\begin{aligned} v_N &= (2C_2\alpha K/\beta_2)^{1/2} - (2C_1\alpha K/\beta_1)^{1/2} + (V_{T2} - V_{T1}) \\ &\simeq (1/2)(2C\alpha K/\beta)^{1/2}(\delta C/C + \delta\beta/\beta) + \delta V_T. \end{aligned} \tag{4.6}$$

It has been reported that α is about 0.5 [4.28]. Note that a PMOS latch-type flip-flop also works an amplifier, with opposite polarity to an NMOS amplifier.

Figure 4.26 shows the relationship between v_N and CK/β for $\delta V_T = 0$. Obviously, v_N becomes smaller with a smaller K; that is, slower activation of the amplifier. A smaller δC, $\delta\beta$, and δV_T are also effective in reducing noise. The following sections discuss some more effective ways of reducing noise.

The amplifier Activation Speed. Two-step driving of the common source [4.29] has been widely used. At the beginning of amplification, a signal is slowly amplified with less noise generation caused by a small K that is realized by driving a small channel width MOSFET, as shown in Fig. 4.27. Then it is quickly amplified by driving a large channel-width MOSFET (usually ten times larger). Although the rapid driving (large K) generates a large imbalance noise, the signal that has been amplified to some extent, accepts the noise. Thus, low-noise, high-speed amplification is realized. Two-step driving applied to the V_{DD} precharge scheme may allow the active restoring circuit to be omitted. A sufficiently high voltage (V_H), which is achieved by having a small K and a large signal, could be rewritten to the cell after completion of amplification without active restoration.

Capacitive Imbalance. Many circuits have been proposed to reduce $\delta C/C$ as much possible. The decoupling scheme, which isolates the amplifier from the data line during amplification, is a typical example. The amplifier reduces not only its load capacitance, but also eliminates the capacitive imbalance of the data line, allowing low-noise, high-speed amplification. Instead, however, careful attention should be paid to the amplifier layout, so as not to introduce another capacitive imbalance in the amplifier itself. The principle of decoupling is to vary the effective resistance by controlling the gate voltages of the decoupling FETs (Q and $\overline{Q}$) shown in Fig. 4.28a [4.31]. The following four kinds of the resistance control [4.30–4.35].

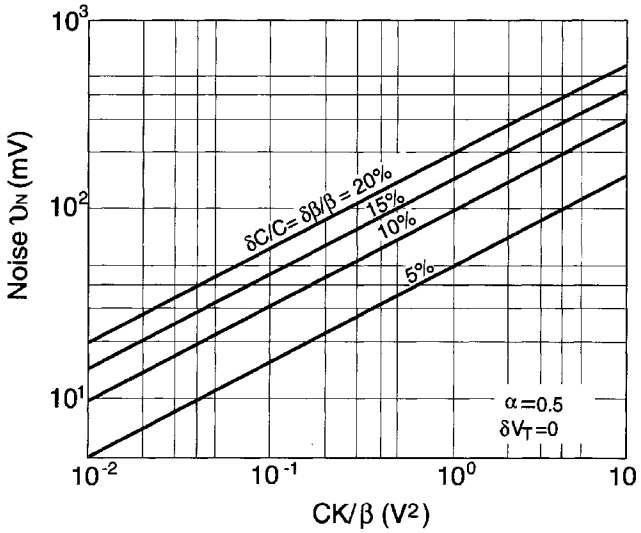


Fig. 4.26. Imbalance noises for a data line and amplifier [4.1]

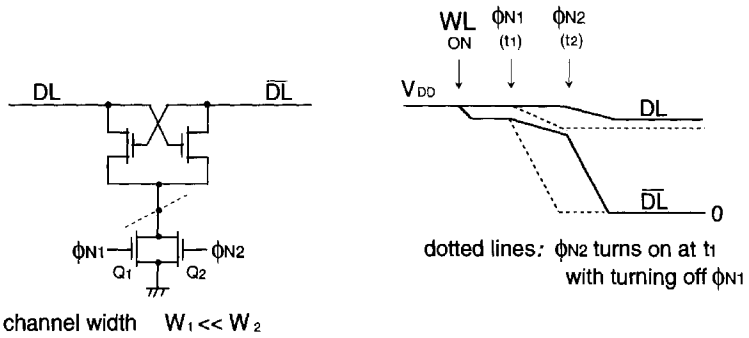


Fig. 4.27. Two-step driving for an amplifier [4.1]

Synchronized Decoupling. The Q and $\overline{Q}$ common gate line is instantaneously driven to a low level from a boosted high level of over V_{DD} by utilizing a capacitive coupling to the gate line. The low level quickly recovers to the original level during amplification. This scheme, combined with V_{DD} data-line precharge, was widely used in the 16 Kb generation.

Diode Decoupling. The gate–source connection of the depletion FETs (Q and $\overline{Q}$) works as a resistor.

High-Level Application to Decoupling FETs. A high level is always applied during amplification, which works as a resistor. This scheme has been popular in modern DRAM designs, as shown in Figs. 4.28b–d. They all employ CMOS latch-type amplifiers and half-V_{DD} precharge. Figure 4.28b [4.33] features the separation of a PMOS amplifier, which is placed on the data line, and an NMOS amplifier. This separation is needed for the application of V_{DD} to the gate of the decoupling FET. Otherwise, a decoupled CMOS amplifier, merged with NMOS and PMOS amplifiers, provides the data line with a degraded high level (i.e. $V_{DD}-V_T$) reduced by the V_T of the decoupling FET, and failing to write or rewrite a full V_{DD} to the cell. A signal amplified to some extent by NMOS amplifier activation by pulse ϕ_N is further amplified to a full V_{DD} by PMOS amplifier activation by a succeeding pulse ϕ_P. Thus, the resulting full V_{DD} on the data line is utilized as a rewrite voltage to the cell. The scheme needs an additional area caused by the separation, despite these being no need for a boosted gate voltage for the decoupling FET. Figure 4.28c [4.34] shows a decoupled CMOS amplifier combined with a boosted level (over $V_{DD}+V_T$) application to the gate of the decoupling FET. The CMOS amplifier is shared, with two pairs of data lines (DL and $\overline{DL}$, and DL′ and $\overline{DL'}$) extended to both sides. Either of two pairs of data lines is selectively connected to the amplifier. For example, if a memory cell connected to the data line (DL, $\overline{DL}$) is selected, ϕ is maintained at the boosted level of the precharging period while $\bar{\phi}$ is turned off. Thus, a read signal, superposed on the half-V_{DD} data-line precharged voltage, is amplified by the CMOS amplifier so that the resulting data-line voltage is a full V_{DD}. The scheme has been popular due to the smaller area needed and the simpler design, although the decoupling effect

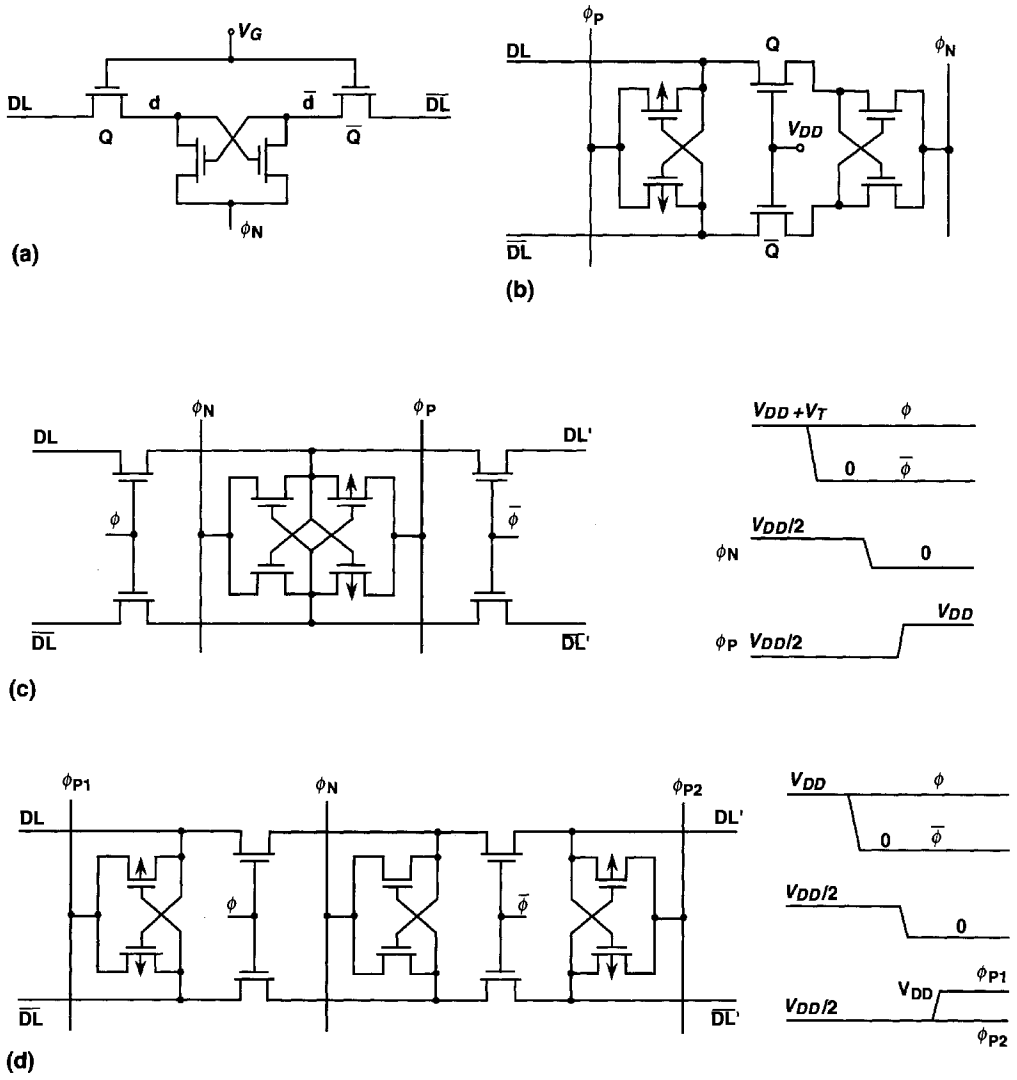


Fig. 4.28. Various kinds of amplification with data-line decoupling [4.1]. (**a**) The principle of data-line isolation; (**b**) CMOS amplifier, $V_{DD}/2$ precharge; (**c**) shared CMOS amplifier, $V_{DD}/2$ precharge; (**d**) shared NMOS amplifier, $V_{DD}/2$ precharge

is weakened by the boost voltage. Figure 4.28d [4.35] is a combination of schemes (b) and (c). The absence of any need for a boosted voltage simplifies the circuit design despite the additional PMOS amplifier.

Complete Decoupling. Amplification starts after completely isolating the amplifier from the data line, with the application of a low enough level to the Q and $\overline{Q}$ gates. This scheme results in a slower cycle despite a maximized decoupling effect, because Q and $\overline{Q}$ are turned off and then turned on again to fully rewrite an amplified signal, necessitating additional clocks.

Reduction of δC is crucial, although the above circuits reduce the detrimental effects. There are three major sources of δC: imbalance between a pair of data lines; an additional imbalance caused by tight layouts of peripheral

circuits such as the amplifier, precharge circuit, and column decoder; and an imbalance due to the cell-storage capacitance (C_S), which occurs when the cell is connected to either of a pair of data lines during reading. To reduce δC as much as possible, the folded data-line arrangement is preferable.

In principle, the folded data line has less imbalance compared with the open data line, because of its parallel running, in close proximity. It also features less additional imbalance because of the layout flexibility of the above peripheral circuits: the circuits can be laid out at both edges of a pair of data lines. On the other hand, the open data line needs them to be positioned at the center of the data lines, causing inevitable imbalances. To cancel the C_S imbalance, various dummy cells, discussed later, have been proposed.

A simultaneous and differential drive of NMOS and PMOS amplifiers [4.36] may effectively cancel δC. For an NMOS amplifier, δC (> 0) on the data line ($\overline{\text{DL}}$) in Fig. 4.25 works as a gain if a signal developed on the other data line (DL) is negative, while it works as a loss (noise) if the signal is positive, as explained before. On the contrary, for the PMOS amplifier δC works oppositely. Consequently, the δC effect is cancelled if both amplifiers are simultaneously activated, with completely differential output waveforms. Fortunately, the resulting differential waveforms also cancel the coupling noises from a pair of data lines to other conductors, enabling a quiet memory array. It would be effective if a completely differential drive against variations on V_{DD}, temperature, and fabrication process could be realized. In this drive, the PMOS amplifier would need a larger area because of conductance matching with the NMOS amplifier.

Conductance Imbalance. The sources of $\delta\beta$ are mainly differences in gate (or channel) length and width between paired MOSFETs in an amplifier. These differences are inevitably introduced in a chip by local variations in the fabrication process, although careful attention is paid to the layout so that any electrical imbalance is eliminated. In the past, both the gate length and width have been enlarged under severe layout limitations, so that a large gate length, which entails a larger gate width for a fixed β, has resulted in a smaller $\delta\beta/\beta$ for their fixed variations. It has been reported for a 16 Mb chip using a 0.5–0.6 μm gate length for the peripheral circuits [4.37] that a noise of 5.5 mV for the standard deviation (σ) is generated by a gate length variation of ± 0.02 μm. In addition to the variations, careful attention should be paid to the layout, so that electrical imbalances regarding parasitic source resistance and contact resistance between the amplifier and the data line are reduced as much as possible. Their imbalances effectively cause $\delta\beta$.

The Amplifier Offset Voltage. The variation in V_T that is unavoidably introduced during volume production is a source of the sense-amplifier offset voltage (δV_T). The reduction of this variation in V_T reduces the offset voltage. There are two kinds of sources of V_T variation; extrinsic variations and intrinsic variations.

Extrinsic Variations. As the device feature size decreases, variation in V_T [4.38] increases because of the increase in short-channel effects. Thus, in addition to enlargement of the gate length, as in reducing $\delta\beta$, stringent controls of L and t_{OX}, and a shallow junction MOSFET, formed by reducing the ion-implantation energy and the process temperature, are essential to reduce the V_T variations. A redundancy technique is also effective in eliminating large-δV_T amplifiers. In an actual 16 Mb chip, a δV_T of 31 mV (3σ) has been reported [4.39].

Intrinsic Variations. Even in the absence of extrinsic variations, intrinsic variations [4.2] in V_T, the S-factor, and the drain current start to be generated and increase rapidly as an FET is scaled down. This is caused by random microscopic fluctuations in the number and location of dopant atoms in the channel region of the MOSFET. A fluctuation-free FET is indispensable, although it is still the subject of research. There are a few techniques that may resolve this issue. One of these is to localize two-dimensional channel doping in a relatively low-impurity-concentration channel region [4.2]. Another technique is to choose a gate material the work function of which is adequate for V_T design as long as the short- and long-term reliabilities are satisfactory. Moreover, despite additional circuit complexity, the following redundancy technique may also partly resolve this issue.

The standard deviation of the intrinsic random V_T variation [4.2] is expressed as follows:

$$\sigma(V_T) = \frac{q}{C_{OX}}\sqrt{\frac{N_A \cdot D}{3L \cdot W}} \tag{4.7}$$

where q is the electronic charge, C_{OX} is the gate-oxide capacitance per unit area, N_A is the impurity concentration, D is the depletion layer width, L is the channel length, and W is the channel width. The calculated $\sigma(V_T)$ values of various sizes of MOSFETs used in DRAMs are shown in Fig. 4.29. The maximum deviation $|\Delta V_T|_{MAX}$, however, depends not only on the device parameters but also on the number of MOSFETs, N, used in a chip. The ratio $m = |\Delta V_T|_{MAX}/\sigma(V_T)$ increases with N according to the following equation:

$$1 - \frac{1}{\sqrt{2\pi}}\int_{-m}^{m} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \frac{1}{N}. \tag{4.8}$$

Table 4.2 shows $\sigma(V_T)$ and $|\Delta V_T|_{MAX}$ for various circuits in a hypothetical 16 Gb DRAM using 0.1 μm technology. The V_T deviation of the memory-cell MOSFETs is the largest because of the small channel area $L \cdot W$ and the large number of MOSFETs. However, this deviation is easily compensated by slightly raising the boosted word-line voltage V_{WL} (by 0.14 V). As for the sub-word driver and the peripheral circuit, V_T deviations are negligible compared with their operating voltages. Therefore, the most serious problem

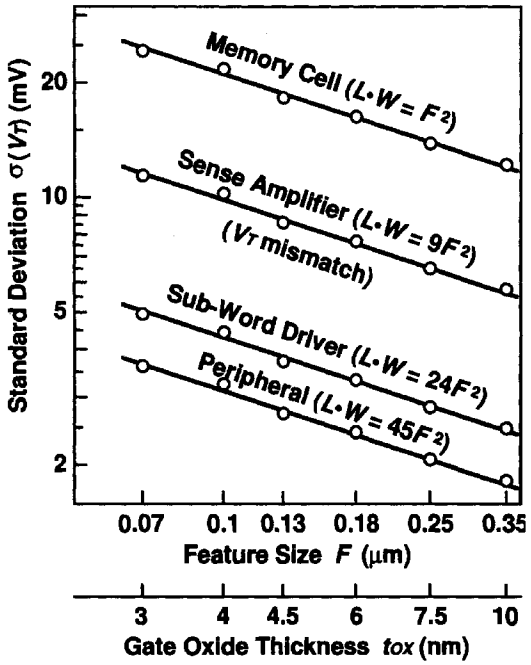


Fig. 4.29. Standard deviations of MOSFET threshold voltages (threshold-voltage mismatch for a sense amplifier) [4.2]

caused by the intrinsic V_T variation is the V_T mismatch (offset voltage) of the sense amplifier pair MOSFETs, as shown in Fig. 4.30. The V_T mismatch δV_T, which becomes an equivalent noise for the sensing operation, will be as much as 50 mV for a 16 Gb DRAM, while the signal voltage, v_S, is usually 100 mV (Fig. 4.3).

To solve this problem, two technologies [4.2] are needed. The first is memory-cell technology to maintain the signal voltage v_S as much as possible, so that a large V_T mismatch is allowed in spite of the scaling down of the memory-cell area and the supply voltage V_{DD}. The dotted lines in Fig. 4.30 show two cases for an acceptable δV_T: the first case is that it is constant with memory capacity, which is made possible by a constant v_S, and the second case is that it is reduced with V_{DD}, which is imposed by a v_S reduction proportional to V_{DD}. Here, an acceptable δV_T of 40 mV is assumed at 64 Mb. If the latter case applies, the DRAM capacity will be limited to only 256 Mb, while if the former case were to apply the limitation would be extended to 1–4 Gb. Hence, the maintenance of v_S is crucial. The second case is an advanced redundancy technique, that eliminates a certain percentage of the sense amplifiers with excessive δV_T values, in order to maintain a constant $|\delta V_T|'_{MAX}/\sigma(\delta V_T)$ ratio. Here, $|\delta V_T|'_{MAX}$ is the maximum δV_T after applying the redundancy technique. For example, if the ratio of spare columns to normal columns is 1/64 (1.6% of the memory-array area penalty), $|\delta V_T|'_{MAX}$ is limited to $2.4\sigma(\delta V_T)$. As a result, the memory-capacity limitation is extended for at least two generations for both cases, as shown in the figure. Note that

Table 4.2. Standard and maximum threshold-voltage deviations of a 16 Gb DRAM [4.2]

Circuit	MOSFET gate area,	MOSFET count,	V_T deviation of 16 Gb				Operational voltage	Design solution
	LW	N	N	m	$\sigma(V_T)$ (mV)	$\lvert\Delta V_T\rvert_{MAX}$ (mV)	(V)	
Sense amplifier	$9F^2$	8 times/2 gen.	4 Mb–8 Mb	5.2	10.2[a]	53.5[a]	$v_{SMAX} = 0.1 - 0.2$	Redundancy and/or ECC
Memory cell	F^2	4 times/2 gen.	16 Gb	6.5	21.7	142.2	$V_{WL} = 2.4$	Raised V_{WL}
Sub-word driver	$24F^2$	8 times/2 gen.[a]	4 Mb–8 Mb	5.2	4.4	23.2	$V_{WL} = 2.4$	–
Peripheral	$45F^2$	Approx. constant	64 Kb	4.3	3.2	14.0	$V_{DD} = 1.2$	–

[a] V_T mismatch of paired MOSFETs ($= \sqrt{2}\sigma(V_T)$). $1 - \frac{1}{\sqrt{2\pi}} \int_{-m}^{m} \exp(-\frac{x^2}{2})dx = \frac{1}{N}$, $m = \lvert\Delta V_T\rvert_{MAX}/\sigma(V_T)$.

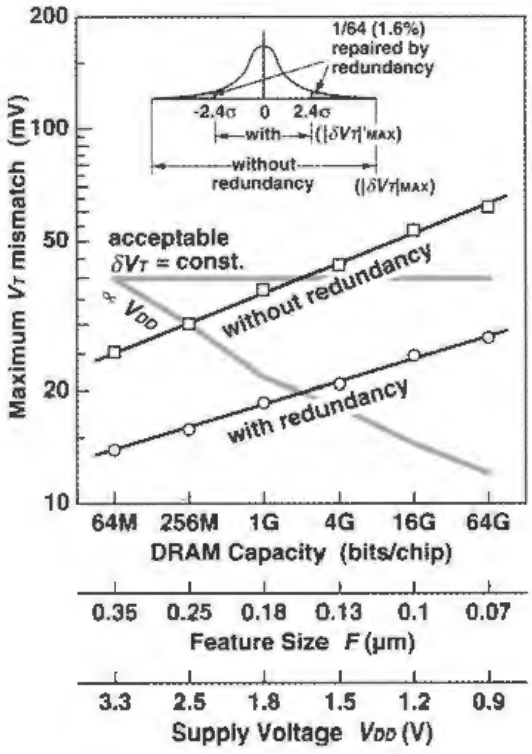


Fig. 4.30. The threshold-voltage mismatch (δV_T) of sense-amplifier MOSFETs [4.2]

v_S would actually reduce more slowly than given in Fig. 4.30, as shown in Fig. 4.3. Hence, the redundancy technique would make the limitation extend beyond 16 Gb. Such a redundancy technique would require small programming elements to minimize the peripheral area penalty, because the number of sparecolumns required would be 64−128 k for a 16 Gb DRAM. On-chip error correction may be another effective solution.

Circuit approaches may be also effective in overcoming the offset voltage issue. Figure 4.31 shows typical δV_T-compensation circuits, although they have not been used yet in commercial chips, due to an area penalty, a complicated timing control, and a slow speed. In circuit (a) [4.40], voltages corresponding to the V_T values of Q and $\overline{Q}$ (their difference is denoted by δV_T) are temporarily stored at nodes N and $\overline{N}$ when ϕ_1 is turned ou after data-line V_{DD} precharging. This is a result of the diode connections of Q and $\overline{Q}$, which are established during discharging toward V_E ($< V_{DD}$) at their common source. The voltage difference (δV_T) between N and $\overline{N}$ is almost maintained even after ϕ_2 is turned off, because the data-line capacitance is much larger than the node capacitance. Eventually, the resultant Q, $\overline{Q}$ flip-flop can cancel δV_T. This scheme requires a sophisticated timing control after precharge, especially for half-V_{DD} precharge, which causes a slow sensing time. Circuit (b) [4.41] precharges the data lines through paired MOSFETs. The flip-flop, which is composed of the same MOSFETs as a result of ϕ_1 activation, can amplify the signal, with cancellation of δV_T by a succeeding

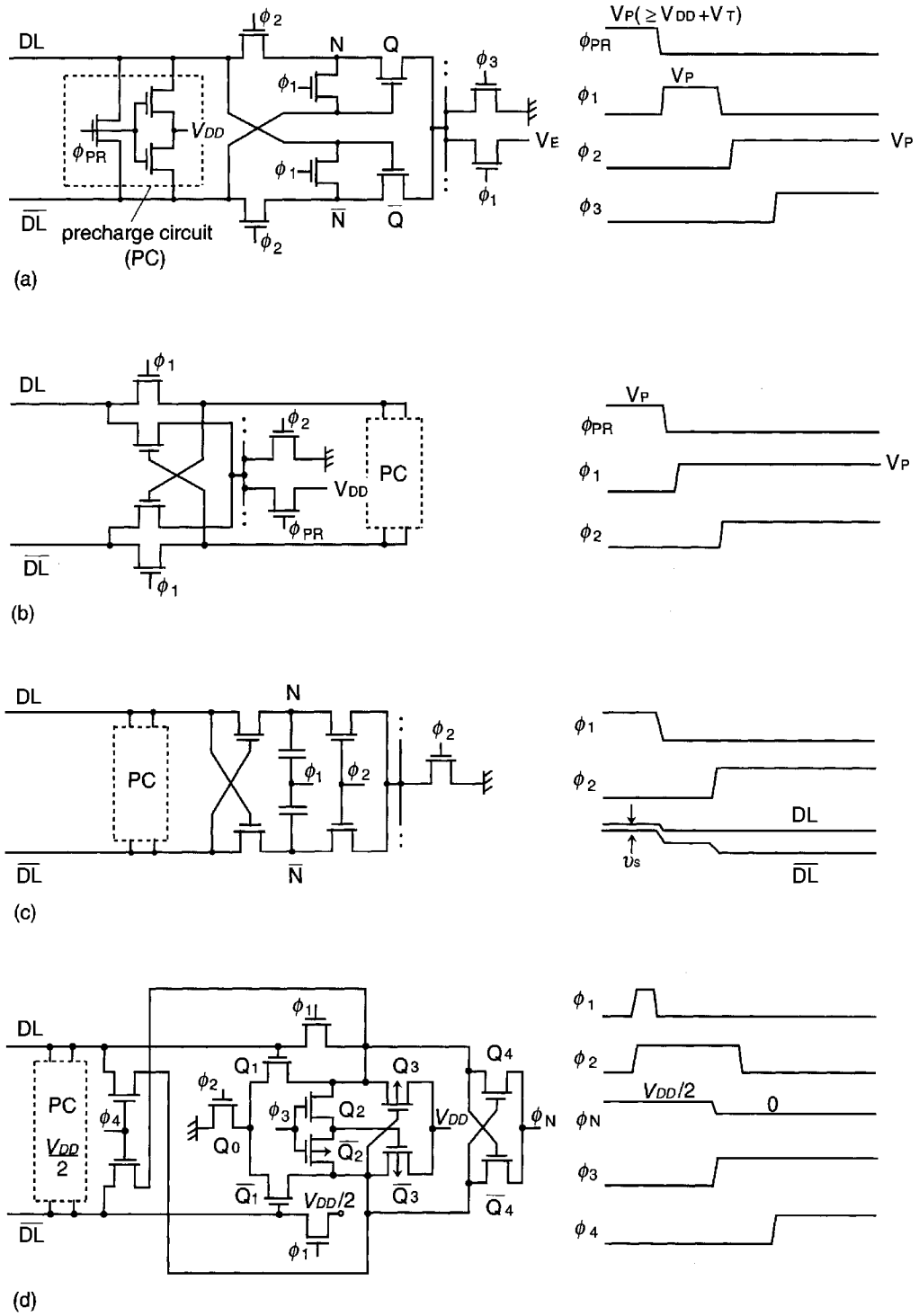


Fig. 4.31. Offset-voltage reductions for a pair of amplifier MOSFETs [4.1]. (**a**) Offset-voltage storing; (**b**) precharging by paired FETs; (**c**) capacitive coupling amplification; (**d**) a current-mirror amplifier

application of ϕ_2. Actually, a one-fourth reduction in δV_T [4.41] has been reported. Complete δV_T cancellation, however, needs a quite long data-line precharge time, because of the source-follower mode of operation. Circuit (c) [4.1] features two-step amplification. Each of nodes N and $\overline{N}$ is precharged to a voltage ($\simeq V_{DD} - V_T$) that corresponds to each V_T through paired MOSFETs. Then, signal is amplified to some extent by the application of ϕ_1, followed by a rapid amplification due to the application of ϕ_2. Note that the precharged N and $\overline{N}$ voltages are never influenced by the signal, because the signal is negative-going for V_{DD} data-line precharge. One drawback is the need for many devices. Circuit (d) [4.38] compensates for the data-line precharge level for δV_T by using a current-mirror differential amplifier. First, ϕ_1 and ϕ_2 are turned on after precharging the data lines to a half-V_{DD}. Thus, a current-mirror differential amplifier consisting of $\overline{Q}_0$, Q_1, Q_1, $\overline{Q}_3$, and Q_3 is constructed, because $\overline{Q}_2$ is on. If $V_T(\overline{Q}_1) < V_T(Q_1)$, the $\overline{Q}_1$ current is larger than the Q_1-current. Thus, the Q_3 gate is more biased to raise its drain voltage, allowing the Q_1 gate voltage to be automatically compensated for the difference, $V_T(Q_1) - V_T(\overline{Q}_1)$. The resulting Q_1 and $\overline{Q}_1$ gate voltages are held on the floating data lines after ϕ_1 is turned off. Thus, the subsequent read signal on the data line is amplified by the Q_1, $\overline{Q}_1$ differential amplifier, so that an amplified differential signal is developed at the Q_4, $\overline{Q}_4$ drains. The signal is finally amplified up to a full V_{DD} by a CMOS amplifier, composed of Q_3, $\overline{Q_3}$, Q_4, and $\overline{Q}_4$, which is activated by application of ϕ_3 and ϕ_N. The full V_{DD} is utilized as a cell rewrite voltage by activation of ϕ_4. The scheme requires many devices, although it has produced a one-sixth δV_T reduction, from 16 mV(σ) to 2.8 mV(σ), for a 0.5 μm CMOS process [4.38].

Figure 4.32 shows a charge-transfer amplifier [4.42, 4.43]: the available cell signal is so large that the δV_T issue is negligible. The small cell signal issue in the 1-T cell, which poses a serious problem in the δV_T issue, originates from the fact that the signal charge at a lightly capacitive cell node is transferred to the heavily capacitive data line. If the cell signal charge is transferred to another lightly capacitive sensing node via the data line, a large signal voltage that is independent of the data-line capacitance can be made available. This

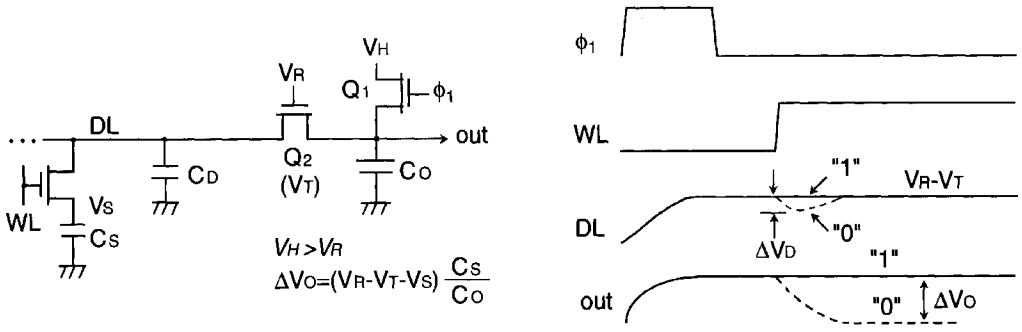


Fig. 4.32. The principle of the charge-transfer amplifier [4.42]. $V_S = V_R - V_T$ is assumed for "1"

is the principle of the charge-transfer amplifier shown in the figure. In the precharge period, the sensing (output) node and data line are precharged by Q_1 and Q_2 to V_H and $V_R - V_T$, respectively. After completing the precharge with ϕ_1 off, a cell signal is read out on the data line. If the storage node voltage (V_S) is small than $V_R - V_T$, a signal voltage swing, ΔV_D, appears on the data line, so as to turn Q_2 on. Here, the initial ΔV_D is expressed by

$$\Delta V_D = (V_R - V_T - V_S)C_S/(C_D + C_S).$$

Hence, Q_2 continues to turn on until it again charges up both C_D and C_S to $V_R - V_T$. The signal-charge decrease (ΔQ) on the data line and at the cell node during this process is given by

$$\Delta Q = \Delta V_D(C_D + C_S) = (V_R - V_T - V_S)C_S.$$

Since ΔQ contributes to the signal voltage, ΔV_O,

$$\Delta V_O = \Delta Q/C_O = (V_R - V_T - V_S)C_S/C_O.$$

Therefore, the cell-stored voltage, V_{S1} or V_{S0}, corresponding to the binary information, is transferred to the output. The voltage difference at the output is $(V_{S1} - V_{S0})C_S/C_O$. If $C_O = C_S$, a large signal voltage at the cell node is completely transferred to the output without detrimental effects due to $V_T(Q_2)$ and C_D. In general, however, the charge-transfer amplifier suffers from an extremely slow speed due to the source–follower mode of operation of Q_2. If a shorter-channel FET is used for Q_2 to enhance the speed, it in turn suffers from a drain-voltage dependence of V_T caused by short-channel effects.

4.3.4 Word-Line to Data-Line Coupling Noise

The word-line to data-line coupling capacitance, C_{DW}, is a noise source [4.1, 4.26]. For example, a large data-line voltage swing during amplification couples the voltage, δ_W, on all of the non-selected word lines through C_{DW}. Then the δ_W's cause a noise, δ_D, on the given data line, again through C_{DW}. δ_D depends strongly on the number of memory cells connected to the selected word line and their stored data patterns. Figure 4.33 shows an experimental result of voltage waveforms δ_W coupled to the word lines [4.27]. A 1.3 µm 64 Kb chip incorporating the folded data-line arrangement combined with V_{DD} precharge was used. The voltages of both the selected and non-selected word lines drop in accordance with a data-line discharge of a voltage swing of 4.5 V, and then recover to their steady states with a certain time constant of the array. This quite large voltage (δ_W) coupled to the word lines causes the following three detrimental effects on the cells. The effects are minimized by the folded data-line arrangement, combined with half-V_{DD} precharge.

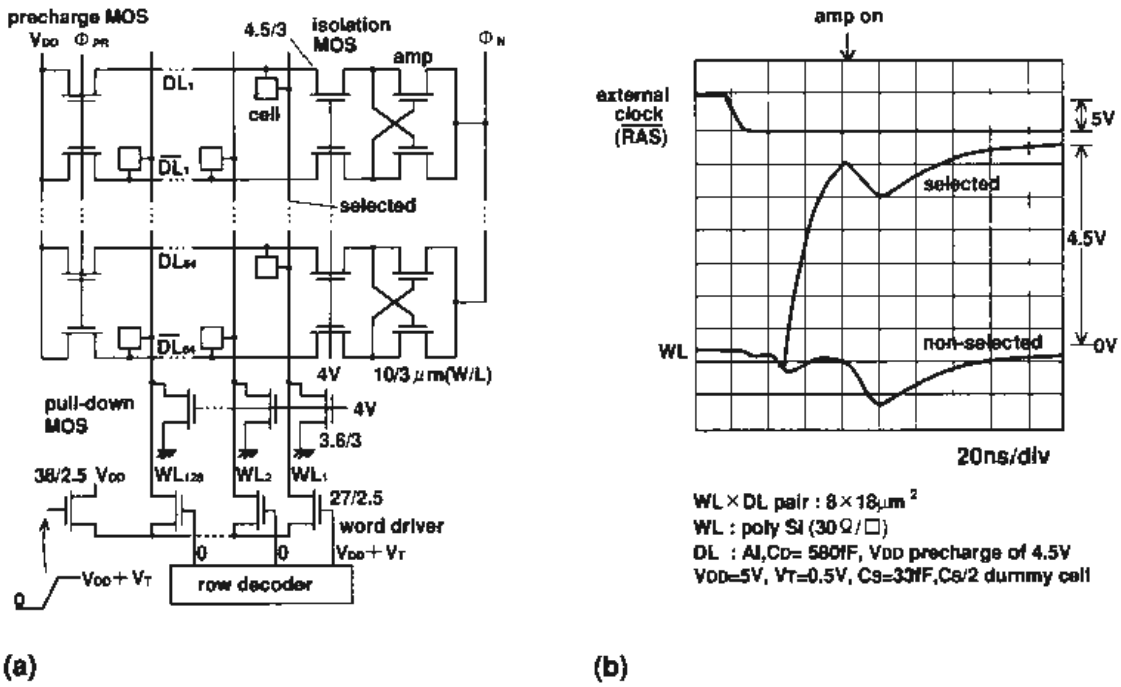


Fig. 4.33. Word-line voltage waveforms of the folded data-line array [4.1]. The memory cell shown in Fig. 3.56b was used. (**a**) 64 Kb circuit; (**b**) waveforms

Differential Noise Generation. The δ_W's generate a differential noise, δ_D, on a pair of data lines. The open data line arrangement, combined with the V_{DD} precharge shown in Fig. 4.34a [4.1], generates an increased δ_D. Furthermore, δ_D depends on the stored data pattern along the selected word line: when a cell on the selected word line that stores a high voltage ("1") is read, δ_D on the cell's data line (DL_1) depends on the number of cells along the same word line that are storing a low voltage ("0"). Thus, it is maximized in the case in which all of the remaining $m-1$ cells store "0", as explained below. The $\delta_W(L)$ coupled to each non-selected word line on the left array in the figure through C'_{DW} is given by

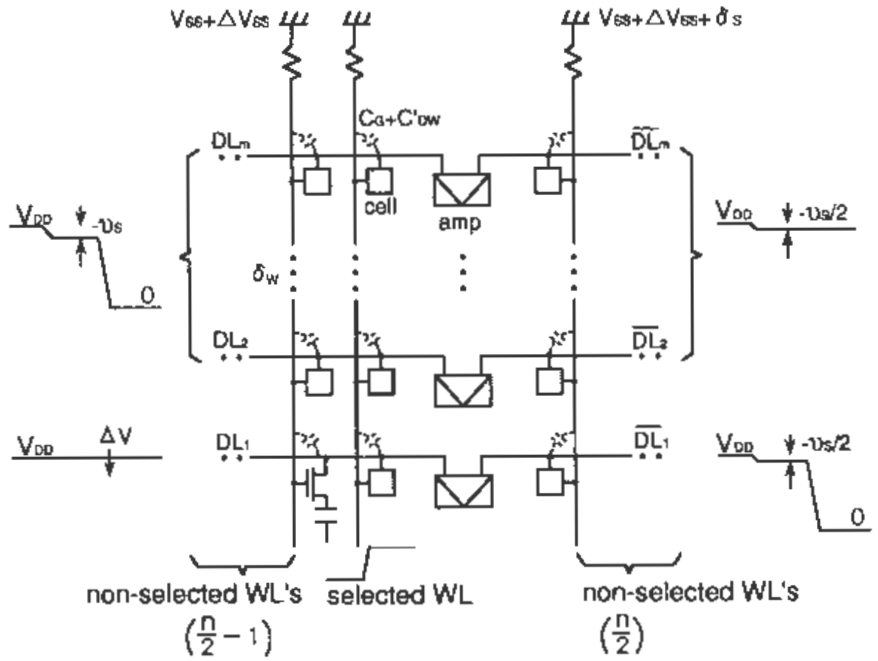
$$\delta_W(L) \simeq (m-1)A\, C'_{DW} V_{DD}. \tag{4.9}$$

Hence, the voltage coupled to DL_1 from $n/2-1$ non-selected word lines is expressed as

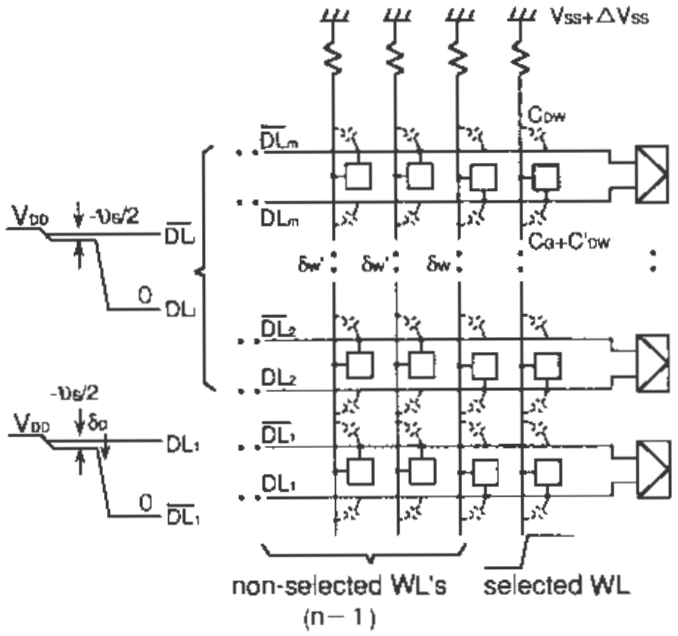
$$\Delta V(DL_1) \simeq (n/2-1)BC'_{DW}\delta_W(L), \tag{4.10}$$

where A and B are constants. Note that $\delta_W(L)$ and $\Delta V(DL_1)$ are negative, because $m-1$ data lines simultaneously discharge from $V_{DD}-v_S$ to 0 V. On the other hand, $\delta_W(R)$ on the right array is at 0 V because all of the data lines $\overline{DL_2}$–$\overline{DL_m}$ stay at $V_{DD}-v_S/2$ (the reference level with a dummy cell), causing

$$\Delta V(\overline{DL_1}) = 0.$$



(a)



(b)

Fig. 4.34. The noise-generation mechanism for V_{DD} precharge [4.1]. (**a**) Open data line; (**b**) folded data line

Thus, the differential noise voltage, $\delta_D(O)$, developed on a pair of data lines $(\overline{DL_1}, \overline{DL_1})$, is

$$\delta_D(O) = \Delta V(DL_1) - \Delta V(\overline{DL_1}) \simeq AB(mn/2)C'^2_{DW}V_{DD} \quad (m, n \gg 1). \quad (4.11)$$

Obviously, the differential voltage works as a noise because its polarity is opposite to the polarity of the signal.

For the folded data line in Fig. 4.34b [4.1], the differential noise, $\delta_D(F)$, is lowered because of cancellation at two crossing points, where a word line and a pair of data lines intersect. For the same data pattern on the selected word line as on the open data line, δ_W and δ'_W are coupled through C'_{DW} and C_{DW}, respectively. Here, C'_{DW} is the WL–DL capacitance at the crossing point where a cell is connected, while C_{DW} is the crossing-point wiring capacitance, as shown in Fig. 4.22. δ_W and δ'_W are given by

$$\delta_W \simeq (m-1)A\ C'_{DW}V_{DD}, \quad (4.12)$$

$$\delta'_W \simeq (m-1)A\ C_{DW}V_{DD}, \quad (4.13)$$

assuming that C'_{DW} and A are the same as on the open data line. Hence, the voltages coupled to the pair of data lines $(DL_1, \overline{DL_1})$, and the differential voltage, $\delta_D(F)$, are expressed as follows:

$$\begin{aligned} \Delta V(DL_1) &\simeq (n/2-1)BC'_{DW}\delta_W + (n/2)BC_{DW}\delta'_W, \\ \Delta V(\overline{DL_1}) &\simeq (n/2-1)BC_{DW}\delta_W + (n/2)BC'_{DW}\delta'_W, \\ \delta_D(F) &= \Delta V(DL_1) - \Delta V(\overline{DL_1}) \\ &\simeq AB(mn/2)(C_{DW} - C'_{DW})^2V_{DD}. \end{aligned} \quad (4.14)$$

Thus,

$$\delta_D(F)/\delta_D(O) = (\Delta C_{DW}/C'_{DW})^2, \quad (4.15)$$

$$\Delta C_{DW} = C'_{DW} - C_{DW}.$$

It is obvious that the noise of the folded data line is over one-tenth lower than that of the open data line if $\Delta C_{DW}/C'_{DW} \simeq 1/5$, as explained before. Moreover, the folded data-line arrangement, combined with half-V_{DD} precharge, almost eliminates the noise due to extremely small δ_W and δ'_W, expressed as

$$\delta_W \simeq (m-1)A\Delta C_{DW}V_{DD}/2, \quad (4.16)$$

$$\delta'_W \simeq -(m-1)A\Delta C_{DW}V_{DD}/2, \quad (4.17)$$

if the simultaneous and differential drive for both the NMOS and the PMOS amplifier, explained previously, is adopted. The open data line never realizes this reduction.

Loss of Stored Data [4.1]. A positive δ_W makes an NMOSFET of the 1-T cell weakly turn on, causing loss of information in non-selected cells. This loss is more prominent with a lower V_T, because the subthreshold current is increased, as explained in Chap. 8. A half-V_{DD} data-line precharge leads to a positive δ_W because of a charging-up operation from $V_{DD}/2$ to V_{DD}. Note

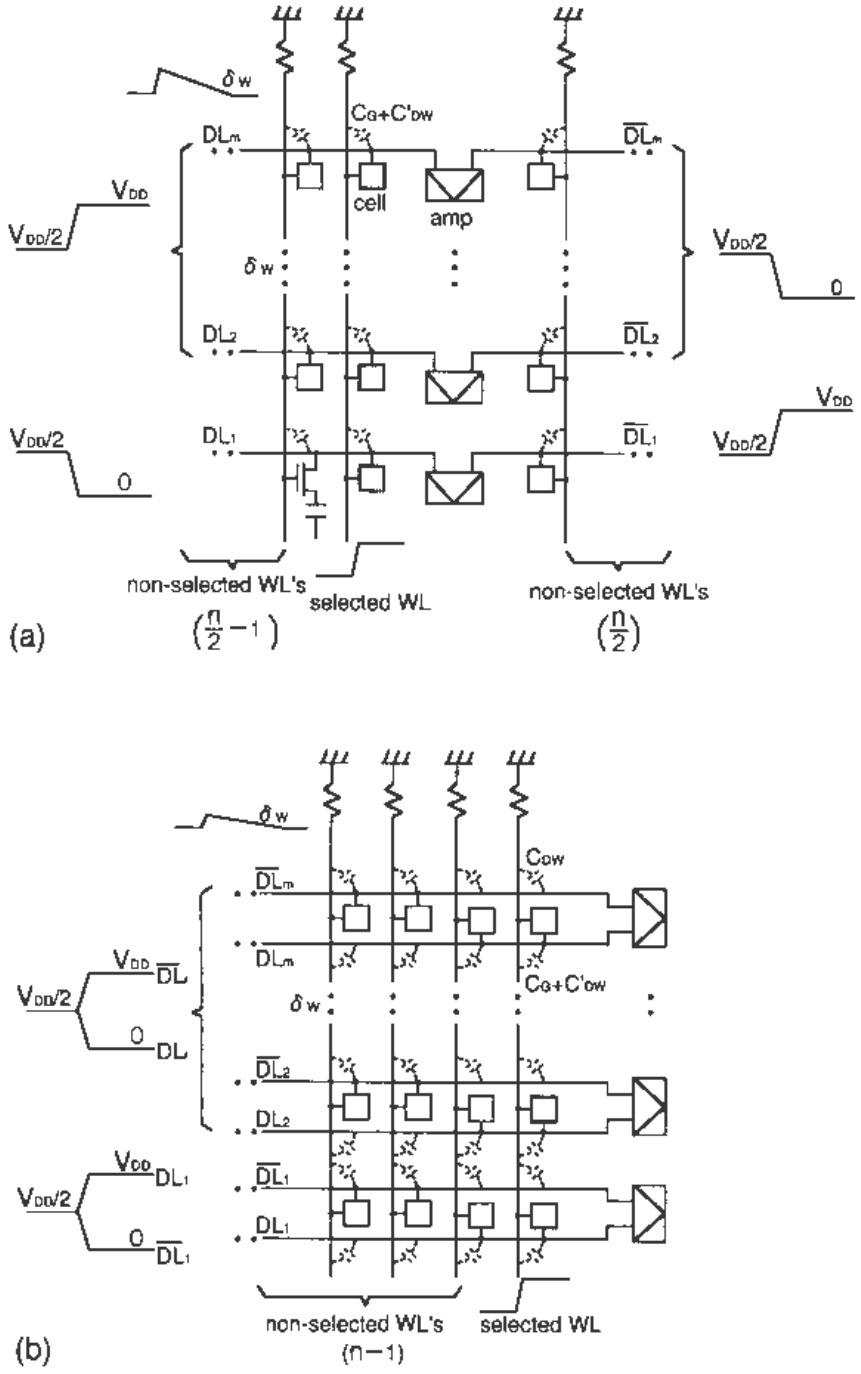


Fig. 4.35. The noise-generation mechanism for half-V_{DD} precharge [4.1]. (a) Open data line; (b) folded data line

that V_{DD} data-line precharge is quite safe, because it always makes all of the non-selected word lines negatively biased, so that the non-selected cell FETs are more deeply cut off. Figure 4.35a shows a δ_W waveform for half-V_{DD} precharge for an open data line. The simultaneous, differential drive of the NMOS and PMOS amplifiers is employed. If all cells except the cell connected to the DL_1 line along the selected word line store a high voltage, a positive δ_W (equal to half of δ_W, as expressed by (4.9)) is coupled to each non-selected word line as a result of successive activations of the cell and the amplifier. Here, consider a non-selected cell, that stores a high voltage on DL_1. Unfortunately, the stored charge of the cell escapes to DL_1 if δ_W still remains positive after DL_1 is completely discharged. On the other hand, for the folded data line shown in Fig. 4.35b, the voltage coupled to non-selected word lines is quite small, as expressed by (4.16) and (4.17), thus ensuring cell-retention characteristics. Note that the δ_W waveform depends on the time constant of the non-selected word line, and on the difference in activation time between the NMOS and PMOS amplifiers. A preceding drive by a PMOS amplifier or a shorter time constant might be hazardous, due to a larger positive δ_W.

Word Pulse Deformation. As a result of coupling from negative-going data-line waveforms to the selected word line, even a word pulse is deformed to a large extent, as shown in Fig. 4.33. This degradation makes it impossible to store a full V_{DD} in a cell, even with a boosted word-line voltage. This is because the floating word-line voltage, which is usually boosted from V_{DD} by using a charge pump, discharges down to V_{DD} due to the above coupling. The poor driving capability of the pumping circuit is responsible for this discharging. The degree of degradation depends on the data-line arrangement, the data-line precharging scheme, and the data pattern along the word line, as shown in Fig. 4.36 [4.1]. For the open data line, the V_{DD} precharge causes no degradation if all of the cells store a high voltage, because all of the data lines that intersect the word line are quiet. If they store a low voltage, however, the degradation is maximized. Note that even if the degradation occurs to the extent that the voltage drops below V_{DD}, it quickly recovers to V_{DD} which is supplied by the word driver. Even half-V_{DD} precharge, combined with the complete differential drive of the data lines, results in degradation, despite a halved swing. For the folded data line, V_{DD} precharge causes almost the same degradation as on the open data line. Half-V_{DD} precharge, however, greatly reduces degradation due to its excellent cancellation capability, as discussed previously. Thus, this scheme has been widely used in commercial chips. Even for half-V_{DD} precharge and the folded data line, if the difference in activation times between the two amplifiers is so excessive as to be larger than the word-line time constant, it may cause additional degradation due to the different noise-generation mechanism involved in non-selected word lines. Negative-going data-line waveforms, caused by the preceding operation of the NMOS amplifier, couple negative voltages (δ_W or δ'_W) to non-selected

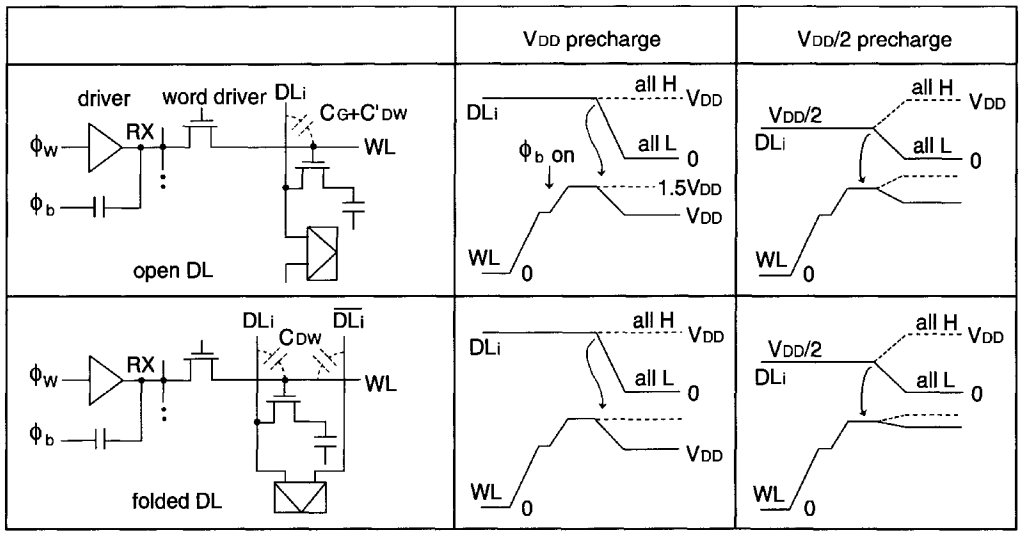


Fig. 4.36. Waveform distortions on the selected word line due to the WL–DL coupling capacitance [4.1]

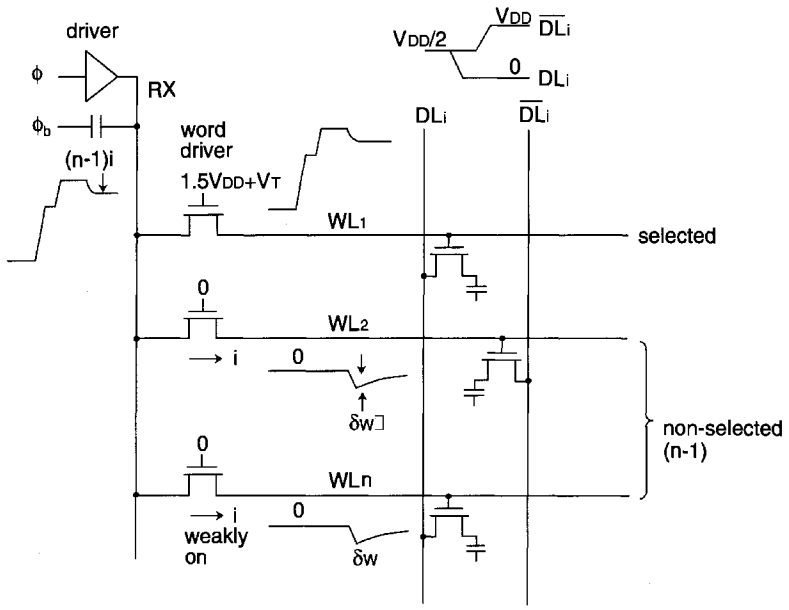


Fig. 4.37. Degradation of the boosted voltage due to the accumulation of currents flowing at non-selected word drivers [4.1]

word lines, as shown in Fig. 4.37 [4.1]. Each negative voltage may weakly turn on the driver-FET that is currently off, the gate of which is at 0 V, especially for a low-V_T FET. The resulting accumulated current, $(n-1)i$, can degrade the boosted waveform. Here, to achieve a full write operation of the cell (i.e. to store a full V_{DD}) despite the degradation, word bootstrapping after amplification would be effective. However, it would lengthen the cycle time, because of a quite slow rise-time pulse on the poly-Si word line.

4.3.5 Data-Line Interference Noise

The interference noise increases with increasing density because of the raised value of C_{DD}. There are two kinds of noise, δ_1 and δ_2 [4.22]: δ_1 is generated by intereference between small signals on the data lines before amplification; while δ_2, whose source is δ_1, is generated during amplification.

Interference Noise Before Amplification (δ_1). A signal voltage ($-v_S$) developed on the data line (DL_1) couples a noise (δ_1) through C_{DD} to the adjacent data line ($\overline{DL_1}$), which is at a floating half-V_{DD}, as shown in Fig. 4.38. Here, if the signal on DL_2 is also $-v_S$, as shown in (a), the same noise ($-\delta_1$) is coupled to $\overline{DL_1}$. Hence, the differential voltage between DL_1 and $\overline{DL_1}$ is $-v_S + 2\delta_1$, which reduces the effective signal by $2\delta_1$. If the signal on DL_2 is $+v_S$, as in (b), however, the noise ($+\delta_1$) couples to DL_1, so that it cancels the noise from $\overline{DL_1}$. Thus, the interference noise depends on data patterns on the adjacent data lines, causing the worst pattern to minimize the effective signal. The ratio, δ_1/v_S, is obtained from Fig. 4.39 [4.23] as

$$\delta_1/v_S = C_{DD}/(C_{DD} + C_1) \simeq 1/\{2 + (C_{D0}/C_{DD})\}, \tag{4.18}$$

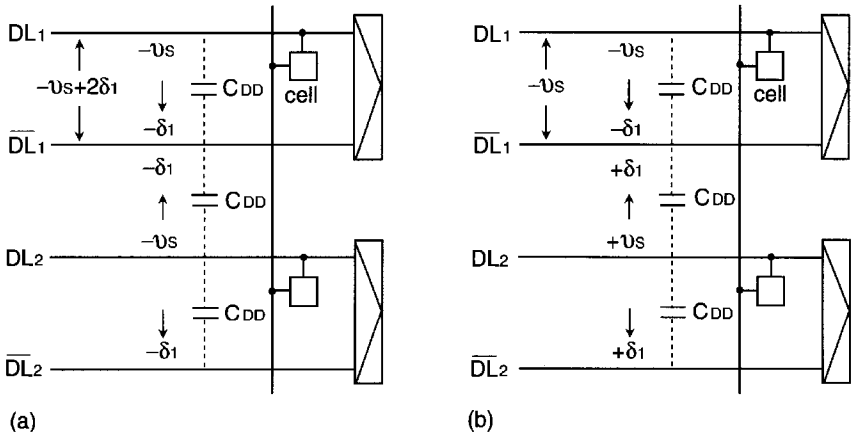


Fig. 4.38. Data-line interference noise, δ_1 [4.22]. (**a**) Same-polarity signals on DL_1 and DL_2 (worst data pattern); (**b**) opposite-polarity signals on DL_1 and DL_2 (best data pattern)

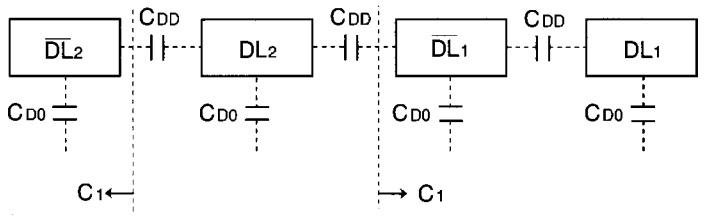


Fig. 4.39. A model for the analysis of interference noise [4.1]. C_{D0}, data line to ground capacitance; C_1, the effective capacitance seen from the C_{DD} electrode to the left or the right

because if the memory array extends to infinity,

$$C_1 = C_{D0} + 1/(1/C_{DD} + 1/C_1);$$
$$\therefore C_1 = (1/2)C_{D0}(1+X) \simeq C_{D0} + C_{DD},$$
$$X = \sqrt{1 + 4(C_{DD}/C_{D0})} \simeq 1 + 2(C_{DD}/C_{D0}).$$

Thus, δ_1 increases with higher density, which entails an increased C_{DD}.

Interference Noise During Amplification (δ_2). The read signals on the data lines are simultaneously amplified by discharging the common sources of the NMOS amplifiers shown in Fig. 4.40 from their precharge voltage (V_P) to 0 V. In this process, the difference in activation time between amplifiers, if any, would generate large noises [4.22]. One source of the difference is the above δ_1. When an activation pulse (Φ_N) is applied to the common source to amplify the signals in the worst case data pattern shown in the figure, the adjacent amplifier SA_2 starts to amplify earlier than amplifier SA_1. This is because the Q_2 gate voltage is higher, by δ_1, than the Q_1 gate voltage. As a result, a rapidly discharging voltage on DL_2 easily couples a large noise (δ_2) through C_{DD} to $\overline{DL}_1$, because the $\overline{DL}_1$ signal is still a small and floating voltage which is subject to coupling. A larger difference in activation time (i.e. larger δ_1) generates a larger noise.

The noise (δ_2) is calculated by using the previous analysis on the amplifier. The voltage change (ΔV_D) on DL_2 developed for the difference in activation time ($t_2 - t_1$) shown in Fig. 4.41 [4.22] is expressed, via (4.4) as

$$\Delta V_D = \{\beta K^2/(6C_D)\}(t_2 - t_1)^3,$$
$$\therefore \delta_2 \simeq (C_{DD}/C_D)\Delta V_D = (C_{DD}/C_D)\{\beta K^2/(6C_D)\}(t_2 - t_1)^3,$$

assuming that the additional activation delay due to δ_2 is negligible. Since $\delta_1 \simeq K(t_2 - t_1)$,

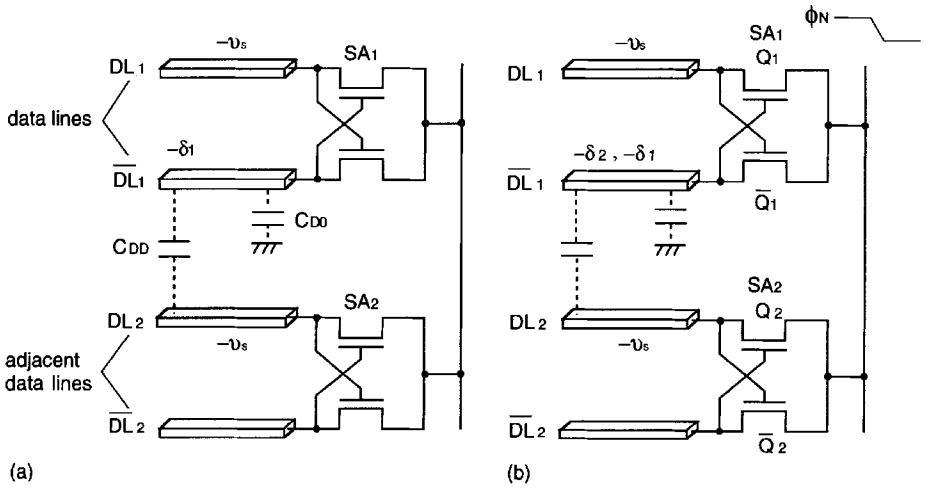


Fig. 4.40. The data-line interference noise, δ_2, triggered by δ_1 [4.22]. (**a**) When reading; (**b**) when amplifying

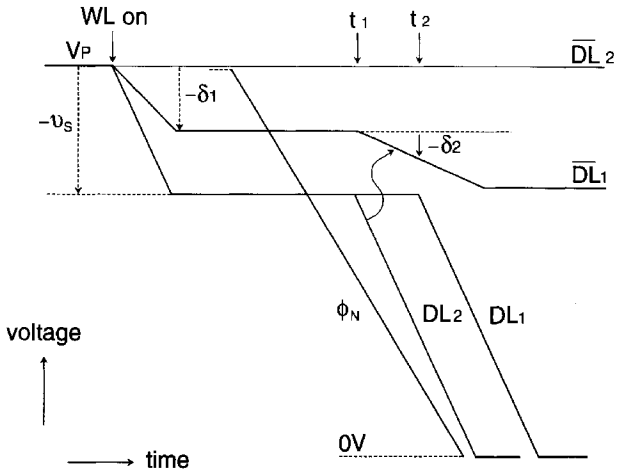


Fig. 4.41. The δ_2 generation mechanism [4.22]. V_P, precharge voltage

$$\delta_2 \simeq (C_{DD}/C_D)\{\beta/(6C_D K)\}\delta_1^3. \tag{4.19}$$

Thus, δ_2 increases rapidly with an increasing δ_1 and C_{DD}. Actually, the δ_2 generation mechanism is more complicated. In half-V_{DD} precharge, δ_2 depends on the signal polarity and the drive sequence of the NMOS and PMOS amplifiers [4.44]. Other sources of δ_2, which make the activation time differ, are an extraordinarily small signal on the specific data line, caused by a large cell-leakage current and α-particle irradiation [4.45], and an exceptionally large V_T for the specific amplifier.

Noise Reduction. The above noises are cancelled or reduced by a transposition (twist) of a pair of data lines or by a shielded data-line structure. Figure 4.42a shows a transposition at the center of data lines (DL_1, $\overline{DL_1}$). The transposition couples the same noise ($-\delta$) from an adjacent data line (DL_2) to each of a pair of data lines (DL_1, $\overline{DL_1}$), thus enabling its cancellation by the differential amplifier. However, the noise ($-\delta_1'$) coupled from $\overline{DL_1}$ to DL_2 is never cancelled, although it is halved by the transposition. It is cancelled by an alternate transposition [4.22, 4.23] shown in Fig. 4.42b. This

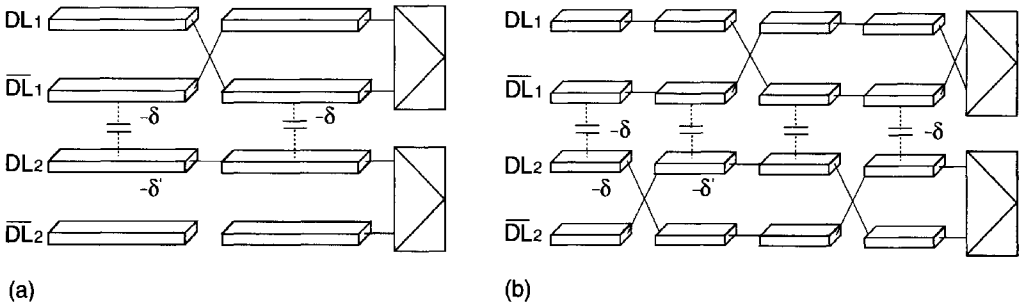


Fig. 4.42. Data-line transpositions [4.22]. (**a**) The principle; (**b**) alternate transposition

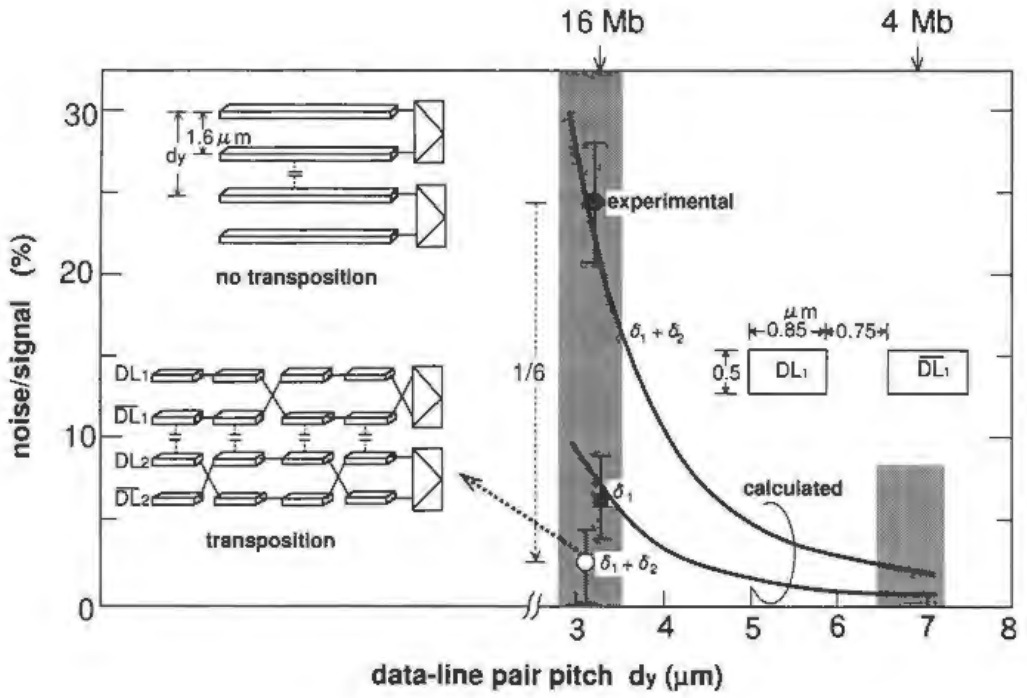


Fig. 4.43. Interference noise versus data-line pair pitch [4.22]. The memory cell in Fig. 3.56b was used

transposition is effective for both δ_1 and δ_2. Figure 4.43 shows experiments on noise reduction [4.22]. Both δ_1 and δ_2 increase rapidly with decreasing distance from an adjacent pair of data lines, reaching a total noise that is as large as one-fourth of the signal voltage for a 16 Mb chip, using a pair pitch of 3.2 μm. At the 3.2 μm pitch, δ_2 is about three times larger than δ_1. The total noise is reduced to one-sixth by transposition. The residual noise comes from an incomplete layout of the amplifier. However, this transposition is not effective for reducing the interference between a pair of data lines, which also reduces the differential signal component. In addition, the transposition requires an additional area. Consequently, a shielded data-line structure is desirable. In fact, it has been reported [4.46] that the structure actually reduces the noise to 7% of the signal voltage. Note that, in principle, the open data-line arrangement never accepts transposition.

4.3.6 Power-Supply Voltage Bounce

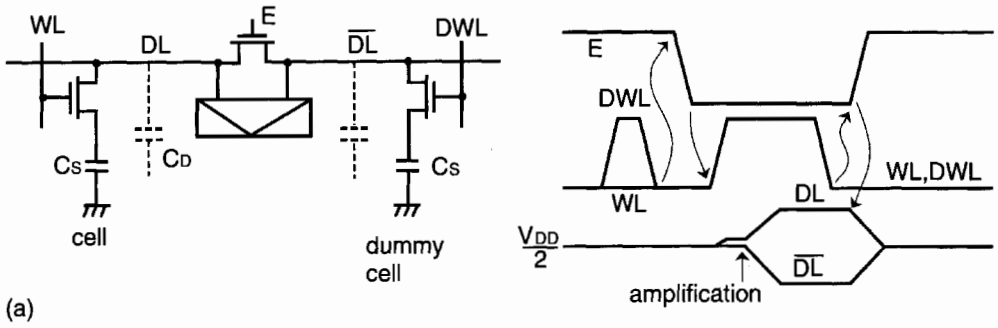
Each non-selected word line is always fixed to ground ($V_{SS} = 0$ V) through a low-impedance pull-down circuit, as discussed in Chap. 3. However, the V_{SS} line may bounce whenever peripheral circuits on the chip operate at random at large spike currents, since the V_{SS} is usually common to those of the peripheral circuits. Thus, the resultanting V_{SS} bounce may couple a voltage to each data line at a floating state just before amplification through C_{DW}

and C'_{DW}. The total C_{DW} and C'_{DW} component along a data line is as large as 32–54% of C_D, and thus quite a large differential noise may couple to a pair of data lines, especially for an open data line, as shown in Fig. 4.34 where a V_{SS} bounce (ΔV_{SS}) is assumed. For an open data line, a ΔV_{SS} difference, δ_S, between the left and right subarrays can exist. This is justified by the fact that the long V_{SS} line that runs along pairs of data lines with a certain impedance, despite the Al wiring, develops a different voltage between two V_{SS}-lines of the subarrays. The resultant differential noise of the open data line equals $\delta_S(nC'_{DW}/2)/\{C_D + (nC'_{DW}/2)\}$, which is as large as 23 mV for $\delta_S = 100$ mV, $n = 128$, $C_D = 300$ fF, and $C'_{DW} = 1.41$ fF. Thus, a more careful V_{SS} line layout becomes necessary as the physical array size increases. The folded data line, however, locally cancels the noises coupled to a pair of data lines, independently of ΔV_{SS}.

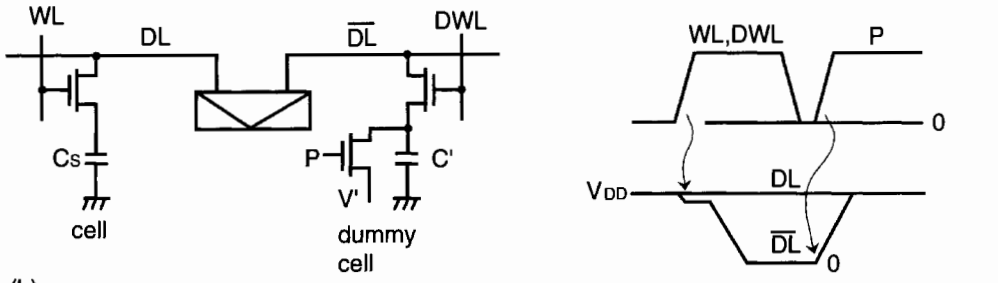
4.3.7 Variation in the Reference Voltage

It is essential for the data-line reference voltage to be accurate against variations of voltage, temperature, and the fabrication process. Its inaccurate setting effectively causes a noise. The dummy cell, which is a kind of reference-voltage generator, plays an important role in generating an accurate reference voltage. In addition, the cell can cancel the word-line drive noise and equalize the electrical characteristics of a pair of data lines.

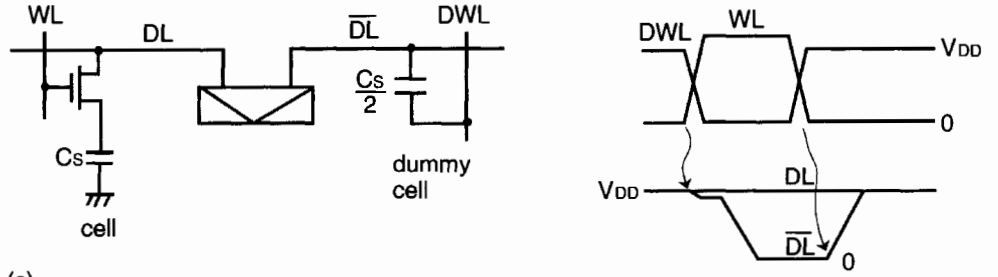
Figure 4.44 shows typical dummy-cell circuits that have been the proposed. Some of the original proposals are slightly modified for ease of understanding. Circuit (a) [4.25] is for half-V_{DD} precharge and the open data line. After storing a half-V_{DD} in the dummy cell and turning off the equalizer FET, the word and dummy word pulses are applied simultaneously. The circuit achieves reduction of word-line drive noise, accurate setting of the reference voltage, and a well-balanced capacitance between a pair of data lines during amplification. However, an additional timing storing a half-V_{DD} in the dummy cell requires a complicated circuit design and a longer cycle time. Circuit (b) is for V_{DD} precharge and an open data line. To obtain a reference level on the data line (DL), there are two combinations. One is $V' = 0$ and $C' = C_S/2$. The other is $V' = V_{DD}/2$ and $C' = C_S$. Both cancel the word-line drive noise. The latter even equalizes the data-line capacitances. For the former, however, there can be different capacitance-variation ratios, caused by different sizes between the cell and dummy-cell capacitors for a given variation in the fabrication process. They eventually create reference-voltage variations on the data line (DL). The latter also suffers from reference-voltage variations caused by a characteristic mismatch between V_{DD} and a half-V_{DD} from an on-chip generator. Circuit (c) is also for V_{DD} precharge and an open data line [4.47]. The dummy cell is just one coupling capacitor, of $C_S/2$. The dummy word line is discharged to 0 V just before cell activation, which differs from the word line in terms of polarity and activation time. Thus, word-line drive noise is never cancelled, but it is the sum of the noises at two crossing points



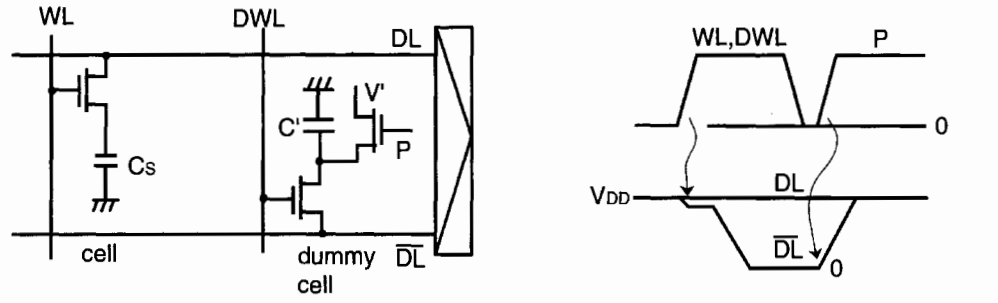
(a)



(b)



(c)



(d)

Fig. 4.44. Various dummy cells [4.1]. $V_{\mathrm{W}} \geq V_{\mathrm{DD}} + V_{\mathrm{T}}$

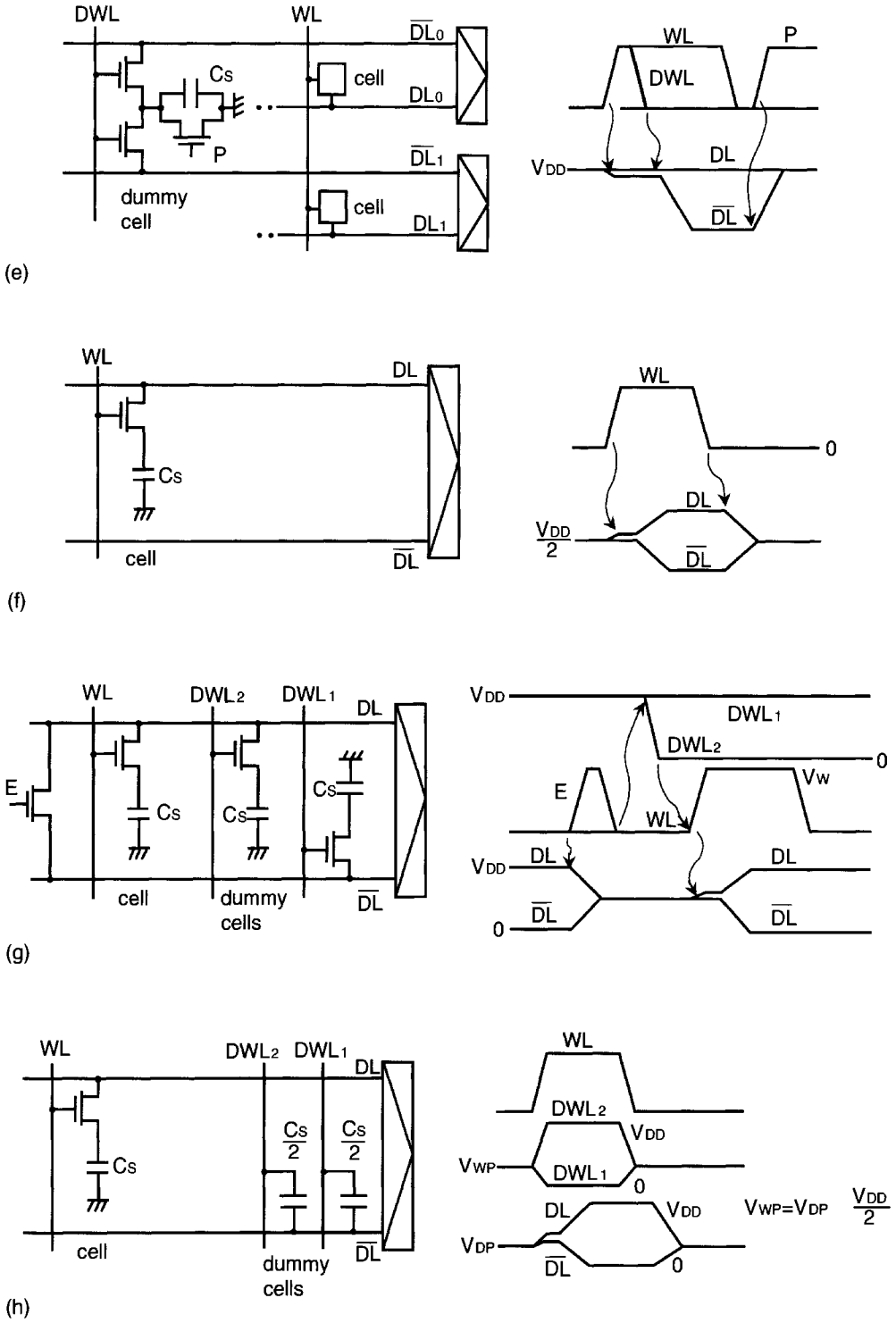


Fig. 4.44. Continued

(i.e. cell and word line, and dummy cell and dummy word line). It also suffers from the drawbacks of reference-voltage variation and a residual capacitive imbalance due to $C_S/2$ between the data lines. Circuit (d) is an application of circuit (b) to the folded data line. Circuit (e) is for V_{DD} precharge and a folded data line [4.48]. The dummy-cell capacitance (C_S) shared with two data lines generates a reference level, with a resulting doubled C_D. However, a differential switching noise is generated just before amplification, when two data lines are isolated from the dummy cell. Moreover, the dummy-cell control slows the access time. Circuit (f) is for half-V_{DD} and a folded data line [4.49, 4.50] without a dummy cell, which has been widely used in modern DRAMs. A half-V_{DD} itself is utilized as a reference level. The drawbacks are residual word-line drive noise and a data-line capacitive imbalance due to C_S, as discussed before, despite the simple circuit design, the reduction in area due to the elimination of the dummy cell, and an excellent setting capability for the reference level. Circuit (g) is for quasi-half-V_{DD} precharge and the folded data line [4.51]. For the precharge period, either of a pair of data lines is fixed at V_{DD} or 0 V, depending on the data of the previous cycle, which differs from the conventional precharging scheme. Just after precharging, both data lines and two dummy-cell nodes are equalized to a floating half-V_{DD} by activating the equalizer (E). The subsequent activation of the word line (WL) with the turning on of one dummy word line (DW_1) enables equalization of the data-line capacitances during amplification. In addition, it would cancel the word-line drive noise with the help of simultaneous discharging of the other dummy word line (DW_2) if the pulse swings of WL and DW_2 were the same. Actually, however,V_{DD} is applied to the dummy word lines, causing residual word-line drive noise. This is because the floating voltage of the dummy word line, boosted by a capacitive coupling, eventually degrades to V_{DD} for the maximum precharge period of 10−20 μs in the catalog specification. The process of equalization and dummy word-line activation prior to word-line activation slow the access time. Circuit (h) is for half-V_{DD} and folded data line [4.52]. The data-line capacitive imbalance during amplification is reduced to some extent despite inaccurate reference-voltage setting due to variation in the fabrication process, incomplete cancellation of the word-line drive noise due to the cell FET, and the need for four dummy word lines.

4.3.8 Other Noises

The voltage bump of the chip may degrade the signal voltage. For example, for the folded data line combined with V_{DD} data-line precharge and V_{DD} word-line activation, a low to high V_{DD} bump after the write operation posed a serious problem in the 16 Kb and 64 Kb generations, where the V_{DD} cell electrode had been popular. A voltage bump (ΔV_D) during the precharge period after write causes a bump ($\Delta V_E = a\Delta V_D$, where a is a constant) in the cell-electrode voltage (V_E), as shown in Fig. 4.45 [4.1]. Thus, the storage-node (N) voltages of non-selected cells are raised by about ΔV_E. However, the

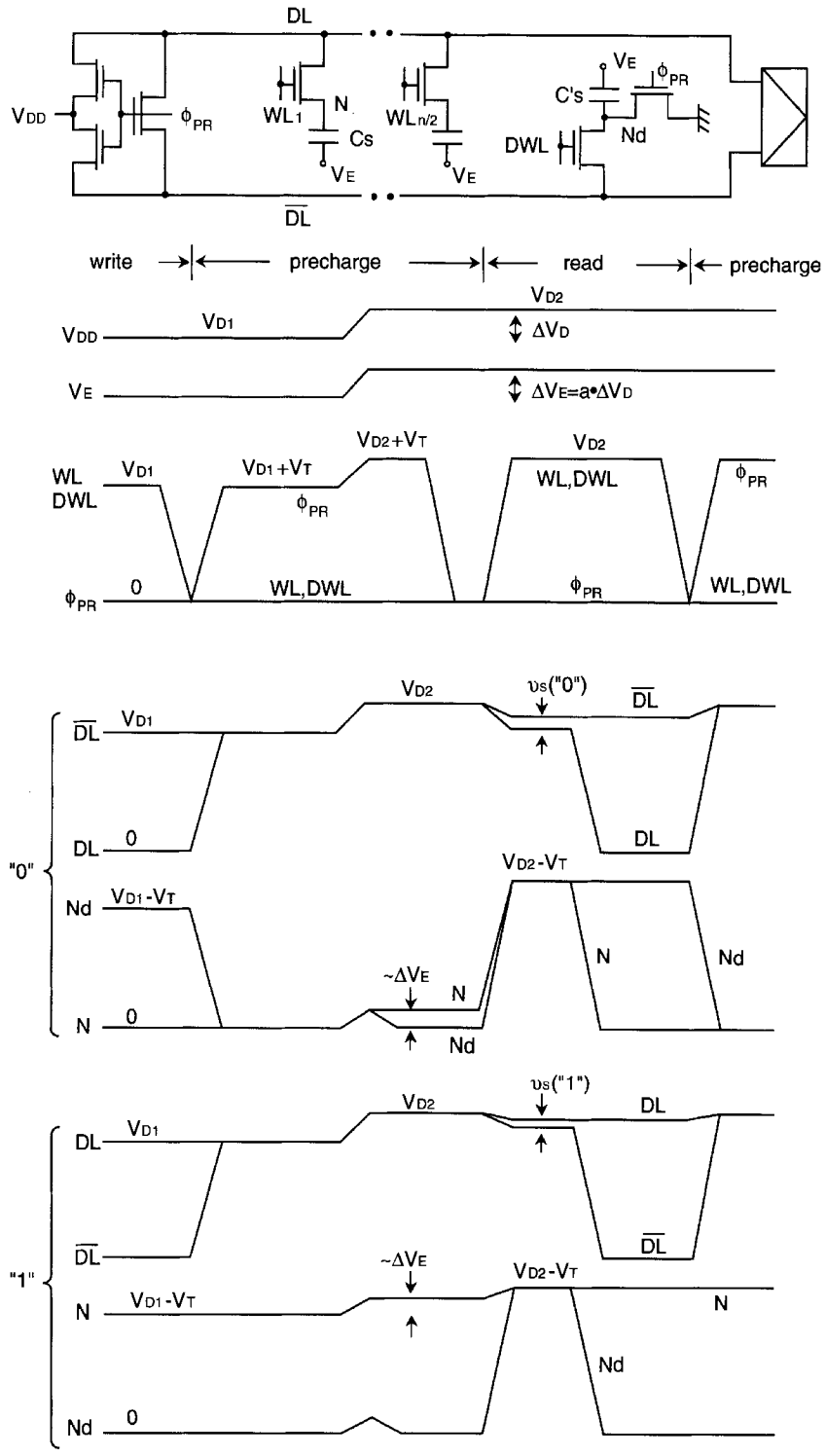


Fig. 4.45. Various node voltages under a voltage bump [4.1]. V_{DD} data-line precharge and no word bootstrapping are assumed

node (Nd) voltage of a dummy cell is always fixed at 0 V by an Nd precharge circuit. The difference in the voltage change between the cell node and the dummy-cell node reduces the signal voltage developed on the data line, as explained below.

When a cell that is storing 0 V is read, the signal voltage, v_S ("0"), is the difference between the voltage changes on a pair of data lines, $\Delta v(\mathrm{DL})$ and $\Delta v(\overline{\mathrm{DL}})$. Thus,

$$\begin{aligned} v_S(\text{"0"}) &= \Delta v(\mathrm{DL}) - \Delta v(\overline{\mathrm{DL}}) \\ &\simeq (C_S/C_D)(V_{D2} - V_T - \Delta V_E) - (C'_S/C_D)(V_{D2} - V_T). \end{aligned}$$

The signal voltage, v_S("1"), when a cell that is storing $V_{D2} - V_T$ is read, is also expressed as

$$\begin{aligned} v_S(\text{"1"}) &= \Delta v(\overline{\mathrm{DL}}) - \Delta v(\mathrm{DL}) \\ &\simeq (C'_S/C_D)(V_{D2} - V_T) - (C_S/C_D)(V_D - V_E). \end{aligned}$$

Figure 4.46 shows the relationship between the signals and coefficient a for $\Delta V_D = 1$ V [4.1]. For $a = 0$ the S/N ratio is determined by v_S("1") because v_S("0") > v_S("1"). The condition $a = 0$ is realized by applying 0 V to the electrode, or by applying a dc voltage [4.53], raised by an on-chip generator so as to be always constant against any V_{DD} variation, to the electrode. For $a = 1$ – that is, for the application of V_{DD} to the electrode (i.e. the cell-electrode V_{DD}) – the voltage margin for v_S("0") is worse because v_S("0") < v_S("1"). For $a = 0.5$ – that is, for a half-V_{DD} electrode – both voltage margins are equalized.

A half-V_{DD} electrode combined with half-V_{DD} data-line precharge [4.54, 4.55] provides excellent signal characteristics that are immune to any V_{DD}

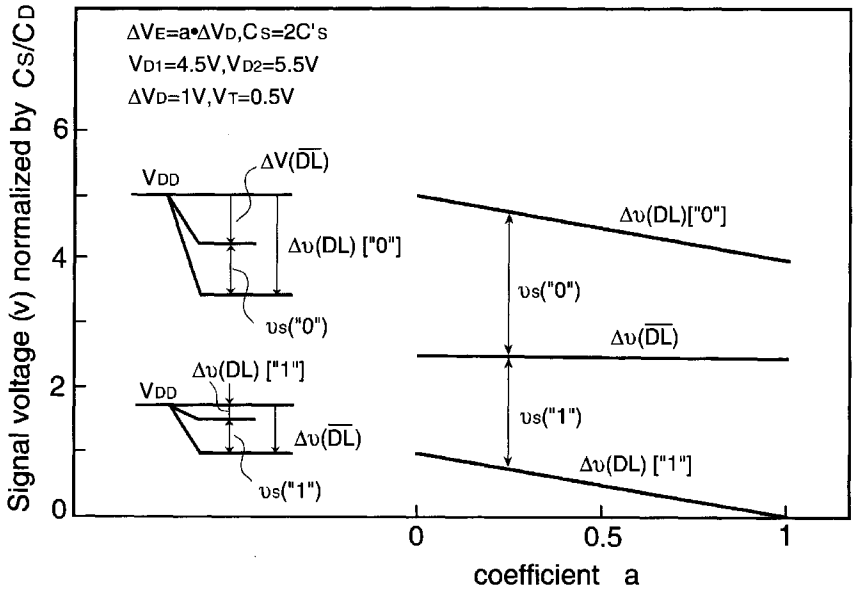


Fig. 4.46. Voltage bump versus signal voltage [4.1]

bump, if half-V_{DD}'s for the data line and electrode are supplied from a common voltage generator, and the word-line voltage is fully bootstrapped. The resultant data-line precharged voltage, which always equals the electrode voltage despite a V_{DD} bump, avoids signal component degradation, with a fixed difference between the signal voltage and the reference voltage. Even in this case, however, the poor driving capability of an on-chip half-V_{DD} generator and its heavy capacitive load (i.e. an extremely large cell-electrode capacitance of the array) may cause two different reference voltages. One is the voltage supplied from a half-V_{DD} generator. This occurs when the precharge period, together with the turning on of the precharge circuit, is long enough for the generator to be able to manage the heavy load voltage. In this case the reference voltage of the data line and the electrode voltage are same, even when V_{DD} occurs during the precharge period. The other is the voltage obtained by equalizing two amplified voltages (i.e. V_{DD} and 0 V) on a pair of data lines every cycle. This occurs when the cycle time is fast enough for the generator to be unable to manage the load voltage. In this case, the reference voltage can differ from the electrode voltage when the bump occurs. Thus, two different reference voltages could be developed by a V_{DD} bump, causing a signal-voltage reduction. The electrode voltage is bumped not only by a V_{DD} bump, but also by various capacitive couplings from other conductors, such as data lines and the substrate during amplification. In principle, the quiet array offered by half-V_{DD} precharge and the folded data-line arrangement eliminates these coupling noises.

Incomplete equalization of a pair of data lines would be another noise source. A sufficiently short equalization time results in residual differential noise, which depends on the write or rewrite data of the previous cycle. Although it is not directly related to the cell S/N ratio, a large number of coupling capacitances between data lines and a pair of I/O lines degrades the stable operation of the main differential amplifier on the I/O lines. A large voltage swing on each data line during amplification couples a voltage on either of the I/O lines through the capacitance. The accumulated voltage works as a noise for the main amplifier. This noise depends on the layout, and is cancelled by transposition of a pair of I/O lines.

4.4 Summary

The high S/N ratio design and technology of the 1-T cell can be summarized as follows. The need for the sophisticated technologies described above stems from the drawbacks of the 1-T cell (i.e. no gain and the existence of a p–n junction), as previously discusssed. The cell S/N ratio will be worsened in the multigigabit era, when the signal voltage will decrease toward 100 mV even with a high-ε dielectric film, as shown in Fig. 4.3, while the total (extrinsic and intrinsic) sense-amplifier offset voltage will increase toward 100 mV. Since there are other signal-voltage degradation components, such as the leakage

current, and other noise components, as discussed so far, S/N design will become tougher in the future. In any event, the key technologies are the development of a large-signal-charge technology – such as high-permittivity and thin dielectric materials for the cell capacitor – the reduction of the sense-amplifier offset voltages through process, device, and circuit innovations, and the realization of a quiet array, that is offered by the folded data-line arrangement combined with half-V_{DD} precharge.

5. On-Chip Voltage Generators

5.1 Introduction

On-chip voltage generation [5.1–5.8] is becoming increasingly important for memory LSI design, as well as for other LSI designs. This importance is being accelerated by a recent trend toward lower-voltage operation. In the past, generators have been widely used in commercial memory chips such as DRAMs and Flash memories, as explained in Chap. 1. For example, DRAM chips have needed various kinds of power-supply voltages, which have been generated internally by using a single external supply-voltage (V_{DD}), as shown in Fig. 5.1. A negative voltage (V_{BB}) is a substrate bias voltage supplied to an NMOS memory-cell array to ensure stable memory-cell operation. A boosted dc voltage (V_{DH} or V_{PP}) is for word bootstrapping, to eliminate the drop in the cell V_T. A half-V_{DD} (or V_{DL}) achieves a half-V_{DD} (or V_{DL}) data-line precharge for power reduction, and a half-V_{DD} (or V_{DL}) cell-capacitor plate for high S/N ratio design, as explained in Chaps. 3 and 4. Modern commercial DRAMs of 16 Mb and beyond even incorporate an on-chip voltage down-converter, which lowers V_{DD} to V_{DL}, by using an internal reference voltage (V_{REF}), to simultaneously achieve high reliability of small devices and power-supply standardization. Flash memories have also used an extremely boosted word-line voltage, and even a negative voltage, to relaxing a stress voltage to a memory-cell FET, as explained in Chap. 1. Schematic circuits for generating these internal voltages are shown in Fig. 5.2. The negative voltage (V_{BB}) and the boosted voltage (V_{DH}) are both generated by using a charge-pumping circuit that consists of a capacitor and diodes working as rectifiers. Charges are injected into the almost capacitive load by pumping the capacitor. The lowered voltage (V_{DL}) is usually generated by using a differential amplifier (op amp), which works as a comparator, and a PMOSFET. V_{DL} is set to be almost equal to an internally generated reference voltage (V_{REF}). In principle, a half-V_{DD} is generated by a voltage division of V_{DD}. For these generators, the load-current delivering capability, output-voltage stability, voltage-setting accuracy, and low power consumption are major design concerns.

In all of the low-voltage CMOS LSI designs of the future, various internal voltages will be indispensable to reduce the subthreshold current that exponentially increases with a decreasing V_T, as discussed in Chap. 8. In fact, for low-voltage logic-oriented CMOS LSIs dynamic V_{BB} controls [5.9,

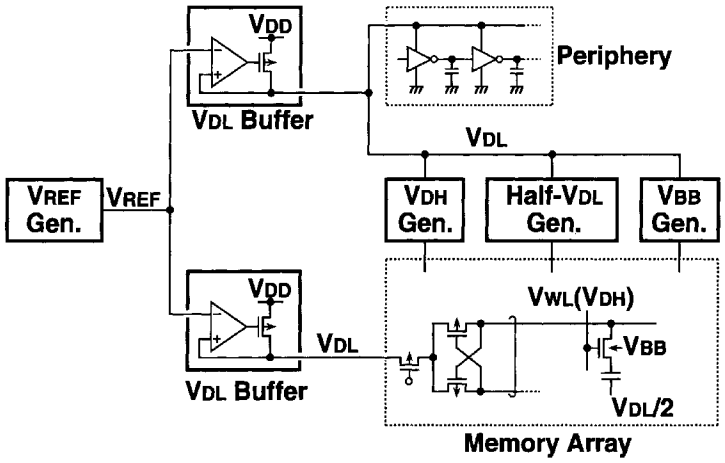


Fig. 5.1. Internal supply voltages for modern DRAMs [5.1, 5.73]

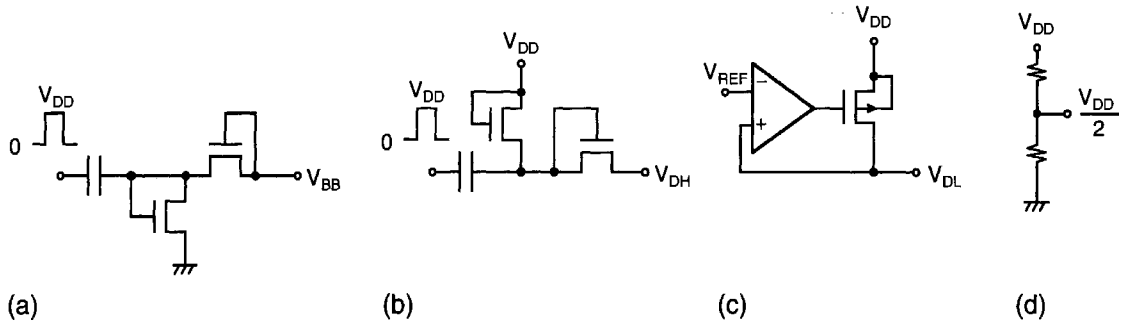


Fig. 5.2. Schematic circuits for a V_{BB} generator (**a**), a V_{DH} generator (**b**), a V_{DL} generator (**c**), and a half-V_{DD} generator (**d**)

5.77, 5.78] through the use of internal voltages have been studied intensively. They control V_T dynamically by changing the V_{BB} of the floating substrate. However, possible problems are expected, since logic LSIs have never used a floating substrate. Thus, the past experiences of DRAM designers who have managed the problems involved in the use of a floating substrate would be instructive to logic LSI designers. In any case, in the future, on-chip voltage generation will be essential not only for memory designers, but also for logic designers.

This chapter describes on-chip voltage generators. First, V_{BB} generators for DRAMs with a floating substrate are discussed. The power-on characteristics, coupling noise, substrate-current generation, substrate (or well) structure, and circuit configuration are the main topics for discussion. Second, voltage up-converters (i.e. V_{DH} generators) are described from the viewpoint of circuit configuration. Third, voltage down-converters (i.e. V_{DL} generators) are explained in terms of their basic design concept, loop stability, reference-voltage stability, burn-in test capability, and voltage trimming. Fourth, half-V_{DD} generators are briefly discussed. Finally, one of the most advanced on-chip generators is explained, as a summary.

5.2 The Substrate-Bias Voltage (V_{BB}) Generator

5.2.1 The Roles of the V_{BB} generator

The application of a substrate bias (V_{BB}) to DRAM chips has changed, depending on the circuit and device technology. Until the 5 μm NMOS DRAM generation, a V_{BB} of -5 V was externally supplied to a p-type substrate. After that, to achieve single-V_{DD} DRAMs, the V_{BB} has been internally supplied by an on-chip V_{BB} generator. Meanwhile, there have been advances in the substrate structure and the application of V_{BB} in accordance with a technology transition from NMOS to CMOS. In the NMOS DRAM era up to the 256 Kb generation, both the memory array and the peripheral circuit were integrated on to one substrate of a chip, and V_{BB} was supplied from the edge of the chip to the substrate through bonding pads or an on-chip V_{BB} generator. Therefore, each substrate of the NMOSFETs in the whole chip was common and was just the same as that of the chip. Even in the succeeding CMOS DRAM era, the memory array has been kept the same (i.e. an NMOS memory array biased at a negative V_{BB}). As for the peripheral circuit, however, the back-biasing scheme has changed. Up to around 16 Mb generation, the NMOSFETs were back-biased at the same V_{BB} as that for the NMOSFETs in the memory array, while the PMOSFETs had no back-biasing (see Fig. 5.11). This was a result of utilizing a double-well structure. Since around the 64 Mb generation, however, each NMOSFET or PMOSFET has had its own substrate, without substrate biasing, by utilizing a triple-well structure. Here, the different biasing schemes between the memory array and the peripheral circuit derives from their different features, as discussed below.

The following is a summary of the main reasons why the V_{BB} generator is needed.

Realization of Stable Memory-Cell Operation. The application of V_{BB} to the NMOS memory array ensures stable memory-cell operation with the highest density, as follows.

Prevention of Memory-Cell FET from being Forward-Biased. To realize a high density, a memory array must consist of single-channel MOSFETs (NMOS or PMOS), although the use of NMOSFETs has been popular. In addition, the substrates of the NMOSFETs must be common and connected to V_{BB} at the edge of the memory array. This can create a voltage difference between the substrate and the source of an NMOSFET in the memory array. In particular, the difference becomes largest when substrate bounces occur, as a result of charging and discharging a huge number of data lines. Since the voltage difference is just the substrate-bias voltage of the NMOSFET, the NMOS characteristics would be dynamically degraded. This degradation can be explained by using Fig. 5.3 [5.10]. Figure 5.3 shows V_T versus V_{BB} for an NMOSFET in a CMOS structure as a parameter of the substrate (i.e. well) doping concentration (N). Obviously, the change in V_T caused by

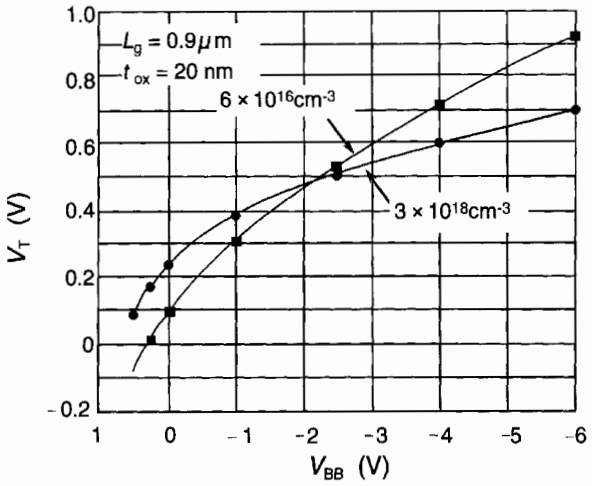


Fig. 5.3. The substrate bias voltage (V_{BB}) versus the threshold voltage (V_T) for a NMOSFET as a parameter of the well doping-concentration [5.10]

a given V_{BB} bounce increases with a shallower V_{BB} setting, and it is enhanced with an increasing N. If the substrate voltage is set to be the same as the source voltage (i.e. $V_{BB} = 0$), V_T is very sensitive to the V_{BB} bounce. Even a small V_{BB} noise would make the NMOSFET forward-biased, and cause a considerable decrease in V_T, so that stored charges in a memory cell easily escape to the data line. This tendency can be also explained by the following expression, which is obtained by differentiating (2.6) with respect to V_{BB}:

$$\Delta V_T / \Delta V_{BB} = (K/2)/\sqrt{|V_{BB}| + 2\Psi}. \tag{5.1}$$

Thus, the substrate must be deeply negatively biased for the NMOS memory array, so that the change in V_T (ΔV_T) caused by the V_{BB} noise (ΔV_{BB}) becomes smaller. This requirement must be maintained for the memory array in a CMOS DRAM chip.

The need to apply V_{BB} to the memory array has been experimentally proven with a 256 Kb CMOS DRAM [5.27] shown in Fig. 5.4, although the chip incorporated PMOS memory cells storing V_{DD} or 0 V. No V_{BB} generator was used. To realize a latch-up immune structure, an n-well layer and an epitaxial layer are formed on a low-resistane, high-impurity p-substrate. V_{SS} (0 V) and V_{DD} (5 V) are applied to the p-substrate and the n-well, respectively. Thus, the sources of all PMOSFETs in the memory array and the peripheral circuit are at V_{DD}, while those of all NMOSFETs are at 0 V. A parasitic transistor (Q_1) is never turned on, because the base is fixed well at 0 V throughout the chip, due to the low-resistane p-substrate. The n-well surrounding the memory array can act as a barrier to protect the memory cells against minority carriers injected into the p-epitaxial substrate. In this scheme, however, local V_{DD} fluctuations tend to be created in the large-sized and resistive n-well layer, while amplification and precharging are performed with a large voltage swing of V_{DD}. Consequently, if negative-going

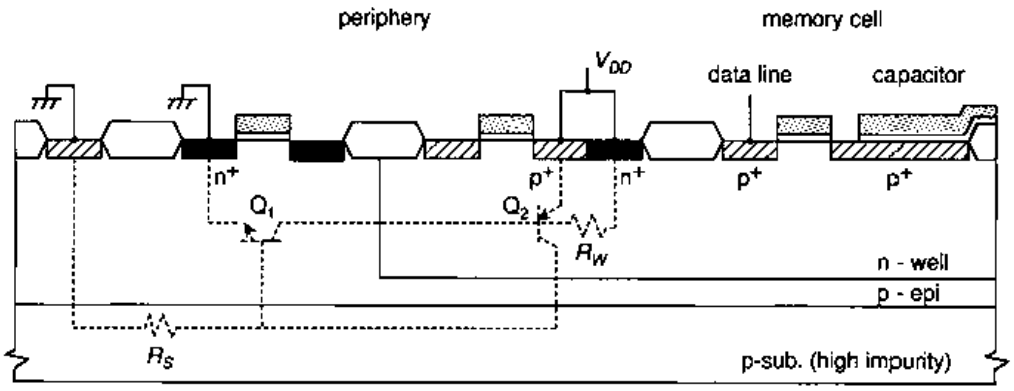


Fig. 5.4. A substrate structure and biasing scheme for a 256 Kb CMOS DRAM [5.27]

pulses are coupled to the n-well through a large number of p–n junctions at data lines, the PMOSFET in a memory cell may be forward-biased, so that V_T is instantaneously small and charges stored in the cell escape to the data line. Moreover, the p–n junctions formed by the n-well and the p^+-V_{DD}-terminals at the rim of memory array are also forward-biased, injecting minority carriers (i.e. holes) so as to attack the cells, as explained later. It has been reported [5.27] that to solve the problems the forward bias must be suppressed to less than 0.1 V. This requires careful attention to layout, such as n-well strapping with low-resistivity metal wiring throughout the n-well layer. Eventually, the biasing scheme was changed to boosting of the n-well voltage over V_{DD} using a well bias generator [5.28].

Other Advantages. The application of V_{BB} has other advantages, of less leakage current in the cell and improved MOSFET characteristics. V_{BB} raises the threshold voltages of parasitic MOSFETs in the memory array. Thus, the impurity concentration of the channel stopper can be reduced, so that the peripheral component of the p–n junction leakage current of the cell storage node is reduced. Moreover, the resulting improved narrow-channel effect and small substrate-bias effect of the cell FET suppresses not only the increase in V_T in a region with a narrower channel, but also variation in V_T. This eventually lowers the necessary word-line voltage. Furthermore, V_{BB} reduces p–n junction capacitance, thus reducing the data-line capacitance of the 1-T cell. It has been reported [5.8] that for an open data-line arrangement, in which the data line is composed of a diffused layer, signal voltage on the data line was increased by 8% by changing V_{BB} from 0 V to −2.25 V.

Realization of Stable Peripheral-Circuit Operations. Application of V_{BB} to the peripheral circuit of DRAMs is also effective for its stable operation, especially for NMOS DRAMs.

Prevention of NMOSFETs from being Forward-Biased. In NMOS DRAMs, the source-to-substrate voltage of each NMOSFET can fluctuate locally due

to the internal noises, especially the V_{BB} noise coming from the memory array. As a result, an NMOSFET can be forward biased, because the substrate is common to all MOSFETs, causing a large variation in V_T, as in the memory array. The application of V_{BB} avoids this forward biasing. In addition, a deeper V_{BB} biasing of the I/O circuit accepts a larger undershoot of the input signal coming from the package input pin, because the p–n junctions in a gate-protection diode or a diffused resistor are more deeply back-biased. Note that if the undershoot of an input pulse exceeds the sum of $|V_{BB}| + V_f$ (i.e. forward drop of the p–n diode), the V_{BB}-level is degraded, because the diode forward current is rather larger than the V_{BB} generator current. On the other hand, in CMOS DRAMs the peripheral circuit does not necessarily need application of V_{BB}. This is because the conventional CMOS circuit allows the source of each MOSFET to be fixed, at V_{DD} for PMOSFETs and at V_{SS} (0 V) for NMOSFETs. Moreover, it has traditionally enabled a tight connection between the source and the well at each MOSFET, so as not to create a source–well voltage difference, thus ensuring a fixed V_{BB} of 0 V throughout the chip. The undershoot issue is also solved as long as a well-fixed ground level is ensured by careful layout. Forward biasing must be avoided, even for a recent low-voltage CMOS design [5.11], in which a local V_{BB} was applied to specific wells in the peripheral CMOS circuit to suppress the subthreshold current, with a resulting increased V_T, as discussed in Chap. 8.

Minority-Carrier Reduction. If electrons (minority carriers) are injected from a certain point of the peripheral circuit into the p-type substrate, they diffuse far away and reach memory cells, because the diffusion length of electrons in typical silicon substrates is greater than 100 μm [5.20]. Then, the electrons are trapped by the potential well of a memory cell that is storing a high voltage, causing loss of information or a shortened data-retention time. Even if a small amount of electrons are trapped at one event, repetitive trapping makes the stored voltage low enough for information to be lost.

There are many sources of electron injection, as shown in Fig. 5.5. The application of V_{BB} is effective in eliminate these sources, and preventing electron

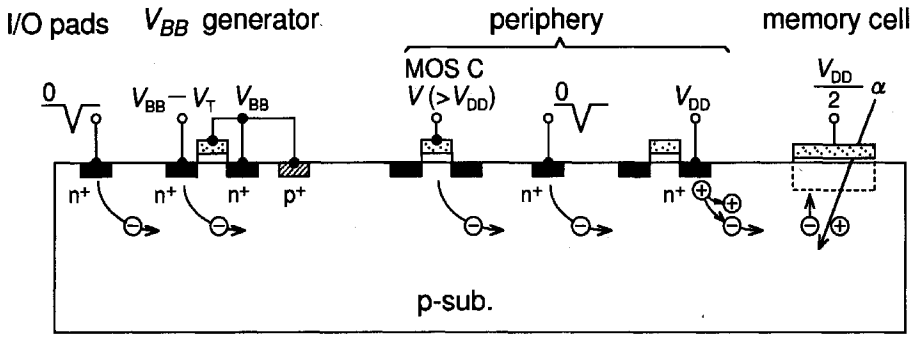


Fig. 5.5. Sources of minority-carrier injection [5.1]

injection. One of the major sources is the forward-biased p–n junction. The forward bias is brought about by pulse undershoots at I/O pads, as discussed previously, and by a diode (D in Fig. 5.6) in the V_{BB} generator. Internal spike noises caused by high-speed pulses may forward bias p–n junctions throughout the chip. The solutions, which are similar to those for CMOS latch-up, are: a deep V_{BB} application to the substrate; the use of a low-impurity epitaxial layer stacked on a high-impurity substrate; a memory array surrounded by a well to isolate it from the peripheral circuit, as in Fig. 5.4; and a V_{DD}-applied n^+ guard ring surrounding the V_{BB} generator, to trap minority carriers from the generator. In addition, a V_{BB} generator placed as far as 500 μm [5.1, 5.8] – which is sufficiently longer than the diffusion length of electrons from the memory array prevents minority carriers from reaching any memory cell. Replacing the output NMOSFETs of the V_{BB} generator with PMOSFETs cuts minority-carrier generation [5.25], as discussed later. Moreover, MOS capacitors in the peripheral circuit can be sources of minority carriers. An inversion layer formed at the bootstrap capacitor or the bypass capacitor might generate minority carriers, although the detailed mechanism remains unknown. The application of V_{BB} reduces the number of minority carriers, and replacement of the inversion layer with a p^+ layer stops minority carrier generation [5.7, 5.16]. Electrons generated by impact ionization at the NMOS drain are also injected into the substrate [5.26]. Enough separation of the sources (i.e. peripheral circuit) from the memory array, or the reduction of

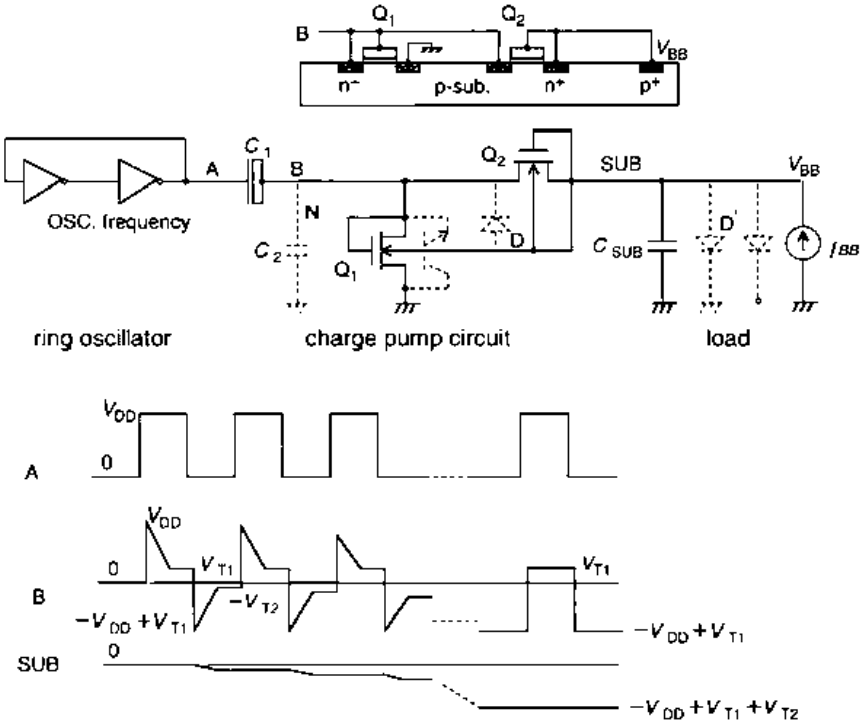


Fig. 5.6. A typical V_{BB} generator and its operating principle [5.1]

I_{BB}, is the solution. Some of electrons generated by α-particle irradiation can attack the memory-cell storage node. The solutions are explained in Chap. 4.

Although an on-chip V_{BB} generator is strongly needed, a deep understanding of both the generation scheme and the load characteristics is required for a successful design. The inherently small current-driving capability of the generator tends to make the V_{BB} level unstable, and the floating V_{BB} load becomes a source of instabilities. In the following sections the relevant issues are discussed, mainly citing a typical example of a DRAM chip comprising an NMOS DRAM cell array and a peripheral circuit integrated on a p-type substrate.

5.2.2 Basic Operation and Design Issues

Figure 5.6 shows a typical V_{BB} generator [5.5, 5.12, 5.13]. It comprises a ring oscillator with an odd number of inverter stages (from five to nine), and a charge-pump-related circuit that is composed of a MOS capacitor (C_1) and MOSFETs (Q_1, Q_2) working as rectifiers. A positive-going V_{DD} pulse at node A pulls up node B to V_{DD} because $C_1 \gg C_2$. Then, node B discharges until Q_1 is turned off and its final voltage is V_{T1} (i.e. the V_T of Q_1). During the discharging process, Q_2 diode remains cut off because of the back-biased diode connection. Next, a negative-going pulse at node A pulls down node B to a negative voltage, $-V_{DD} + V_{T1}$, turning Q_1 off and Q_2 on. Thus, node B is charged up to $-V_{T2}$ (V_{T2} is the V_T of Q_2) by injecting the same amount of charge (electrons) as the charge involved at node B, $C_1(V_{DD} - V_{T1} - V_{T2})$, to the substrate (SUB), giving the substrate a negative voltage. Although the resultant substrate-bias voltage is quite small because of the huge value of the substrate capacitance ($C_{SUB} \gg C_1$), repetitive pumping causes the substrate to be gradually discharged. However, this discharging stops when V_{BB} reaches the level, $-V_{DD} + V_{T1} + V_{T2}$, and Q_2 cut off. Here, careful attention should be paid so as not to turn on the parasitic devices in the generator: a diode (D) formed between node B (diffused layers at the Q_1 drain and Q_2 source) and the substrate, and an npn bipolar transistor formed at Q_1. Note that to cut off the diode (D) V_{T2} must be smaller than the built-in potential (= 0.6 V, the voltage drop of the pn diode) of the diode junction. In spite of this, the n^+ layer of the diode, which is weakly forward biased by V_{T2}, injects minority carriers (electrons) into the substrate. A small V_{T2} also prevents the parasitic bipolar transistor from being turned on. The setting of a larger channel length for Q_1 is also effective in avoiding activation of the bipolar transistor.

In actual practice, the V_{BB} level is settled by the charge (electron) pumping current (I_{CP}) of the V_{BB} generator and the substrate current (I_{BB}) flowing from NMOSFETs in the internal core circuits to the substrate. I_{BB} is the hole component of the impact ionization current in the vicinity of the drain of the NMOSFETs, as explained later, which is subject to shallowing of V_{BB}. Figure 5.7 shows the relationship between I_{CP} and V_{BB} as a parameter

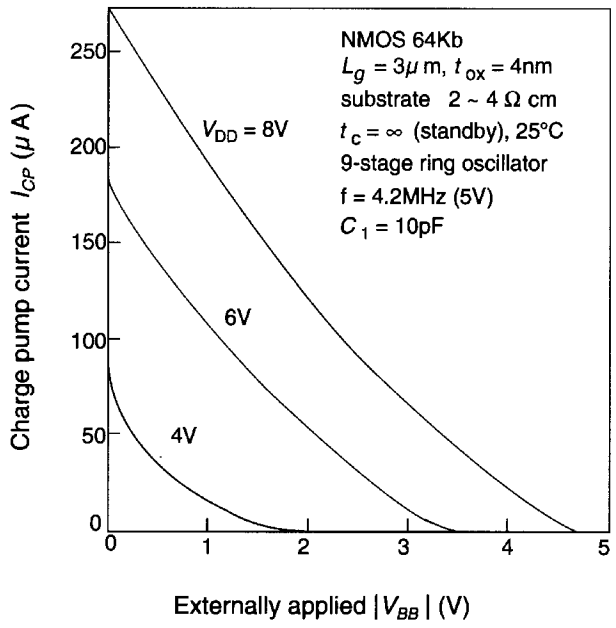


Fig. 5.7. I_{CP} versus V_{BB} [5.12]

of V_{DD} [5.12]. The V_{BB} generator is for a 64 Kb NMOS DRAM, using a channel length of 3 μm, a gate-oxide thickness of 40 nm, and a p-type substrate resistance of 2–4 Ω · cm. The frequency (f) of a nine-stage ring oscillator is 4.2 MHz at 5 V and 25 °C. The capacitances, C_1 and C_{SUB}, are 10 pF and 1300 pF, respectively. The pumping current, I_{CP}, which is proportional to $C_1 V_{DD} f$, increases with V_{DD}, but decreases with increasing $|V_{BB}|$ because of a decrease in the effective gate voltage of Q_2. Beyond a certain value of V_{BB}, Q_2 is cut off.

Figure 5.8 shows the relationship between I_{BB} and V_{DD} as a parameter of the memory cycle time (t_c) for 64 Kb NMOS DRAMs [5.8, 5.12]. Obviously, I_{BB} increases exponentially with an increasing V_{DD}, while it is almost inversely proportional to t_c. When a V_{BB} generator is incorporated in a chip, the actual V_{BB} is the one at which $I_{CP} = I_{BB}$. Thus, the relationship between V_{BB} and V_{DD} can be expected to be as follows. When V_{DD} is still small, and thus I_{BB} is small enough to accept a small I_{CP}, V_{BB} becomes deeper with V_{DD}, as shown in the regions of $I_{CP} \simeq 0$ in Fig. 5.7. However, when V_{DD} is increased, I_{BB} starts to exceed I_{CP} at a certain V_{DD}, since I_{BB} increases exponentially with V_{DD} while I_{CP} increases proportionally with V_{DD}. This implies that at this value of V_{DD} the number of holes injected into the substrate exceeds the number of electrons pumped into the substrate by the V_{BB} generator. As a result, V_{BB} starts to drop to a shallower level. An excessively shallow V_{BB} results in a low V_T, especially for NMOSFETs with a large substrate-bias effect coefficient (K). Thus, the source–drain current (I_{DS}) of MOSFETs in the major internal circuits starts to increase, causing an increase in I_{BB} since I_{BB} is proportional to I_{DS}. The increased I_{BB} drives the initial V_{BB}

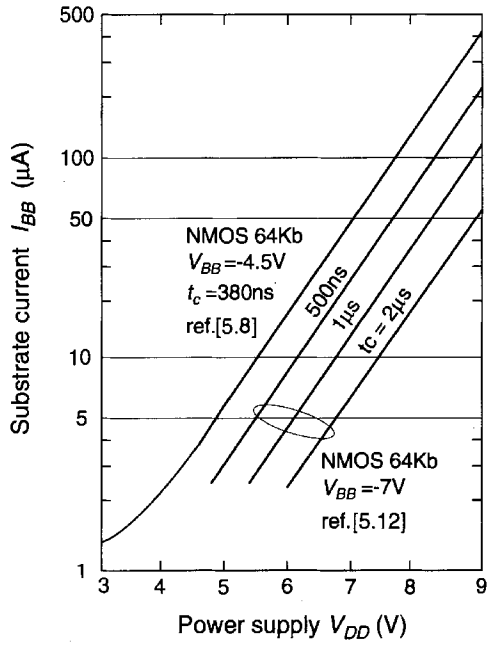


Fig. 5.8. I_{BB} versus V_{DD} for NMOS 64 Kb chips

to a further shallow value. As a result of the positive feedback, V_{BB} rapidly becomes positive, but is eventually clamped to 0.6 V, which is the built-in potential of a p–n junction diode, such as diode D′ in Fig. 5.6. For a smaller value of t_c (i.e. a faster cycle time) V_{BB} starts to drop at a lower V_{DD}, since I_{BB} is inversely proportional to t_c. A larger I_{CP} shifts the V_{BB} drop point to a higher V_{DD}, despite the increase in the power dissipation of the V_{BB} generator. This expectation has been experimentally verified, as shown in Fig. 5.9 [5.7, 5.10, 5.12].

In the on-chip V_{BB} approach there are several possible problems: a large rush current or CMOS latch-up; the incapability of burn-in test at a sufficiently high stress voltage due to the degraded V_{BB} level; unstable chip operation caused by a large V_{BB} bounce during amplification of the cell-signal voltage; the shortening of the cell data-retention time, due to minority carrier injection from the V_{BB} generator; and an increase in chip stand-by power due to the addition of the V_{BB} generator. In the following sections, the V_{BB}-relevant problems and solutions are discussed in detail.

5.2.3 Power-On Characteristics

A large rush current would be developed during power-on. This is because the substrate is biased to a positive voltage by capacitive coupling between the power-supply node and substrate. This positive bias would make the V_T of the NMOSFET negative and thus increase the I_{DS} of the NMOSFET, causing a rush current in the chip. Even for CMOS design, the capacitive

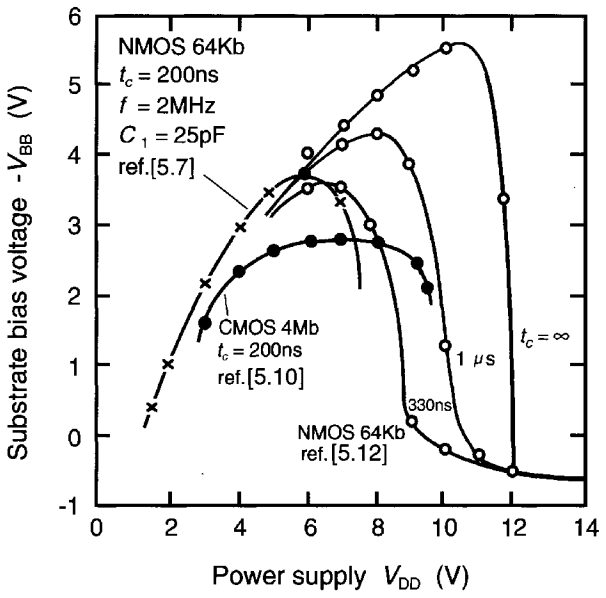


Fig. 5.9. V_{BB} versus V_{DD} for DRAM chips incorporating a V_{BB} generator

coupling during power-on may cause CMOS latch-up, if the substrate is at a floating V_{BB}. In the following, power-on characteristics will be explained.

The substrate starts to be discharged from an initial floating voltage of 0 V toward a final deep V_{BB} by charge-pumping of the V_{BB} generator. However, it takes quite a long time to reach a final V_{BB}. In addition to a heavy C_{SUB}, a small number of charges pumped to the substrate per ring-oscillator cycle, which stems from a low V_{DD} during power-on and the resultant low pumping frequency, are responsible for the slow discharging. Moreover, during V_{BB} discharge the positive-going V_{DD} couples a positive voltage to the substrate, which is still at a shallow V_{BB}, through the V_{DD} node–substrate capacitance. As a result, on the way to a final value V_{BB} is instantaneously raised to a positive value, causing an unexpectedly large surge current.

Figure 5.10 shows the well-known substrate structure of an NMOS DRAM. The current flowing from the V_{DD} node to the substrate during power-on is expressed as

$$I_1 = \frac{C_{PS}C_{SUB}}{C_{PS}+C_{SUB}}\frac{\Delta V_{DD}}{\Delta t}, \tag{5.2}$$

where C_{PS} is the total capacitance formed between the substrate and other conductors such as the memory-cell plate (electrode) and data lines. During this period, V_{BB} is increasing toward the voltage, $C_{PS}V_{DD}/(C_{PS}+C_{SUB})$, that could be coupled to the substrate through C_{PS}. However, the p–n junctions clamp V_{BB} to 0.6 V, which is the built-in potential of a p–n junction. Therefore, a parasitic npn bipolar transistor is forward biased with a base current of I_1. Thus, as a result of current amplification, a large collector–emitter current (I_2) flows. I_2 increases as the amplitude of the V_{DD} waveform is increased and

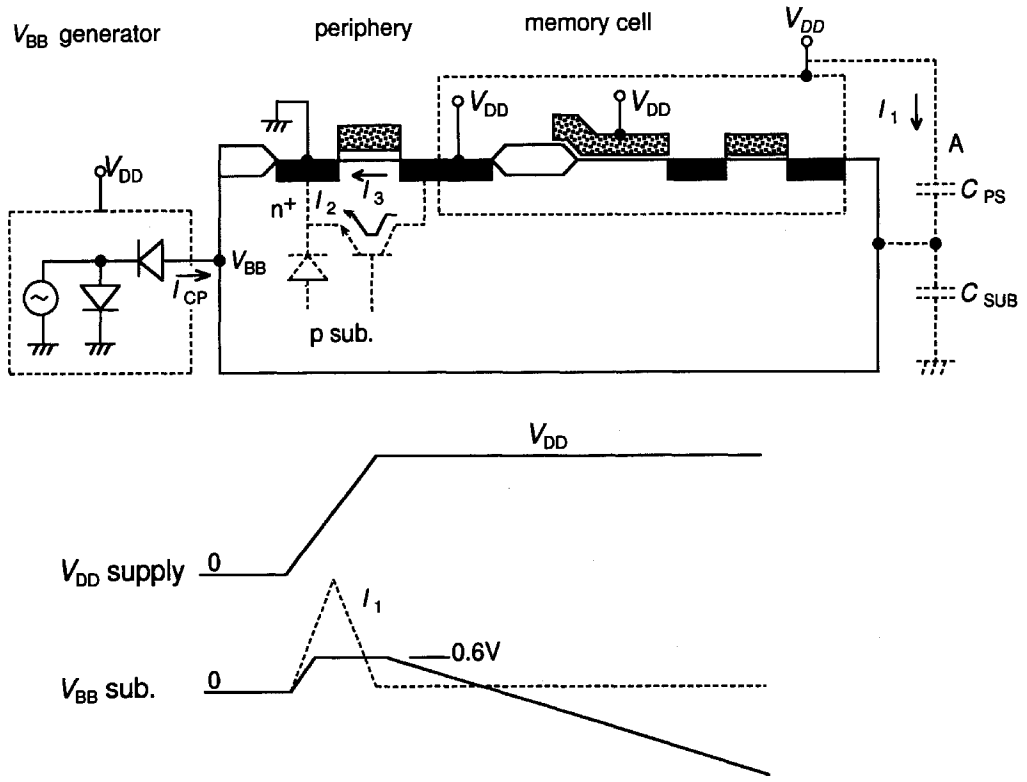


Fig. 5.10. The power-on characteristics of a V_{BB} generator [5.1]

the rise time is shortened. If the V_T of the NMOSFET is sufficiently negative during the period of positive V_{BB}, it keeps on latching the V_{BB} to the clamped positive level [5.10] even after V_{DD} reaches the final V_{DD} value of power-on. This is because the source-drain current (I_3) caused by the negative V_T is too large to be managed by the V_{BB} generator. If V_T is not so negative, or is still positive, even at $V_{BB} = +0.6\,V$, I_3 is quite small or zero. Thus, after the transient currents (I_1, I_3) relevant to the capacitive coupling have decayed sufficiently, V_{BB} starts to fall to a negative value. In any case, the currents (I_1, I_2, I_3) simultaneously generated by the above mechanisms are observed as the power-on surge current. A considerable surge current may prevent the power supply of the system from turning on if it exceeds the current-delivering capacity of the supply. It has been reported [5.1, 5.6] that the surge current of the V_{DD} capacitor plate of the folded data-line arrangement can be reduced to one-third by the V_{SS} plate. Note that during power-off V_{BB} changes to a more negative value due to capacitive coupling, and then discharges to ground with a long time constant. During this period, the V_{BB} generator remains off, and there are fewer problems, if any, because of the ever-decreasing value of V_{DD}.

In the case of a double-well CMOS structure, a positive V_{BB} is subject to CMOS latch-up. Figure 5.11 illustrates the cross-section of a double-well CMOS 4 Mb DRAM [5.14]. NMOSFETs and stacked-capacitor memory cells

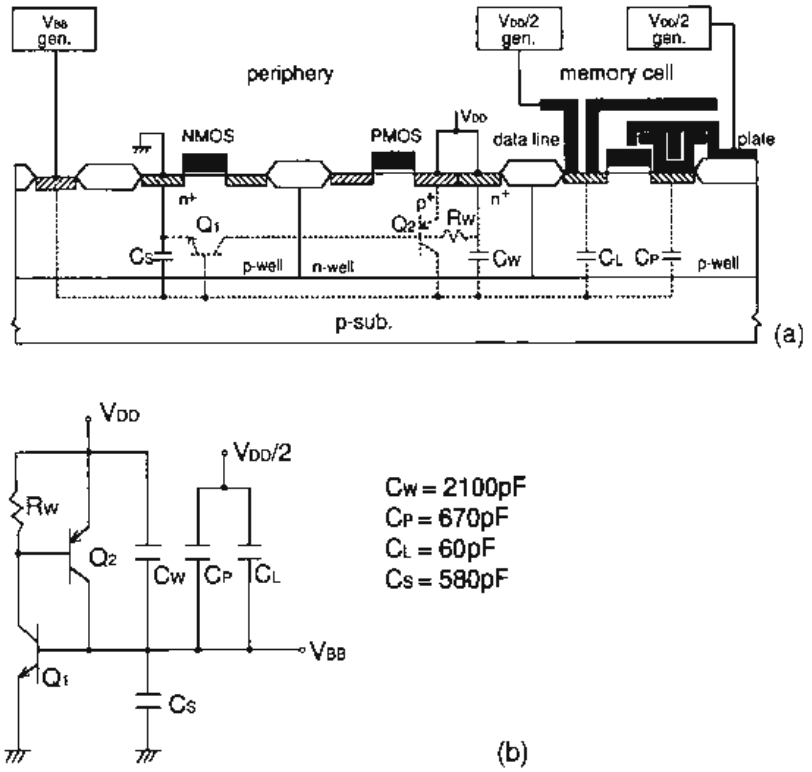


Fig. 5.11. A cross-section of a double-well CMOS 4 Mb DRAM chip (**a**) and the equivalent circuit in relation to latch-up (**b**) [5.14]

are embedded in a p-well biased at $V_{BB} = -2.5\,V$, which is generated by an on-chip generator. A half-V_{DD}, also generated by another on-chip generator, is applied to both capacitor plates and data lines through precharge circuits. The n-well for PMOSFETs is at V_{DD}, and thus V_{BB} is 0 V. Despite the built-in structure of npn and pnp parasitic bipolar transistors (Q_1, Q_2), Q_1 is cut off in normal operation because the Q_1 base is deeply back-biased at $V_{BB} = -2.5\,V$. During power-on, however, CMOS latch-up could occur due to the forward biasing of the base-emitter of Q_1 by a positive V_{BB}, which is caused by coupling capacitances between the n-well and the substrate (C_W), the capacitor-plate and the substrate (C_P), and the data line and the substrate (C_L). The variation in V_{BB} (ΔV_{BB}) due to variation in V_{DD} (ΔV_{DD}) during power-on is given by:

$$\Delta V_{BB} = \frac{C_W + (C_P + C_L)/2}{C_W + C_P + C_L + C_S} \Delta V_{DD}. \tag{5.3}$$

Thus, $\Delta V_{BB} = 0.72 \Delta V_{DD}$ is obtained by using the parameters in the figure. This implies that latch-up starts easily, because V_{BB} is soon clamped to around 0.6 V at a ΔV_{DD} as small as 0.8 V, on the way to the maximum ΔV_{DD} ($= V_{DD} = 5\,V$). The current (I_1) flowing from the V_{DD} node to the V_{BB} node after clamping is given by

$$I_1 = \left(C_W + \frac{C_P + C_L}{2}\right) \frac{V_{DD}}{\Delta t}. \tag{5.4}$$

Only a part of I_1 becomes the Q_1 base current (I_B), and thus the collector current ($= h_{fe} I_B$, where h_{fe} is the current gain) flows across the well-resistor (R_W). The voltage drop at R_W could turn Q_2 on, causing the latch-up. Therefore, the latch-up susceptibility is proportional to

$$\left(C_W + \frac{C_P + C_L}{2}\right) \frac{\Delta V_{DD}}{\Delta t} \times h_{fe} \times R_W. \tag{5.5}$$

To improve latch-up immunity, each term in the equation must be as small as possible, although $\Delta V_{DD}/\Delta t$ must be larger in terms of ease of use. The following are device and circuit techniques to meet this requirement.

Figure 5.12 shows latch-up characteristics [5.14] as a parameter of an n-well phosphorous dose. Obviously, a shorter V_{DD} rise time causes latch-up more easily and a larger dose improves the immunity with a reduced R_W. Here, if V_{BB} can be enforced to be clamped to less than 0.6 V, latch-up never occurs because the npn transistor is always cut off. The clamping is achieved by an additional low-V_T MOSFET, as shown in Fig. 5.13 [5.15]. In normal operation at a negative V_{BB}, the MOSFET is cut off because the drain voltage and the gate–source voltage are 0 V, causing no detrimental effects to the V_{BB} generator. If V_{BB} is raised to a positive value, the MOSFET is turned on, because the gate–source voltage is positive, so that V_{BB} is clamped to its V_T. At $V_{BB} = +0.5$ V the drain current is as large as 2 mA, but the p–n junction is still off. Note that the MOSFET must be large enough to sink a large I_{BB}. In addition, quite large V_T variations caused by physical and temperature variations of the MOSFET eventually make the difference between V_T and the built-in potential small, and the efficient clamp is lost.

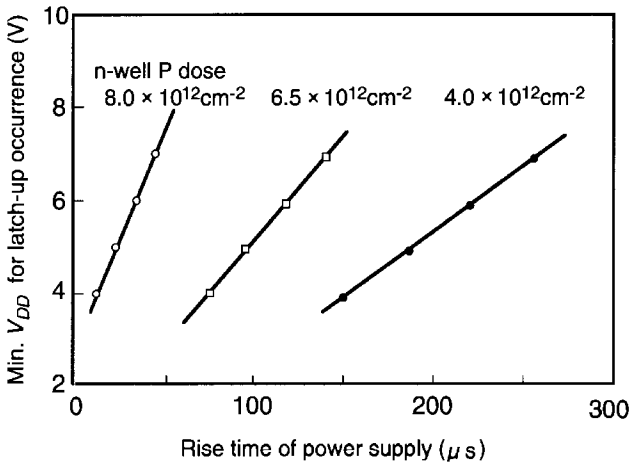


Fig. 5.12. The latch-up immunity of a CMOS 4 Mb DRAM as a function of n-well concentration [5.14]

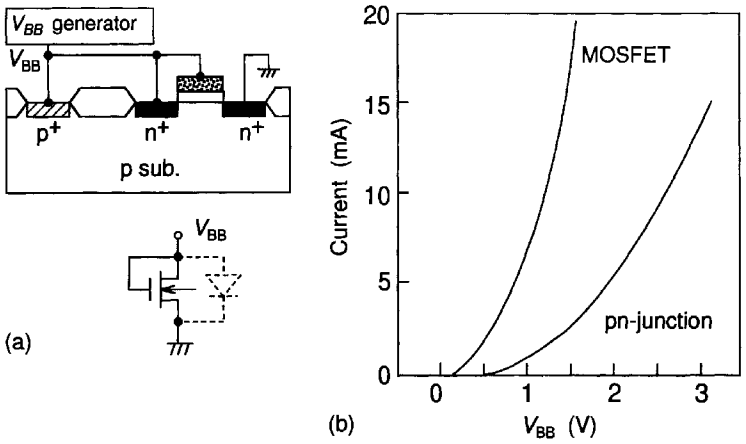


Fig. 5.13. Clamping of the substrate voltage by a MOSFET [5.15]. (**a**) Cross-section; (**b**) current versus V_{BB}

Figure 5.14 shows a circuit for clamping V_{BB} to ground [5.14]. The circuit gives a well-fixed V_{BB}, which is free from the above-described V_T variations, with a large difference from the built-in potential. The V_{BB} generator consists of a conventional V_{BB} generator and a power-on reset-signal generator. The reset-signal generator for pumping C_3, which is similar to the V_{BB} generator, is composed of a seven-stage ring oscillator, a charge-pumping capacitor (C_2), and MOS diodes. During power-on the reset-signal generator starts to pump node N_2 earlier than the V_{BB} generator, because $C_3 \ll C_{SUB}$. While node N_2 and thus the reset-signal $\overline{POR}_1$ are still at a low level, Q_3 is cut off. However, the rising V_{DD} turns on Q_4, allowing the substrate to be clamped to 0 V (i.e. V_{BB}-clamping). When the N_2 voltage exceeds the logic threshold of the next inverter, $\overline{POR}_1$ goes up to a high level. Thus, Q_3 is turned on, and node N1 is discharged to the substrate voltage (V_{BB}) of 0 V, so that Q_4 is turned off. After that, the substrate starts to be discharged by the V_{BB} generator, as usual. The succeeding discharge is fast, because V_{DD} has been high enough at this time. Here, the reset-signal generator tracks the rise time of V_{DD} well, allowing the suppression of latch-up: when V_{DD} rises slowly, the resulting low frequency of the ring oscillator makes the V_{BB}-clamping period long, avoiding latch-up caused by the early release of V_{BB}-clamping. The V_{BB}-clamping suppresses the V_{BB} noise coupled during power-on through the n-well to substrate capacitance (C_W in Fig. 5.11). However, it would not be valid for other V_{BB} noises from the memory-cell capacitor plate and the data lines, because V_{DD} may be indirectly applied to them through a kind of delay circuit, which differs from C_W-coupling. In fact, V_{DD} is applied to them through an on-chip half-V_{DD} generator, with a delay after the application of V_{DD}. In this case, although the V_{BB} noise through C_W is successfully suppressed and V_{BB} starts to fall to a negative value, the subsequent other noises positively coupled to the substrate through C_P and C_L in Fig. 5.11 prevent V_{BB} from falling to the value. Therefore, the $V_{DD}/2$ generator must rise up completely

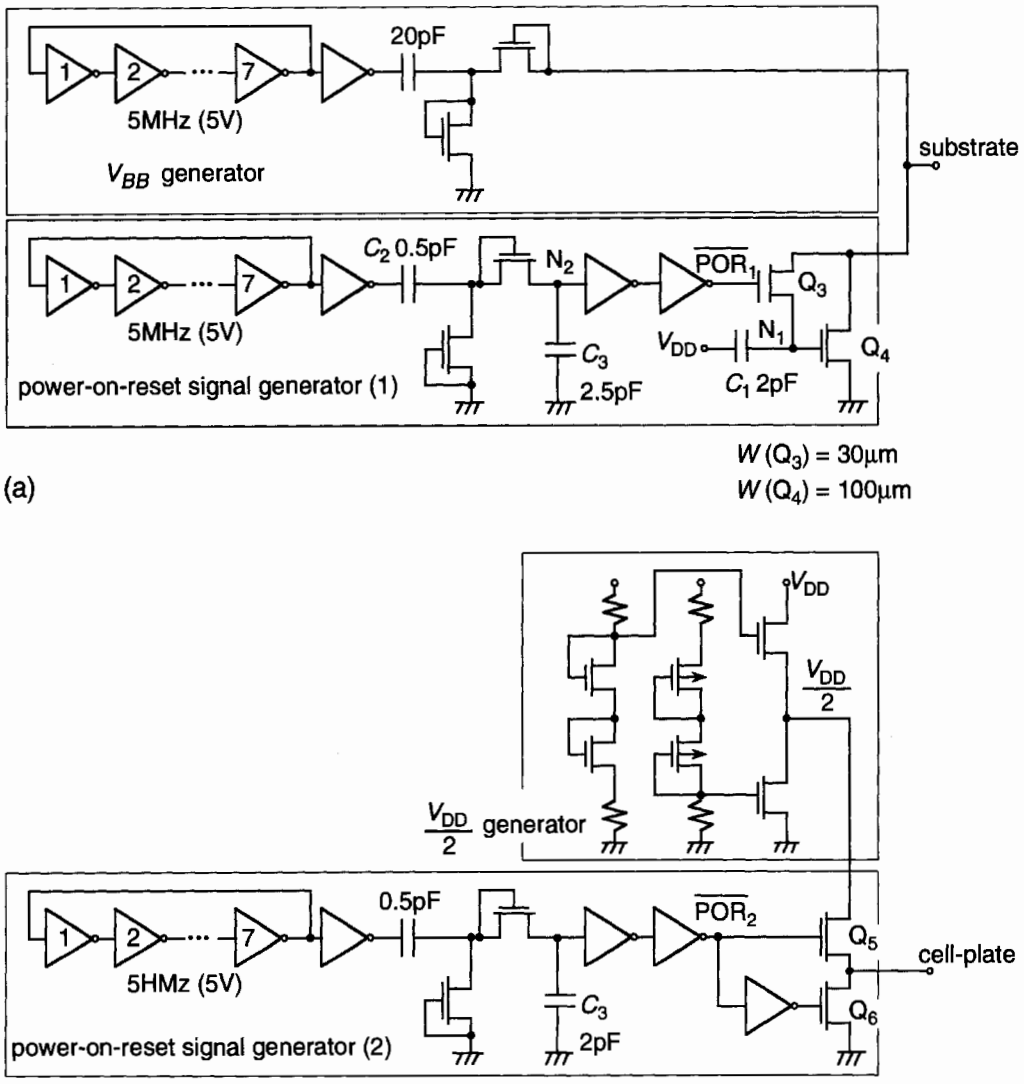


Fig. 5.14. Control circuits for V_{BB} (**a**) and a half-V_{DD} (**b**) [5.14]

to its final voltage while the substrate is clamped at 0 V. Figure 5.14b shows a $V_{DD}/2$ generator to meet this requirement. $\overline{POR_2}$-generation preceding $\overline{POR_1}$-generation is always ensured, because C_3 for $\overline{POR_2}$ is smaller than that for $\overline{POR_1}$. Figure 5.15 shows experimental waveforms of a CMOS 4 Mb DRAM using the above circuits [5.14]. No latch-up was observed for all of the doses shown in Fig. 5.12, despite wide ranges of variation of V_{DD}, of 4–7 V, and rise-time variations of 10–500 µs.

5.2.4 Characteristics in the High-V_{DD} Region

The voltage margin in the high-V_{DD} region must be guaranteed in order to perform the burn-in test with a high stress voltage. Since the test is an

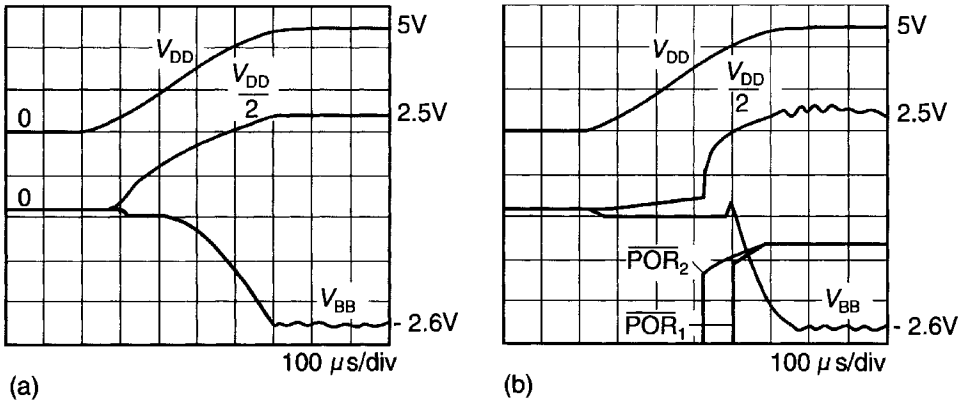
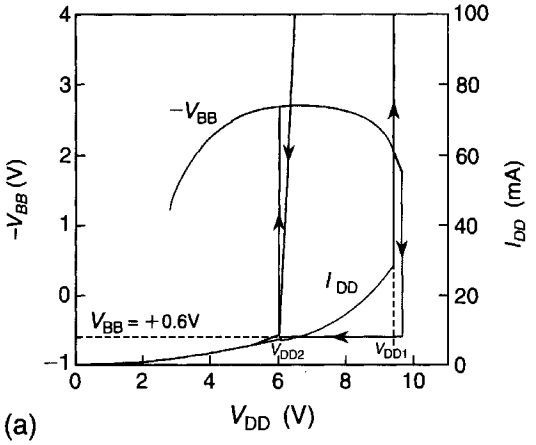


Fig. 5.15. Various waveforms at power-on timing, without a clamping circuit (**a**) and with a clamping circuit (**b**) [5.14]



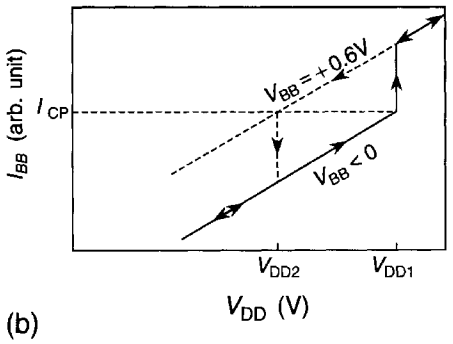


Fig. 5.16. The failure mode of a CMOS 4 Mb chip, showing hysteresis in V_{DD} versus I_{DD} (**a**) and V_{DD} versus I_{BB} (**b**) [5.10]

acceleration test, all of the voltages including V_{BB} must be sufficiently deeper during the test. However, a positive V_{BB} is generated at high V_{DD}, as seen in Fig. 5.9. Thus, to meet this requirement, a deep understanding of the behavior of the V_{BB} generator and of the MOSFETs in the high-V_{DD} region is essential.

Figure 5.16 shows the V_{BB} characteristics of a CMOS 4 Mb DRAM [5.10], whose chip cross-section is similar to that shown in Fig. 5.11. They are similar

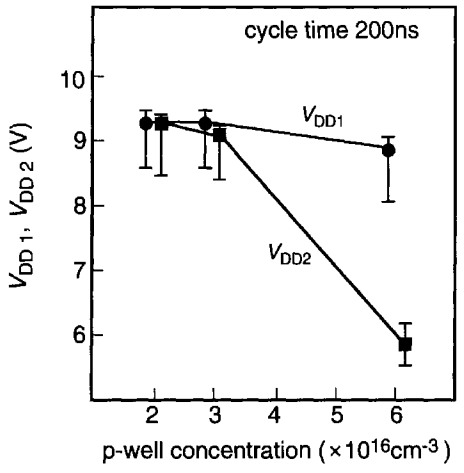


Fig. 5.17. The dependence of V_{DD1} and V_{DD2} on p-well concentration [5.10]

to the power-on characteristics described before. When V_{DD} is increased from a nominal value of 5 V, V_{BB} starts to be shallow at a certain V_{DD} (i.e. V_{DD1}) due to the mechanism described earlier. The resulting reduced V_T increases the source–drain current (I_{DD}) of each NMOSFET, and thus I_{BB}. In turn, the increased I_{BB} makes V_{BB} shallower. Finally, such positive feedback results in a positive V_{BB} of 0.6 V, as explained previously. If V_T is negative at $V_{BB} \geq 0$, the feedback is accelerated and V_{BB} stays at 0.6 V although V_{DD} is increased beyond V_{DD1}. On the other hand, when V_{DD} is decreased from a high value, V_{BB} remains at 0.6 V even below V_{DD1}, although the decrease in V_{DD} reduces I_{BB}. When V_{DD} reaches V_{DD2}, at which $I_{BB} = I_{CP}$ (i.e. the pumping current of the V_{BB} generator), V_{BB} starts to return to a negative value. Thus, V_{BB} has hysteresis characteristics for V_{DD}. These characteristics are avoided by device design, as shown in Fig. 5.17. A low p-well concentration of below $(2 - 3) \times 10^{16}\,\mathrm{cm}^{-3}$ is appropriate for the NMOSFETs that are the same FETs as those shown in Fig. 5.3. This is because the substrate bias effect coefficient (K) is less and V_T is still positive even at $V_{BB} \geq 0$.

5.2.5 The V_{BB} Bump

Even during normal operation after power-on, voltage changes at nodes in a chip cause voltage bumps (ΔV_{BB}'s) to the substrate through the node–substrate p–n junction capacitances. This is because the output impedance of the V_{BB} generator is so high that the substrate is substantially regarded as a floating body. The major nodes that contribute to a large ΔV_{BB} are the data lines and outputs of decoders, because of a huge number of nodes and a quite large capacitance at each node. For example, in an NMOS DRAM using the V_{DD} data-line precharge scheme, the simultaneous charging or discharging of many data lines with an amplitude of V_{DD} couples a large ΔV_{BB}. All of the nodes of the decoders, except for the selected one, which have been precharged to V_{DD}, are also simultaneously discharged and cause a large ΔV_{BB}. Figure 5.18 shows ΔV_{BB} components calculated for NMOS 64 Kb

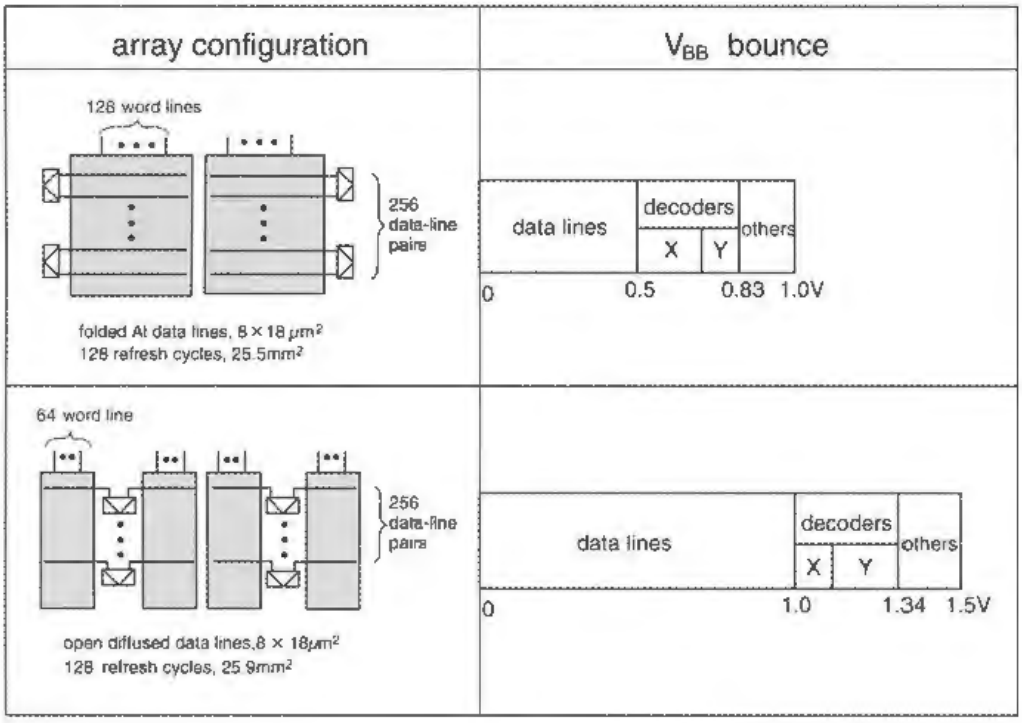


Fig. 5.18. The V_{BB} bounce components of NMOS 64 Kb DRAMs [5.7]

DRAMs [5.7]. A folded Al data-line arrangement and an open diffused data-line arrangement are compared. The data lines and decoders occupy 80–90% of the whole ΔV_{BB}. Even for Al data-lines, ΔV_{BB} is as large as 1 V, as shown in Fig. 5.19 [5.7]. As long as the substrate is deeply biased to around −3 V and the substrate-bias effect coefficient (K) of the NMOSFET is not so large, variation in the V_T caused by a ΔV_{BB} as large as 1 V is so small that the operation of the peripheral circuit is still stable. For a memory-cell operation, however, a large ΔV_{BB} would cause degradation or instability for some kinds of memory cell, as discussed next.

The V_{BB} bump during precharging degrades the voltage margin of the low-stored information of the 1-T cell. Figure 5.20a shows a cross-section of a memory cell with an open diffused data-line arrangement [5.16]. It features the Hi-C capacitor structure – that is, an n^- layer surrounded by a p^+ layer under the plate – in order to increase the cell capacitance and provide immunity from α-particle-induced soft errors. Figures 5.20b,c show an array structure using the cell and various capacitor components [5.16]. C_1 is the total capacitance of the substrate to power (V_{DD} and V_{SS}) lines, C_2 is the total capacitance of the substrate to all data lines, C_3 and C_4 are the total capacitances of the substrate to all row-decoder outputs and to all column-decoder outputs, respectively, and C_5 is the total capacitance of the substrate to the memory-cell capacitor plates and to all nodes in dynamically operating peripheral circuits. Here, we assume that a low level (0 V) has been written

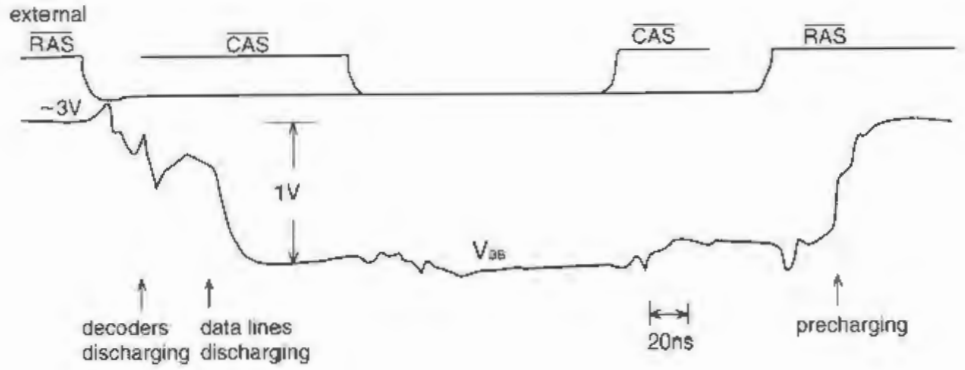


Fig. 5.19. The V_{BB} waveform of a 64 Kb DRAM using folded Al-data lines and an on-chip V_{BB} generator [5.7]. Chip area = $3.43 \times 7.52\,\text{mm}^2$, $C_D = 515\,\text{fF}$, 128 refresh cycles, $V_{DD} = 5\,\text{V}$, data-line precharge voltage = $V_{DD} - V_T = 4.5\,\text{V}$, $\rho_{SUB} = 10\,\Omega\text{-cm}$

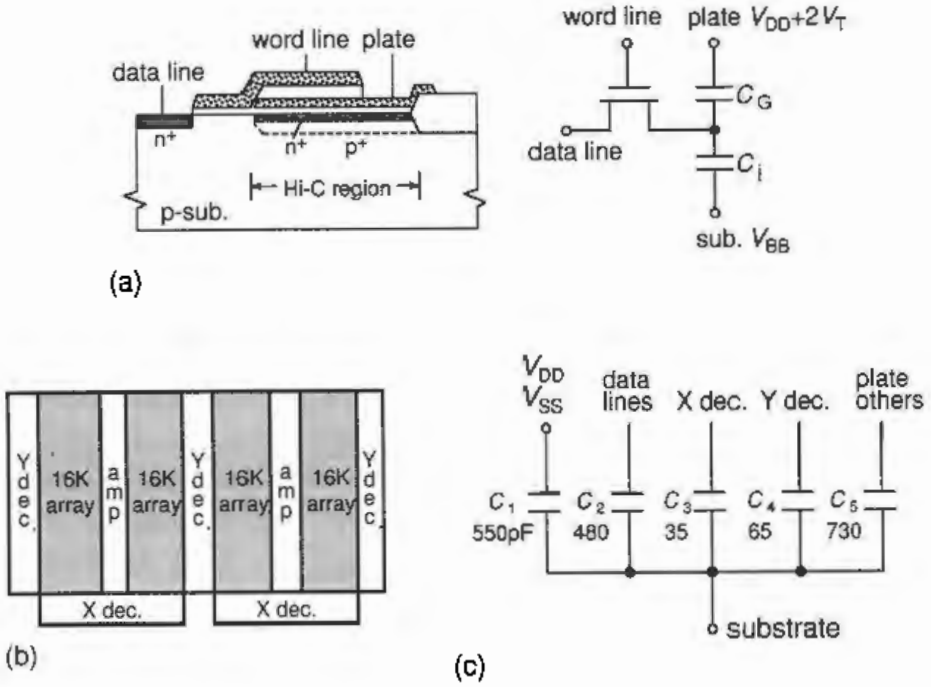


Fig. 5.20. The substrate capacitance of an NMOS 64 Kb DRAM using the Hi-C cell [5.16]. (**a**) Cross-section of Hi-C cell with an open diffused data line; (b) array organization; (**c**) capacitors to the substrate

to the cell by turning off the word line. The subsequent data-line precharging from 0 V to V_{DD} causes a V_{BB} bump to the substrate, so that the bump makes the floating 0 V of the storage node increase by a positive voltage, ΔV, which can be expressed as

$$\Delta V = \Delta V_{BB} \frac{C_j}{C_G + C_j}. \tag{5.6}$$

When the cell is subsequently read, the signal voltage of the cell is degraded by ΔV. Here, a ΔV_{BB} as large as 1 V at $V_{DD} = 5$ V was experimentally obtained with a 64 Kb chip using the cell, although this is smaller than the calculated value shown in Fig. 5.18. Thus, the resultant ΔV is as large as 0.32 V, since the ratio of the junction capacitance (C_j) to the total cell capacitance is large ($\simeq$ 32%) in Hi-C capacitor. Thus, a cell voltage margin of a low level loses by 13%, for an original margin of 2.5 V.

The large V_{BB} bump involved in the V_{DD} data-line precharging scheme would result in unstable operation of the whole chip if the resulting V_{BB} were to become positive, as discussed previously. For example, it would be hazardous if high-speed column modes, each of which generated a large substrate current (I_{BB}), were to be applied after V_{BB} dropped by ΔV_{BB} (i.e. a negative V_{BB}-bump) due to data-line discharging. If successive column modes continue for the maximum duration as given in the catalog specifications (i.e. 16 μs) with the fastest cycle time (for example, less than 50 ns) and the resulting I_{BB} almost exceeds I_{CP}, I_{BB} makes V_{BB} gradually shallower, causing the shallowest V_{BB} at the ends of the modes. The succeeding data-line precharge creates a positive V_{BB} bump, so that a positive V_{BB} is eventually developed as a result of adding the bump and the shallowest V_{BB}.

To reduce the V_{BB} bump, a smoothing (bypass) capacitor built in the substrate [5.16] has been proposed. However, a more effective solution is the use of a combination of a half-V_{DD} data-line precharge, a folded Al-data-line arrangement, and CMOS technology. In principle, a half-V_{DD} data-line precharge cancels the bump, as discussed in Chap. 4. The use of an Al data line suppresses ΔV_{BB} with minimization of the diffused area of the data line. The CMOS circuit allows only one of the decoders to be discharged, and thus minimizes ΔV_{BB}, as explained in Chap. 3.

5.2.6 Substrate-Current Generation

One of the most effective solutions to cope with the problems related to the on-chip V_{BB} generator is to reduce the substrate current (I_{BB}) of the whole chip. I_{BB} is generated only when the source–drain current flows in proportion to the FET size (i.e. the channel width). Thus, the peripheral circuit is the major source of I_{BB}, because it is composed of large MOSFETs that operate during the memory cycle. On the other hand, iterative circuit blocks relevant to the memory array, such as decoders and word drivers, are the minor sources, especially in CMOS design, although their total channel width is huge. This is because only a small number of MOSFETs, which are smaller in size than those in the peripheral circuit, operate selectively. Thus, reduction of the I_{BB} of each MOSFET in the peripheral circuit is crucial. Note that this reduction is also essential to ensure the reliability of MOSFETs, as discussed later. In this section, I_{BB} characteristics and I_{BB} reductions [5.1, 5.17–5.19] are discussed.

I_{BB} Characteristics. As the MOSFET is scaled down for a fixed V_{DD}, the electric field near the drain strengthens. Consequently, electrons flowing from the source to the drain obtain a high energy from the high electric field (so-called "hot electrons"), and generate electron–hole pairs as a result of impact ionization at the drain, as shown in Fig. 5.21. Some electrons from these pairs flow into the drain. The others are injected into the gate insulator as the gate current (I_G) and are trapped there, causing a gradual change in V_T and a decrease in the transconductance of the MOSFET. On the other hand, some holes of the pairs flow into the substrate, resulting in the substrate current (I_{BB}). The I_{BB} of an NMOSFET is three orders larger than that of a PMOSFET, because of a large impact ionization coefficient and a higher electric field due to a sharper impurity profile near the drain. Thus, the degradation of NMOSFET device parameters is more prominent. The maximum I_{BB}, I_{BBmax}, which is usually developed at $V_{GS} = V_{DS}/2$, is expressed as follows:

$$I_{BBmax} \propto \exp(-\gamma/V_{DS}), \tag{5.7}$$

where γ is a constant. The life time (τ_{HC}) of an NMOSFET, which is defined as the time when V_T degrades by 100 mV due to hot electrons, is given by

$$\tau_{HC} \propto (I_{BBmax}/W)^{-n}/(f \cdot t_{SUB}), \quad n = 2.5 - 3.0, \tag{5.8}$$

where W is the channel width, t_{SUB} is the pulse width of the I_{BB} pulse with an amplitude of I_{BBmax}, and f is the pulse frequency. Obviously, the most efficient way to reduce I_{BB} is to lower V_{DS} (i.e. V_{DD}), because I_{BBmax} reduces exponentially with a reducing V_{DS}. A one-order reduction of I_{BBmax} extends the life of the MOSFET by three orders. Even for a given V_{DD}, I_{BB} can be reduced considerably by the following improvements in the MOSFET structure and circuit.

The Stress-Released Drain Structure. Figure 5.22 shows the well-known LDD (Lightly Doped Drain–Source) NMOSFET [5.20]. A low-concentration n(n^-) layer relaxes the electric field at the drain. As a result, the

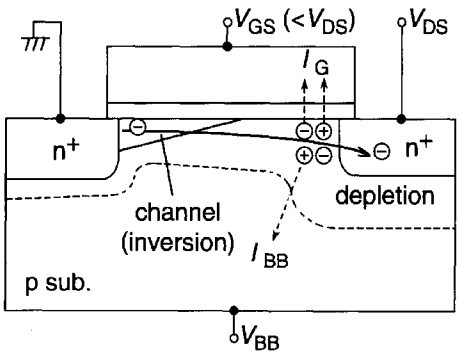


Fig. 5.21. The mechanism of hot-carrier injection [5.1]

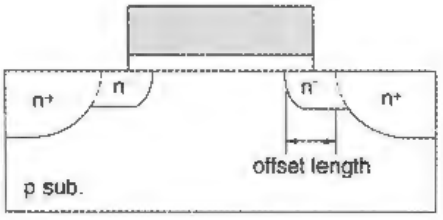


Fig. 5.22. A lightly doped drain-source (LDD) NMOSFET [5.20]

breakdown and hot-carrier characteristics, and short-channel effects, are improved by adjusting the n^- concentration and the offset length. I_{BB} is also reduced.

The Drain–Source Voltage-Reduction Circuit. In a DRAM chip consisting of dynamic circuits, the NMOSFET for discharging a floating node at a boosted voltage (i.e. a higher V_{DS}) generates a large I_{BB}. The MOSFETs in a word driver, in a bootstrapped NMOS driver (Q_2 in Fig. 2.28), or in a voltage up-converter are typical FETs that are exposed to such a high voltage. I_{BB} is reduced remarkably by the insertion of an NMOSFET, the gate of which is connected to a dc voltage, between the boosted node and the discharging MOSFET [5.21–5.23]. The reduction can be explained by nsing the conventional CMOS inverter shown in Fig. 5.23 [5.23]. Q_2 is the inserted NMOSFET. When a low voltage (0 V) is inputted, the output voltage becomes V_{DD}, while node N is charged up to $V_{GG} - V_T$. Here, $V_{GG} - V_T$ is quite low, because the substrate bias effect due to the raised Q_2 source voltage increases the V_T value of Q_2. When inputting V_{DD}, node N is discharged from

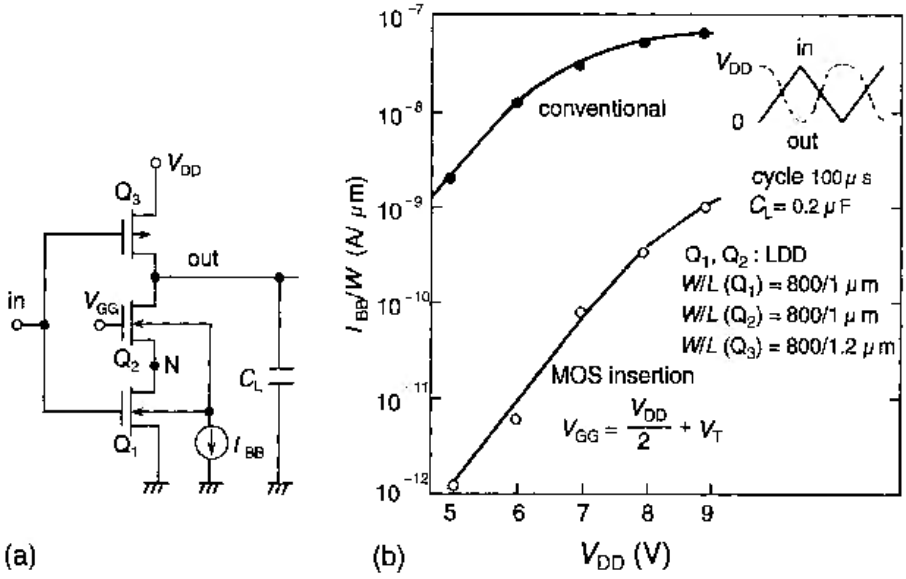


Fig. 5.23. The reduction of I_{BB} due to the insertion of an NMOSFET [5.23]. (**a**) Circuit; (**b**) I_{BB} versus V_{DD}

$V_{GG} - V_T$ to 0 V. Thus, the initial drain–source voltages of Q_1 and Q_2 at the discharging are $V_{GG} - V_T$ and $V_{DD} - (V_{GG} - V_T)$, respectively. Obviously, the total I_{BB} of Q_1 and Q_2 is minimized when the lowest drain–source voltage is simultaneously applied to both FETs. It is realized by $V_{GG} = V_{DD}/2 + V_T$, which is obtained by equalizing both of the initial drain–source voltages. This application of V_{GG} reduces I_{BB} by three orders, as shown in Fig. 5.23b. Here, the use of a longer channel length for Q_1 instead of Q_2 insertion is not so effective. For example, the change from 1 µm to 2 µm reduces I_{BB} only to one-third. This is because the strength of the electric field that contributes to I_{BB} is mainly determined by the drain junction depth and potential profile and the gate-oxide thickness, rather than the channel length. In the conventional CMOS inverter, the I_{BB} generated during discharging the output is larger than that generated during charging up of the output, because of the larger I_{BB} value of the NMOSFET. I_{BB} increases with an increase in the output capacitance.

Although the major source of the chip I_{BB} is the peripheral circuit, the I_{BB} of MOSFETs relevant to the memory array must be reduced. It has been reported that half-V_{DD} data-line precharge has an advantageous I_{BB} reduction capability, with a halved drain–source voltage [5.21]. Not only does it reduce the I_{BB} of the I-T cell FET to three orders less than that of the V_{DD} data-line precharge, but also the I_{BB} of the sense-amplifier MOSFETs. However, attention should be paid to equalizing the MOSFETs [5.24] on a pair of data lines. This is because the source–drain voltage, which is V_{DD} in the initial stage of precharging, is so high that hot electrons are generated and the resulting degradations in V_T and transconductance increase the imbalances of a pair of data lines.

5.2.7 Triple-Well Structures

To solve the problem of I_{BB} overloading in the on-chip V_{BB} generator, triple-well structures are useful. These structures enable the separation of two substrates, one belonging to the memory array and the other for the peripheral circuit, so that different V_{BB}'s are applicable. For example, if the substrate of the peripheral circuit can be well fixed to 0 V throughout the chip while that of the memory array is biased at V_{BB}, the problem is almost solved.

Figure 5.24 shows a triple-well structure proposed for a 16 Mb DRAM [5.29]. A negative V_{BB} is applied only to the NMOS memory-cell array and relevant circuits that are embedded in a p-well on n-type substrate, while no back bias is applied to NMOSFETs and PMOSFETs in the peripheral circuit. In order to operate a core circuit at V_{INT} (i.e. V_{DL}) reduced by an on-chip voltage down-converter, a p-well is added to the conventional double-well structure (Fig. 5.11). Here, the converter and data-output buffers in the n-well operate directly at V_{DD}. In this scheme, simultaneous applications of V_{DD} to both the n-substrate and the n-well solve the power-on issue. However, the influence of the V_{DD} bump, which is never suppressed by the V_{BB} generator, from the

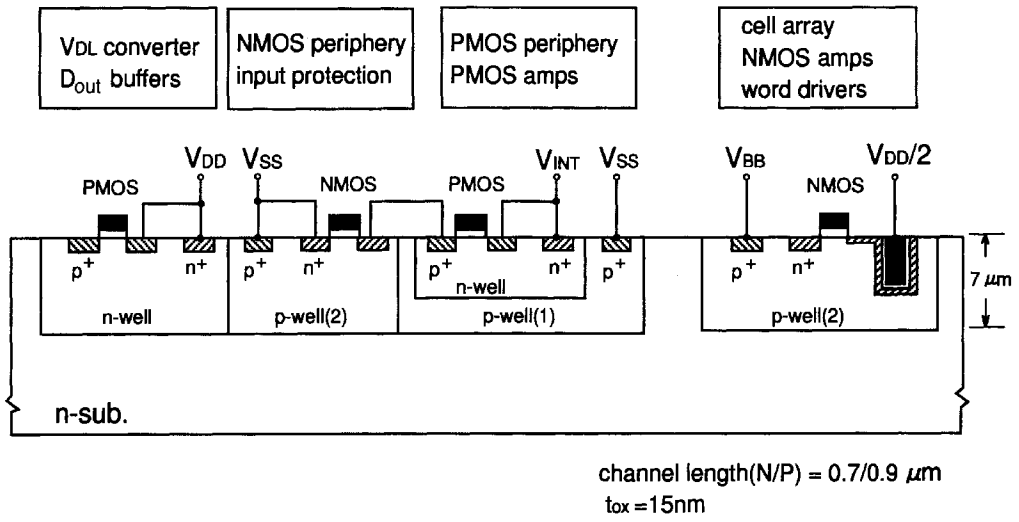


Fig. 5.24. The triple-well structure and biasing scheme [5.29]

whole n-substrate to the p-well of the memory array during power-on, remains unknown. Another triple-well structure (see Fig. 8.28) proposed for a 256 Mb DRAM [5.11] might be a strong candidate for future substrate structures.

5.2.8 Low-Power V_{BB} Generators

A low-power V_{BB} generator is strongly needed to keep chip stand-by power low. The reduction of minority-carrier injection and low-voltage operation are also important design issues. Various low-power techniques have been proposed based on the circuit shown in Fig. 5.6. One example is the use of mode switching, from a high-power mode in the active period to low-power mode in the stand-by period. In the active period the generator must provide a I_{CP} large enough to compensate for the large I_{BB} of the peripheral circuit, while in the stand-by period it needs only a small I_{CP} to compensate for the small I_{BB} caused by refresh-only operations with an extremely slow cycle. Another example is the use of a level monitor, which detects the degradation of the V_{BB} output level so that the V_{BB} generator is switched to a high-I_{CP} mode to recover the level.

Figure 5.25 shows a V_{BB} generator that features two sets of charge-pump circuits [5.30]: a slow cycle ring oscillator (1) for supplying a small current during the retention and stand-by modes, and a fast cycle ring oscillator (2) for supplying a sufficiently larger current during the active cycle or when the level monitor detects that the V_{BB} level is high. Thus, it minimizes the retention current by shutting down the fast cycle circuit. Here, the CMOS inverters (IV_1 and IV_2) are waveform shapers, and inverter IV_3 is a buffer to drive a heavy capacitance (C_2). To suppress the current flowing into the substrate, Q_3 must be small (i.e. small W/L).

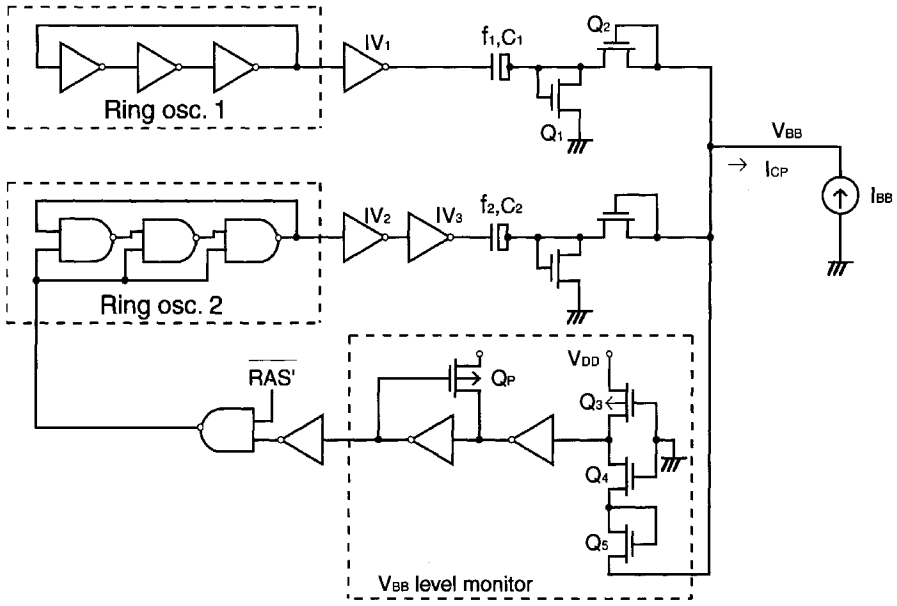


Fig. 5.25. A V_{BB} generator with mode switching and a level monitor [5.30]

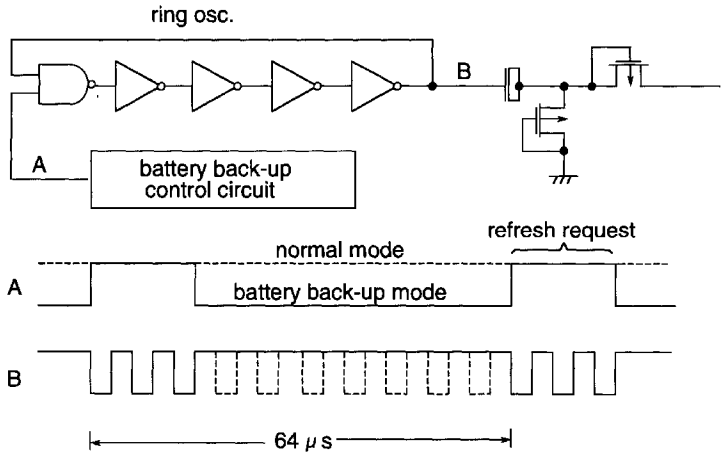


Fig. 5.26. A V_{BB} generator with pump-frequency control and output PMOS diodes [5.31]

Figure 5.26 shows another alternative which stops the oscillation of the V_{BB} generator while the DRAM chip is not in an active cycle [5.31]. In normal mode the ring oscillator provides a sufficient I_{CP} at a high frequency, while in battery back-up mode it oscillates intermittently only during the burst requests of the refresh operation. A negative V_{BB} is realized by extracting holes from the p-substrate by PMOSFETs at the output stage. The use of PMOSFETs there avoids minority-carrier (electrons) injection from the diode in the generator to the NMOS memory cells.

Figure 5.27 shows another V_{BB} generator [5.25] in which charges are alternately pumped twice a cycle by using PMOSFETs (Q_2, Q_4). In addition,

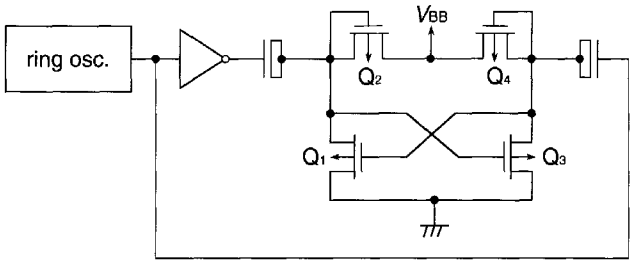


Fig. 5.27. A PMOS cross-coupled V_{BB} generator [5.25]

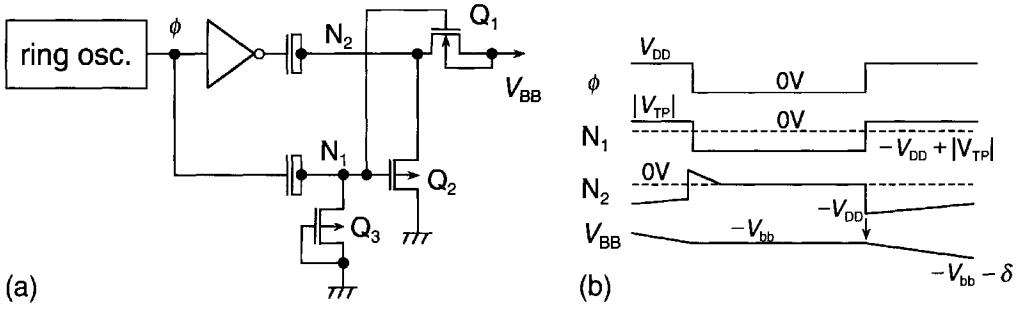


Fig. 5.28. A hybrid pumping V_{BB} generator [5.32]. (**a**) Circuit; (**b**) timing

the raised gate voltages of the cross-coupled Q_1 and Q_3 eliminate their V_T loss, offering a deeper V_{BB} of $-V_{DD} + V_T(Q_2, Q_4)$. The V_T drop is still an obstacle to obtaining a deep V_{BB}.

Figure 5.28 shows a V_{BB} generator that is suitable for low-V_{DD} operation [5.32]. All FETs except Q_1 are PMOSFETs. When clock Φ decreases to 0 V, the N_1 voltage also goes down to a negative voltage of $-V_{DD}+|V_{TP}|$ (V_{TP} is the V_T of Q_3), and thus Q_2 is turned on, so that the N_2 voltage is clamped to 0 V, and Q_1 is cut off. When Φ increases to V_{DD}, the N_1 voltage is clamped to $|V_{TP}|$ and thus Q_2 is cut off. At the same time, the N_2 voltage instantaneously drops to $-V_{DD}$ due to capacitive coupling. Then, N_2 and the substrate continue to be charged up and discharged, respectively, until both voltages are equilibrated, allowing the charge at N_2 to be pumped to the substrate. As a result, one clock cycle deepens V_{BB} by δ. Here, the pumping is performed without drop in V_T at Q_1, because the Q_1 gate voltage is sufficiently higher than the N_2 and substrate voltages during the equilibration process. By repetitive clock applications, the substrate is gradually discharged and the pumping ceases when V_{BB} reaches as low as $-V_{DD}$. Thus, perfect conversion of the power-supply voltage is established. In fact, a V_{BB} of -1.44 V at $V_{DD} = 1.5$ V has been obtained experimentally.

5.3 The Voltage Up-Converter

5.3.1 The Roles of the Voltage Up-Converter

The voltage up-converter has been widely used to eliminate the V_T drop in DRAM circuits and for programming Flash memory cells, as explained in Chap. 1. The requirements for the converter, however, are different for the two memories. The raised voltage (V_{DH}) necessary for DRAM is rather lower than that required for Flash memory. Instead, the cycle time (t_{cyc}) of the V_{DH} pulse for DRAM is rather shorter: In modern memory design, with $V_{DD} = 3.3\,V$ or $5\,V$, V_{DH} for the DRAM circuit ranges from $4\,V$ to $8\,V$ with a boost ratio (V_{DH}/V_{DD}) of $1.3 - 1.5$, while V_{DH} for the Flash memory is $12-20\,V$ with a boost ratio of $2.4 - 6.0$. The cycle time is about $100\,ns$ for DRAM, while it is $1\,\mu s$ at most even for the V_T-verify operation of the Flash memory. Thus, the V_{DH} generator for DRAM must provide more charge to compensate for the charge loss at the load every cycle, if the same load capacitance and conversion efficiency of the V_{DH} generator are assumed. In addition to the existing memories, future CMOS LSIs will need a V_{DH} generator to control V_T, as discussed in Chap. 8. In the following section, V_{DH} generation focused on DRAM applications is described.

In the past, three kinds of circuits in a DRAM chip have needed a raised voltage. They are the memory cell, the data-line isolation circuit, and the data-output buffer, as shown in Fig. 5.29.

- *The Memory Cell.* To perform a full write and full read operation while eliminating a drop in V_T at the memory cell, the word voltage must be raised to V_{DH} (so-called word bootstrapping).

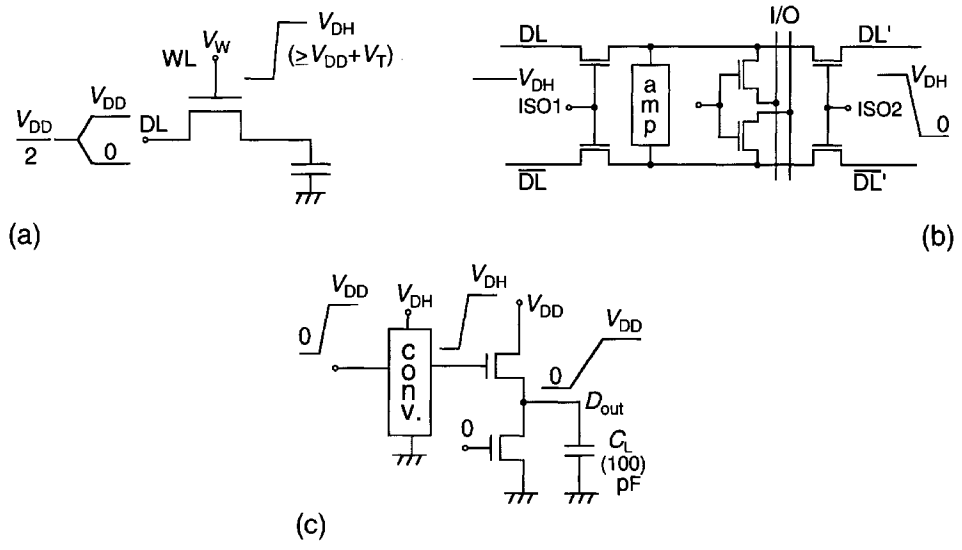


Fig. 5.29. Circuits using the boosted voltage (V_{DH}). (**a**) Memory cell; (**b**) data-line isolation circuit; (**c**) data-output buffer

- *The Data-Line Isolation Circuit.* In the shared I/O and CMOS amplifier (see Fig. 3.60) a raised voltage (V_{DH}) must be applied to the gate (ISOI or ISO2) of the data-line isolation MOSFETs, so that a data-input voltage from the common I/O line or an amplified cell-signal voltage is completely transferred to the cell without a drop in V_T.
- *The Data-Output Buffer.* In a conventional floating substrate, the data-output buffers have used only NMOSFETs [5.63] to avoid a possible CMOS latch-up caused by an inherently large spike current at the I/Os. Consequently, the gate voltage of the NMOS must be raised to cope with a drop in V_T that degrades the speed, especially at a lower V_{DD}. Note that triple-well structures that enable a grounded substrate permit the use of PMOSFETs without using V_{DH}.

The memory cell necessitates the highest V_{DH} of the three, which stems from the special requirements of the memory cell. The narrow channel of the cell FET, the narrowest within a chip, not only increases V_T and the substrate bias effect coefficient (K), but also the variation in V_T caused by variations in the fabrication process. As a result, for example, the maximum cell V_T is as large as 1.5 V whereas the V_T of standard FETs in the peripheral circuit is 0.5 ± 0.2 V, neglecting the K effect. What makes the matter worse is the fact that the necessary V_T of the memory cell gradually increases with increasing memory capacity, which comes from the requirement of the ever-increasing refresh time (t_{REFmax}), as seen in Fig. 4.17. V_{DH} must satisfy the following equation:

$$\begin{aligned} & V_{DH} \geq V_{DD} + V_T; \\ & \therefore \quad V_{DH}/V_{DD} \geq 1 + (V_T/V_{DD}). \end{aligned} \tag{5.9}$$

Thus, the boost ratio (V_{DH}/V_{DD}) is increased as V_{DD} is lowered under the ever-increasing V_T of the memory cell, although the ratio has been about 1.5 in conventional designs for $V_{DD} = 5$ or 3.3 V. The relationship between the necessary word voltage (V_W) and the boost ratio, and the data-line voltage (V_{DD}), is shown in Fig. 5.30 where it is assumed that $V_W = V_{DH}$ and that V_T is fixed. Here, it will be shown below that a fixed boost ratio would be hazardous in terms of device reliability: the worst S/N ratio of a memory cell is established at the minimum data-line voltage (V_{Dmin}), thus calling for $V_W \geq V_{Dmin} + V_T$. If the line of $V_W = 1.5V_{DD}$ with a fixed boost ratio of 1.5, which is for actual designs, is drawn at the V_{Dmin} value shown in the figure, the value of V_W at the maximum data-line voltage (V_{Dmax}) would be so high that device reliability could not be ensured due to a high stress voltage. Moreover, the excessive V_W – which would be higher than the device breakdown voltage (BV) at the burn-in voltage (V_{DB}) – would destroy devices. Thus, a boost ratio that is variable in accordance with V_{DD} is needed, as seen in the ideal case in the figure.

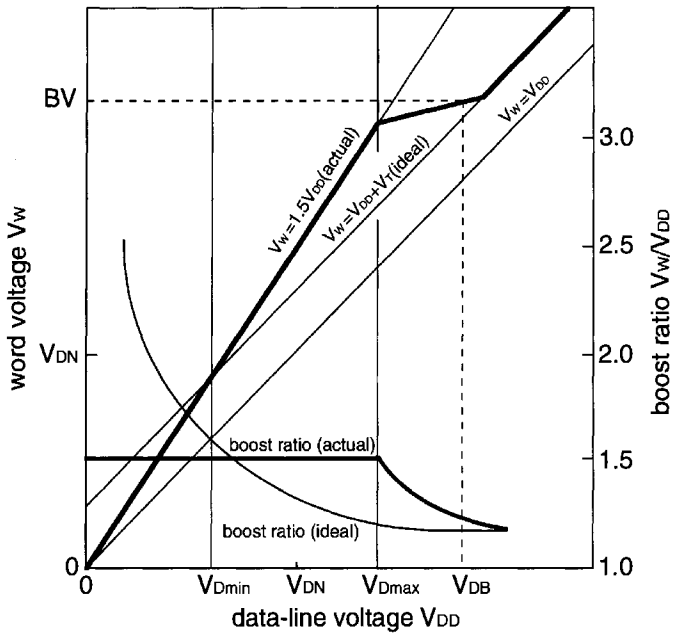


Fig. 5.30. The necessary word voltage and boost ratio [5.1]. V_{DN}, nominal data-line voltage; V_{Dmin} and V_{Dmax}, minimum and maximum data-line voltages with a voltage tolerance (usually $V_{DN} \pm 10\%$)

5.3.2 Design Approaches and Issues

Circuit Approaches. There are two approaches for generating the V_{DH} pulse. One is to boost by ΔV_B with a capacitor (C_B) after outputting a V_{DD} pulse, as shown in Fig. 5.31a. The other is to use a raised quasi-static power-supply voltage (V_{DH}) from an on-chip V_{DH} generator, as shown in Fig. 5.31b. The latter has the advantages of high speed and less V_{DH}-level variation: it eliminates the timing margin (τ_{D1}), which must be larger than the rise

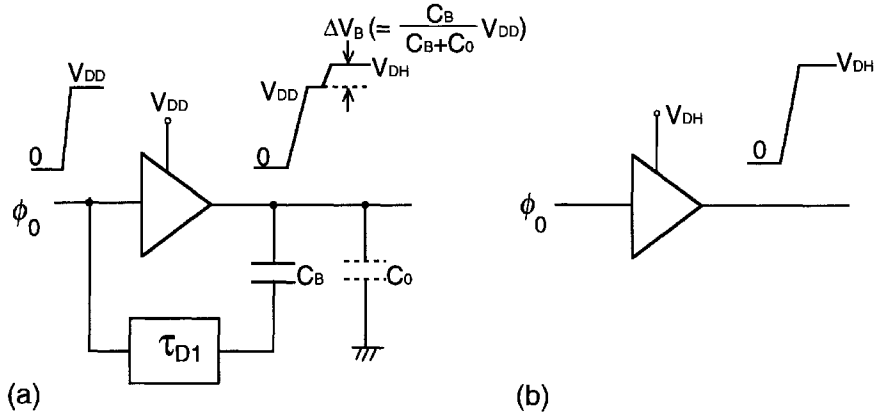


Fig. 5.31. The generation of boosted pulses using a boost capacitor (**a**) and a raised power-supply voltage (**b**) [5.1]

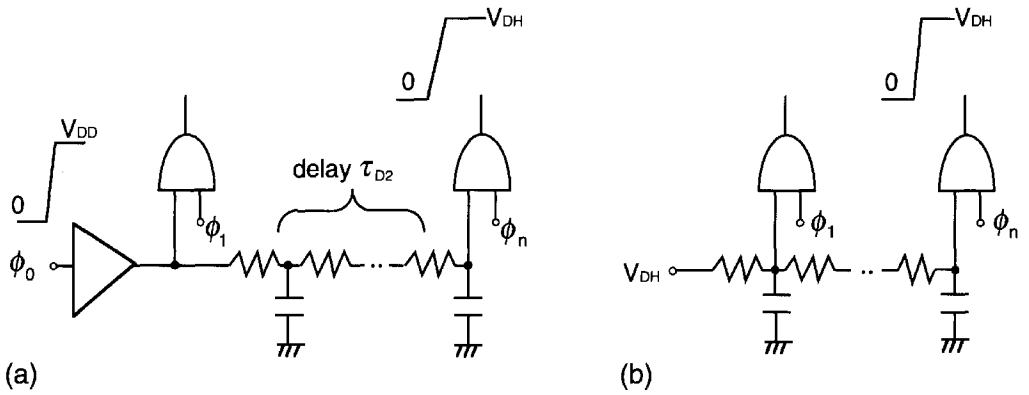


Fig. 5.32. Boosted pulse generation through AND logic [5.1]. (**a**) AND logic with a boosted pulse after traveling on a long line; (**b**) AND logic with a boosted power-supply voltage

time of the V_{DD} pulse. Note that the rise time is quite large for a heavy load capacitance, as in the word line. It also eliminates the line delay (τ_{D2}) for AND logic of the V_{DH} pulse traveling on a wiring line and a high-speed pulse (Φ_n), as shown in Fig. 5.32. This case is for word-line driving by using a decoder output pulse Φ_n. Furthermore, it could generate a stable V_{DH} of $V_{DD} + V_T$, which enables a variable boost ratio, by using a level detector, as discussed later. On the other hand, C_B boosting suffers not only the drawback of slow speed, but also has a fixed boost ratio that might destroy devices, as discussed before. C_B boosting makes it difficult to achieve a variable boost ratio, although this is partly realized by changing the number of pumping circuits according to V_{DD} [5.59], and by combining double boosting in a low-V_{DD} region with single boosting in a high-V_{DD} region, as discussed in Chap. 3 (see Fig. 3.51). It also suffers from large C_B/C_O variations, and thus V_{DH}-level variations, due to variations in the fabrication process. Thus, the V_{DH} generator approach is preferable.

Design Issues of the On-Chip V_{DH} Generator. In this section, the design issues of the V_{DH} generator are discussed, after the concept of V_{DH} generation has been explained. To design the generator, both the characteristics of the generator circuit itself and its load must be clarified, because they eventually determine the V_{DH} level.

The Generator Circuit. Figure 5.33a shows a generator composed of a basic capacitive voltage doubler and the load. The doubler consists of a pumping capacitor (C_P) and four switches ($S_1 - S_4$). The load capacitor C_L is much larger than the pumping capacitor C_P. Two non-overlapping clocks Φ_1 and Φ_2 drive the switches. During the turning on of Φ_1, V_{DD} is applied to capacitor C_P. During the turning on of Φ_2, capacitor C_P is charged and then discharged, so that a charge transport of $(2V_{DD} - V_{DH})C_P$ takes place between C_P and C_L. When the output voltage V_{DH} is equal to $2V_{DD}$, the charge pump is in

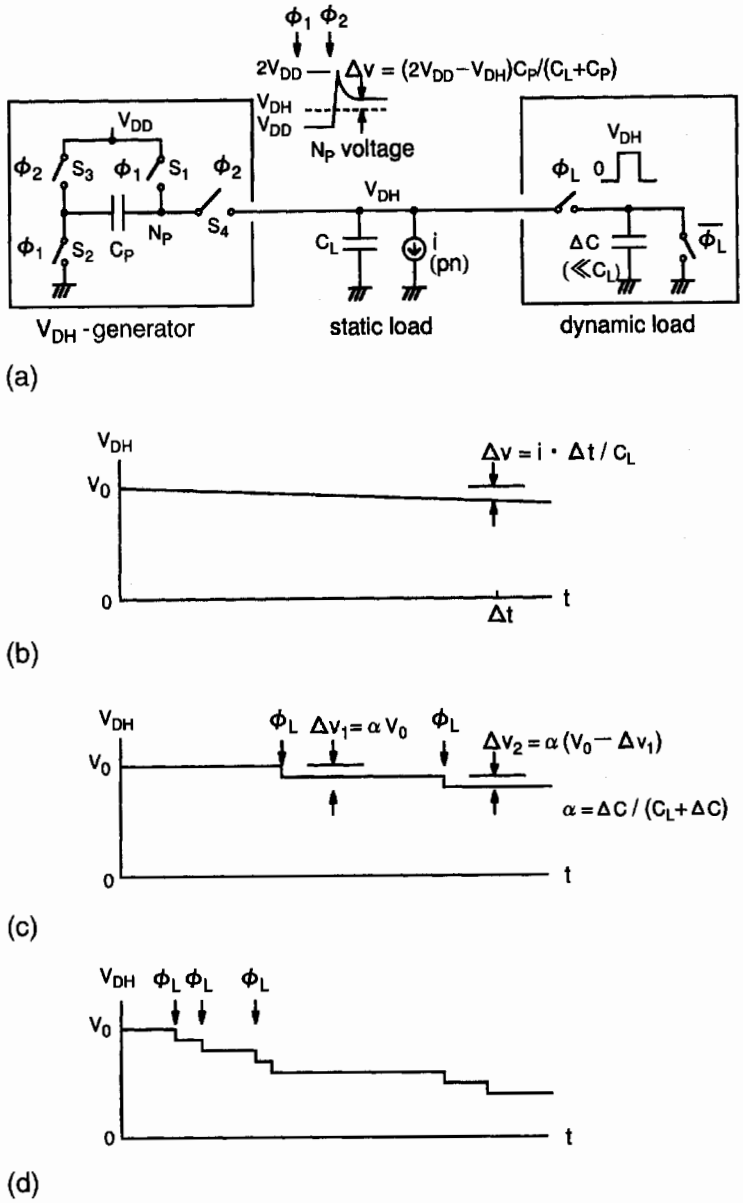


Fig. 5.33. V_{DH} degradation of the circuit (**a**) for various operational modes. It is assumed that there is no supply current from the generator. (**b**) is due to only the p–n junction current of the static load, (**c**) is due to only the refresh-operation current of the dynamic load, and (**d**) is due to only the random-access operational current of the dynamic load

equilibrium, and does not transport any charge. This is the case for no sink current (i.e. load current).

The Load Characteristics. Actually, two kinds of sink current exist at a DRAM load that consists of a static load and a dynamic load. One is a small dc current of $1-100\,\mu A$ that is always flowing at the back-biased p–n junctions in the static load. The other consists of large spike currents that flow only

when the dynamic load is activated. The spike current flows sequentially during every 16 µs cycle time for refresh operations of the stand-by mode, or flows at random during active periods. Here, the load currents for a 0.5 µm 16 Mb DRAM [5.1] are estimated as follows. The word line and the data-line isolation circuit, which operate at V_{DH}, consume 80 pC and 110–320 pC every cycle, respectively, thus consuming average currents of 0.7 mA and 1–3 mA at a cycle time of 110 ns. As for the data-output buffers of ×4 bits configuration, the charge and current are 60 pC and 1.3 mA at a page-mode cycle time of 45 ns. Thus, the total load current is 3–5 mA, although wider data-bit configurations and higher-speed burst modes further increase the current. During the stand-by mode, in which the data-output buffers are inactive and the chip operates at a slow refresh cycle of 16 µs, the total current is reduced to 15–30 µA, because it is inversely proportional to the cycle time. The currents degrade the V_{DH} level, as shown in the figure, assuming that there is no supply current from the generator.

The Combined Characteristics. When the generator is connected to the actual DRAM load, the output voltage V_{DH} settles at a certain voltage at which the supply current of the generator and the load current (I_L) are equalized. If I_L is a dc current, V_{DH} is calculated for the voltage doubler [5.72] in Fig. 5.34 as follows. In the voltage-doubling process, four phases can be distinguished. The first phase, T_{1e}, is when the Φ_1 switches are turned on. The second phase, T_1 is when they stay turning on. The third phase, T_{2e} is when the Φ_2 switches are turned on. The fourth phase, T_2 is when they stay turned on. During T_1, the charge on C_L decreases by $I_L T_1$. During T_2, the charge on C_L decreases by $I_L T_2 C_L/(C_P + C_L)$. In the stationary condition, during T_{2e}, the charge on C_L thus increases by the sum of the above-decreased charges, which is expressed as

$$\Delta Q(C_L, T_{2e}) = I_L \left(T_1 + T_2 \frac{C_L}{C_P + C_L} \right). \tag{5.10}$$

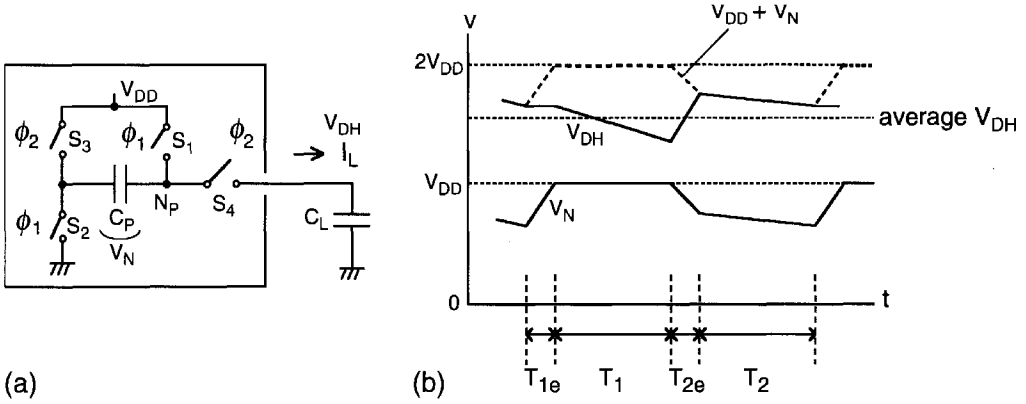


Fig. 5.34. A voltage doubler (**a**) and node voltages (**b**) [5.72]. V_N, voltage difference between C_P electrodes

On the other hand, during T_{2e} the charge on C_P decreases by the amount of charge that C_L has delivered to the load. The decreased charge is thus given as

$$\Delta Q(C_P, T_{2e}) = I_L \left(T_1 + T_2 \frac{C_L}{C_P + C_L} \right). \tag{5.11}$$

During T_2, the charge on C_P decreases by $I_L T_2 C_P/(C_P + C_L)$. Hence, the total decrease in charge on C_P equals $I_L(T_1 + T_2)$. This charge is delivered to capacitor C_P during T_{1e}. As can be seen in the figure, the average output voltage V_{DH} is given by

$$\begin{aligned} V_{DH} &= 2V_{DD} - \frac{\Delta Q(C_P, T_{2e})}{C_P} - \frac{1}{2}\frac{\Delta Q(C_L, T_{2e})}{C_L} \\ &= 2V_{DD} - I_L \left(T_1 + T_2 \frac{C_L}{C_P + C_L} \right) \left(\frac{1}{C_P} + \frac{1}{2C_L} \right) \\ &\simeq 2V_{DD} - I_L T_C / C_P, \end{aligned} \tag{5.12}$$

if $C_P \ll C_L$ and $T_C = T_1 + T_2$.

The ripple on the voltage V_{DH} is obtained as

$$V_{ripple} = \frac{\Delta Q(C_L, T_{2e})}{C_L} = \frac{I_L}{C_L} \left(T_1 + T_2 \frac{C_L}{C_P + C_L} \right) \simeq \frac{I_L T_C}{C_L}. \tag{5.13}$$

Note that the ripple must be suppressed to be as small as possible, because a large ripple may cause a high stress voltage in devices.

To obtain a well-regulated V_{DH} with a small level degradation and ripple, a small I_L and a large C_L are preferable. A fast pumping cycle and a large C_P are also essential, although they increase the power of the generator. For the dynamic load of DRAMs, however, V_{DH} regulation is more complicated. A small load current in the refresh operation makes regulation easy with the low power caused by slow cycle pumping. However, a large random-access current may result not only in a high power level in the converter, but also fatal degradation of the V_{DH} level. The V_{DH} level monitor scheme, which is discussed later, is an effective way to compensate for the level degradation. For example, as soon as a large load-current mode starts to operate, the V_{DH} level gradually decays toward the next equilibrated level. The level monitor detects a desired level on the way to the equilibrated level, and makes a further V_{DH} degradation stop, with more pumping current. The scheme eventually causes another ripple, which is different from the above. It is also desirable to have this compensated for within one active cycle by using a chip-enable clock (see Figs. 5.41 and 5.42).

Design Issues. The key issues of the generator can thus be summarized as a high but variable boost-ratio design, and the concurrent achievement of low power and high supply current, while keeping a well-regulated output voltage against dynamic load operation and variations in the fabrication process. Over-voltage protection is also indispensable.

5.3.3 High Boost-Ratio Converters

Figure 5.35 shows a conventional charge-pumping circuit [5.73]. The clock Φ_1 works as switches S_2 and S_3, controlled by clocks Φ_1 and Φ_2, in Fig. 5.34; while a precharge NMOS M_P works as switch S_1. A transfer NMOS M_1 acts as the series switch S_4. In this circuit configuration, however, the V_T drops at M_P and M_1 and the substrate-bias effect of M_1 reduce V_{DH}, causing degradation of the boost ratio V_{DH}/V_{DD}, especially at low voltages. A feedback charge-pumping circuit [5.38], shown in Fig. 5.36, solves this problem. By utilizing a feedback MOSFET, node N_1 is precharged at V_{DD}, instead of $V_{DD} - V_T$, while P_1 is "high". Thus, node N_1 is boosted to $2V_{DD}$ when P_{1B} goes up, a full V_{DD} is stored at node N_2 while P_1 is "low", and node N_2 is boosted to $2V_{DD}$ when P_1 goes "high". This feedback technique can eliminate a voltage loss due to the threshold voltage of the NMOSFETs. Figure 5.37 shows the whole V_{DH} gnerator. The major components of the converter are a main charge-pumping circuit, as shown in Fig. 5.36, a transmission gate M_1, and a control pulse generator for the transmission gate. The control pulse generator consists of a $3V_{DD}$ driver to drive the gate of the transmission gate, a $3V_{DD}$ booster to generate a voltage of $3V_{DD}$ and to feed it to the $3V_{DD}$ driver, and a $2V_{DD}$ booster, which generates a gate control signal for the $3V_{DD}$ driver. A timing diagram for the drive pulse is shown in Fig. 5.37b. Node N_3 is boosted from

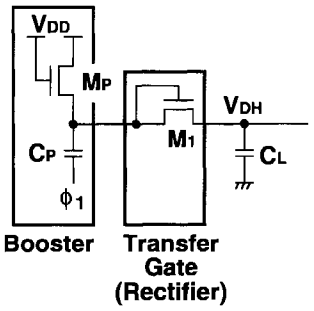


Fig. 5.35. A conventional pumping circuit [5.73]

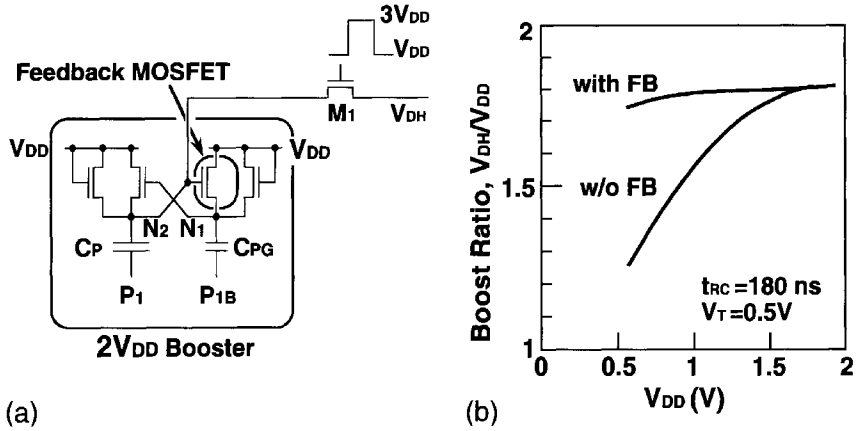


Fig. 5.36. A feedback charge-pump circuit [5.38]. (**a**) Circuit; (**b**) boost ratio

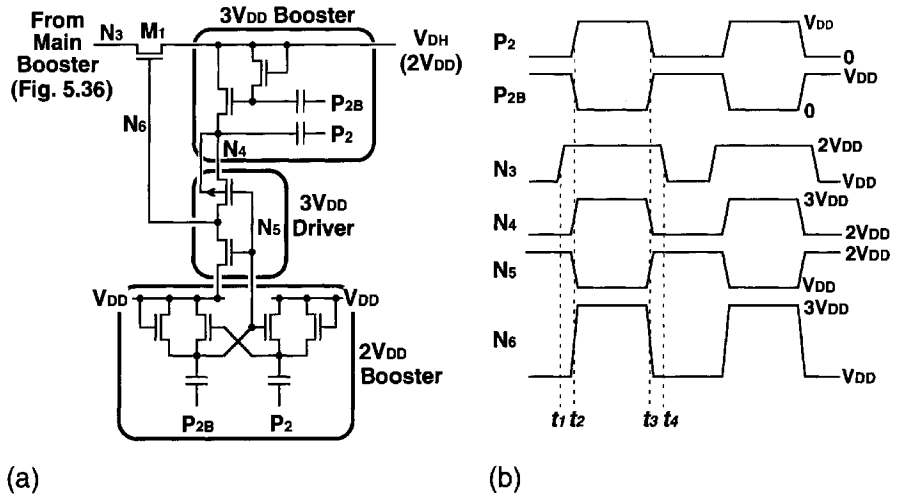


Fig. 5.37. A V_{DH} generator with a feedback charge-pump circuit [5.38, 5.73]. (**a**) Schematic of circuit; (**b**) timing diagram

V_{DD} to $2V_{DD}$ at t_1. Before t_1, node N_6 stays at V_{DD}, since both the node N_4 and node N_5 voltages are $2V_{DD}$. When P_2 goes "high" and P_{2B} goes "low" at t_2, node N_4 is boosted from $2V_{DD}$ to $3V_{DD}$ and node N_5 is pulled down from $2V_{DD}$ to V_{DD}. Then, node N_6 goes from V_{DD} to $3V_{DD}$, the transmission gate M_1 is turned on, and $2V_{DD}$ is obtained at the output of the V_{DH} converter. In actual, the boost ratio is degraded by parasitic capacitances at the nodes. However, an almost constant boost ratio as large as 1.8 was obtained at $V_{DD} = 0.5-2\,V$ for a V_{DH} generator with feedback, while a decrease in the boost ratio was observed when V_{DD} was less than 1.5 V for a conventional generator without feedback [5.38], as shown in Fig. 5.36b. One drawback is that the transfer NMOSFET M_1 needs a highly boosted gate voltage to eliminate the drop in V_T.

Figure 5.38 shows another voltage doubler [5.74], using PMOS transfer gates. Although a boosted gate voltage is not needed, reverse biasing of the junctions must be ensured, with the n-well (i.e. substrate of the PMOSFET)

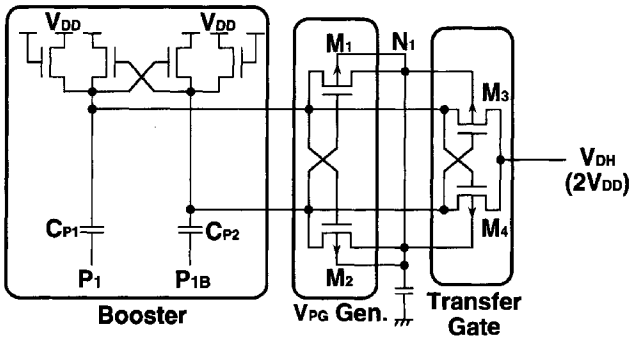


Fig. 5.38. A voltage doubler, using a dual PMOS series switch and n-well switching [5.74]

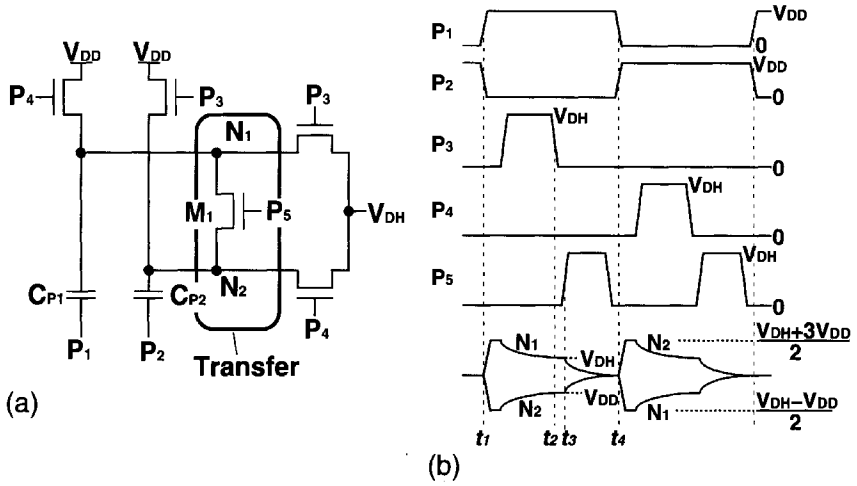


Fig. 5.39. A charge transfer pumping circuit [5.73, 5.75]. (**a**) Schematic of circuit; (**b**) timing diagram

voltage always higher than the source voltage. This is done by two sets of cross-coupled PMOSFETs. M_3 and M_4 are the transfer gates to alternately pump the load. M_1 and M_2 quickly give the highest voltage to either of the n-wells of M_3 and M_4, to ensure the reverse bias of the junctions.

Figure 5.39 shows a charge transfer pumping circuit [5.73, 5.75]. The circuit would operate as a pair of conventional pumps if the transfer MOSFET M_1 was eliminated. In this circuit, one pumping node (for example, N_1) is finally at V_{DH}, while the other pumping node (N_2) is at V_{DD} after applications of clocks P_1 and P_2 followed by clock P_3 or clock P_4, if each of clocks P_3 and P_4 has a high enough level. When clock P_5 is activated, both nodes are equalized at $(V_{DH} + V_{DD})/2$ by M_1. They are boosted to $(V_{DH} + 3V_{DD})/2$ or dropped to $(V_{DH} - V_{DD})/2$ by the succeeding clocks P_1 and P_2. The boosted voltage contributes to the charge pumping to the load by activating P_3 or P_4. If we suppose that C_{P1} and C_{P2} both equal C, the delivered current I_{DH} to the load is $C(3V_{DD} - V_{DH})/t_c$ (where t_c is the pumping cycle time). This is large enough compared with $2C(V_{DD} - V_{DH})/t_c$ of a pair of conventional pumps. In fact, experiments have revealed that it generates more than twice as much I_{DH} compared with the conventional pump at $V_{DH}/V_{DD} = 2.5/1.5\,\mathrm{V}$, and even at $V_{DH}/V_{DD} \geq 2.0$ it can still generate a charge.

5.3.4 Low-Power, High Supply Current Converters

Practical generators based on the above-described generator design have been proposed for commercial DRAMs using a V_{DD} of either 5 V or 3.3 V. Their common concept, which is similar to that of the V_{BB} generator, is the use of a level monitor, feedback techniques to obtain a well-regulated V_{DH}, and switching of the supply-current of the generator. An over-voltage protection circuit is also commonly used.

Figure 5.40 shows a V_{DH} generator for increasing the voltage from 5 V to 7 V [5.57], featuring two sets of converters. To maintain the V_{DH} level with a low power during the stand-by period, only low-frequency pumping is carried out by a ring oscillator (OSC). When the chip is activated by enabling an external clock ($\overline{RAS}$), the resultant internal clocks (Φ_1, Φ_2, and $\overline{\Phi_{PH}}$) activate another pumping circuit. Thus, charges are injected into the load every active cycle. A voltage clamper, which is omitted in the figure, protects the devices from over-voltage by connecting node G_4 to a constant voltage such as V_{DL} through series-connected diodes. The current consumption of the converter is 3 mA at the 60 ns cycle time of a 5 V 1 Mb BiCMOS DRAM.

Figure 5.41 shows a V_{DH} generator with a level monitor [5.65] to detect the V_{DH} level. When the level is low, a ring oscillator starts to operate so that a pumping capacitor C_P is driven and thus V_{DH} is raised. When V_{DH} reaches a certain level, the level monitor turns the ring oscillator off, so that charge injection to the V_{DH} load stops. A more detailed explanation is as follows. The whole circuit is activated by an ENABLE signal, which is a slow-cycle pulse, such as a refresh control pulse of 16 µs cycle time during the stand-by period, or a random pulse generated synchronously with activation of a chip-enable signal. Here, it is assumed that the V_T of the level-detecting MOSFET Q_M is the same as that of the memory-cell FET. First, operation during the stand-by period is explained, with the inputting of a slow-cycle refresh-control (ENABLE) pulse. This input makes the Q_M gate high, because node a rises to a high level. When refreshing the cells, a word-line activation lowers V_{DH}. As long as the resulting V_{DH} is higher than $V_{DD} + V_T$, however, Q_M continues to be switched on, allowing Q_1 and Q_2 to form a current mirror. Thus, node b is held at a high level if the transconductance of Q_2 is sufficiently larger than that of Q_0 or Q'_0, and the detecting speed (i.e. the time until Q_2 detects a high level at node a) is fast enough. Thus, the ring-oscillator enable

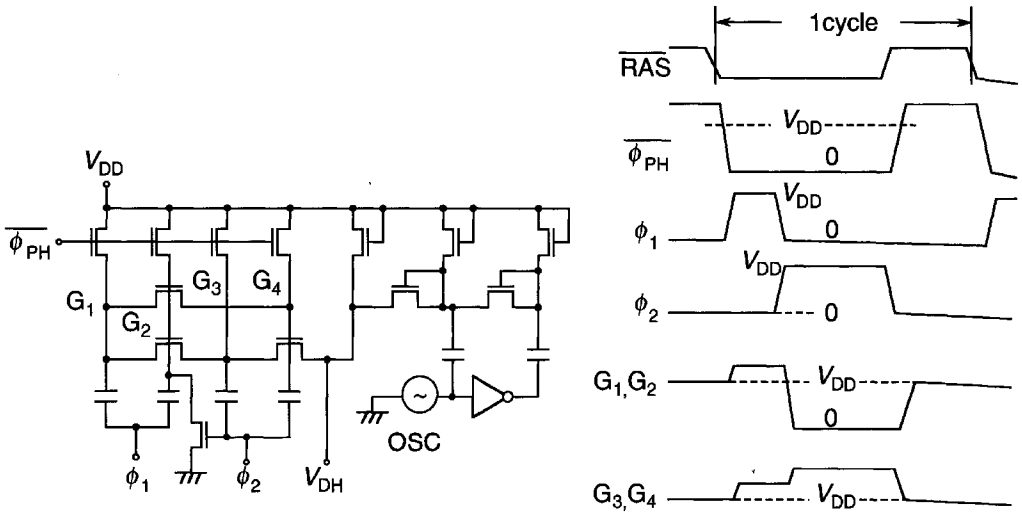


Fig. 5.40. A V_{DH} generator consisting of two sets of generators [5.57]

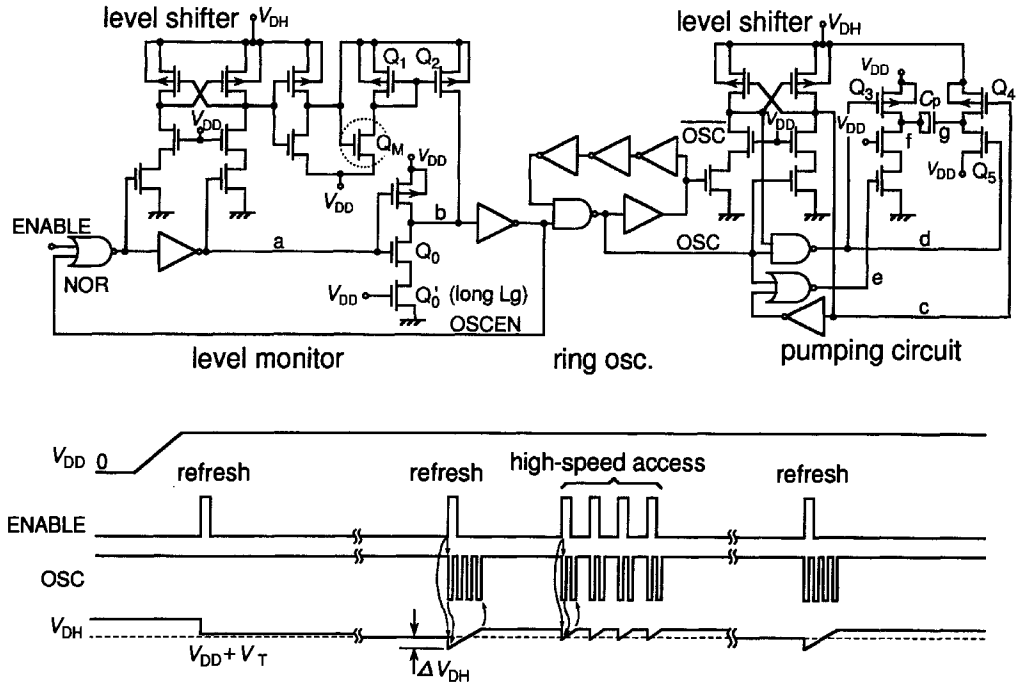


Fig. 5.41. A V_{DH} generator with a level monitor [5.63]

signal (OSCEN) is kept at a low level, turning the oscillator off. If V_{DH} is lower than $V_{DD} + V_T$, however, node b is discharged by Q_0 and Q_0' because Q_2 is off, which allows the oscillator to start to oscillate with a high-level OSCEN. The resultant low OSC makes not only nodes c, d, and e high, but also nodes f and g change from V_{DD} and V_{DH} to 0 V and $V_{DD} - V_T$, respectively. Even when ENABLE is turned off, the input NOR gate allows node a and OSCEN to maintain the same levels, allowing the oscillator to continue oscillating and the Q_M gate to be at V_{DH}. Thus, the level monitor is ready for V_{DH} level detection. The succeeding high OSC eventually makes node f change from 0 V to V_{DD}. Since Q_4 is on while Q_5 is off at this moment, the voltage change by V_{DD} that is developed across C_P injects charges into the V_{DH} load, so that V_{DH} is raised slightly. Subsequent repetitive OSC pulses gradually increases V_{DH}, until it reaches $V_{DD} + V_T$. As soon as V_{DH} exceeds $V_{DD} + V_T$, Q_M is turned on, so that the oscillation stops and thus V_{DH} stays at $V_{DD} + V_T$. Thus, V_{DH} is maintained at the same value by application of the refresh-control pulses. Even when the chip is accessed at random, V_{DH} degradation is compensated for in the same manner by application of an ENABLE pulse generated synchronously with a chip-enable pulse. Here, the insertion of the NMOSFETs (for example, Q_0') with their gates biased to V_{DD} is to protect the relevant MOSFETs from the high stress voltage of V_{DH}, as discussed previously (see Fig. 5.23). The separation of Q_0 and Q_0' not only reduces the stress voltage of Q_0, but also allows Q_0', whose channel length is sufficiently large, to almost determine the effective transconductance of the

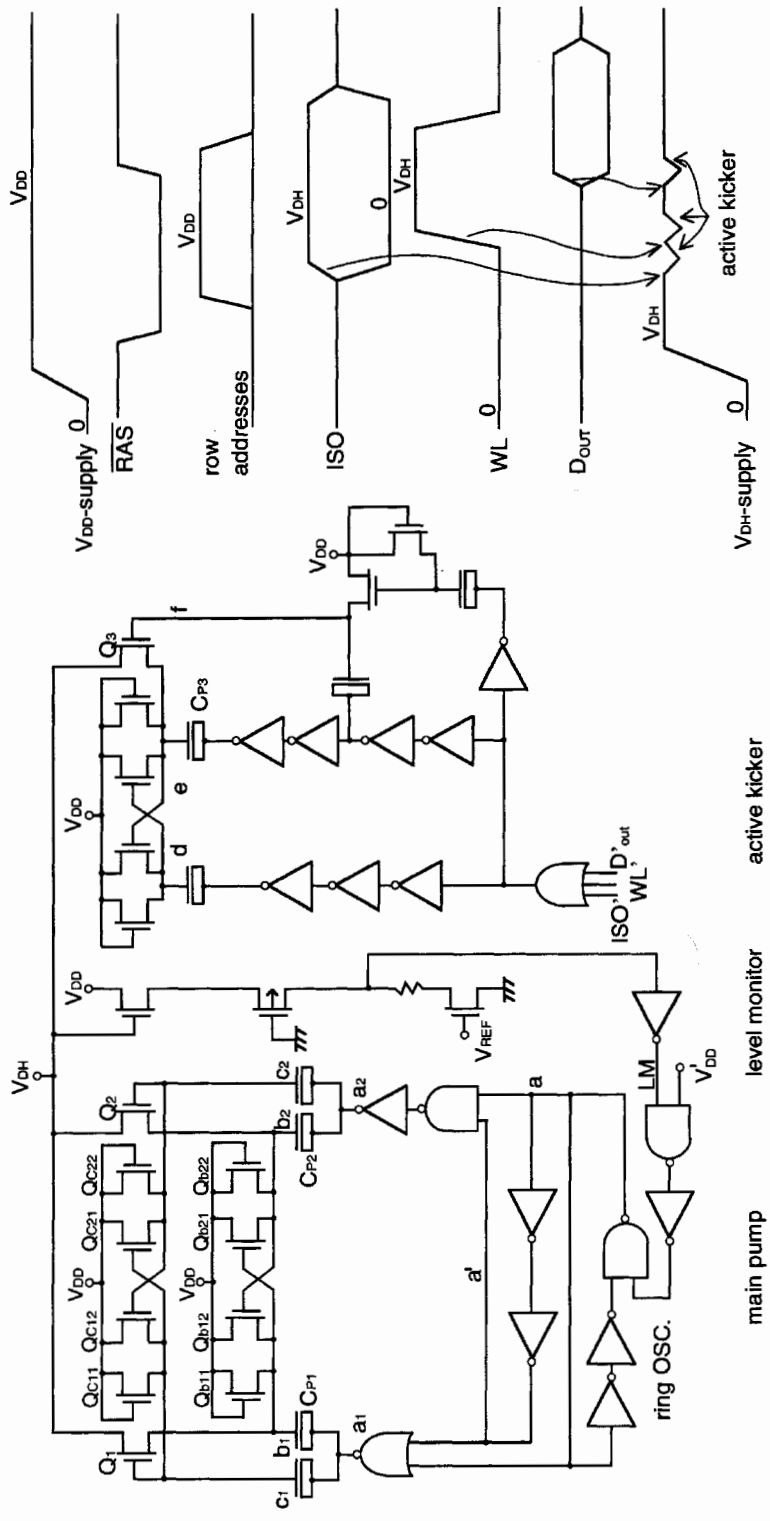


Fig. 5.42. A V_{DH} generator with two sets of pump circuits and a level monitor [5.64]

series-connected Q_0 and Q'_0. Thus, the transconductance is less sensitive to variations in the fabrication process, and the capacitance of node a is reduced.

Figure 5.42 shows another V_{DH} generator [5.64]. It features the use of two kinds of charge-pumping circuits, a main pump and an active kicker, to provide charges for a purely capacitive output load. The main pump compensates for a small charge loss due to the leakage current of the load. It is driven by a ring oscillator, which is activated when V_{DH} is lower than the level determined by the level monitor. The active kicker operates synchronously with the load-circuit operations, such as ISO driving (Fig. 5.28), word-line driving, and output buffer driving. This circuit compensates for a large charge loss due to the load circuit operations. Slow-cycle operation of the main pump and extremely slow-cycle operation of the active kicker in data-retention mode provide a minimized retention current.

The V_{DH} operation of data-output (D_{out}) buffers is becoming more difficult. The ever-increasing need for wider data-bit I/O and higher-speed burst (column) modes are responsible for this difficulty. Figure 5.43 shows a D_{out} buffer scheme [5.65] in which a level-compensation circuit (LCC) is attached to each D_{out} buffer. When a high (H) data that consumes V_{DH} charge is inputted after V_{DH} has been degraded, V_{DH} is compensated by the charge-pumping. For $V_{DD} = 3.3\,V$, a 25 ns cycle time, and $\times 16$ data-bits I/O, this V_{DH} generator consumes 8 mA for all "H"s, 4 mA for eight "H"s and 0 mA for all "L"s, enabling reduction of the average current.

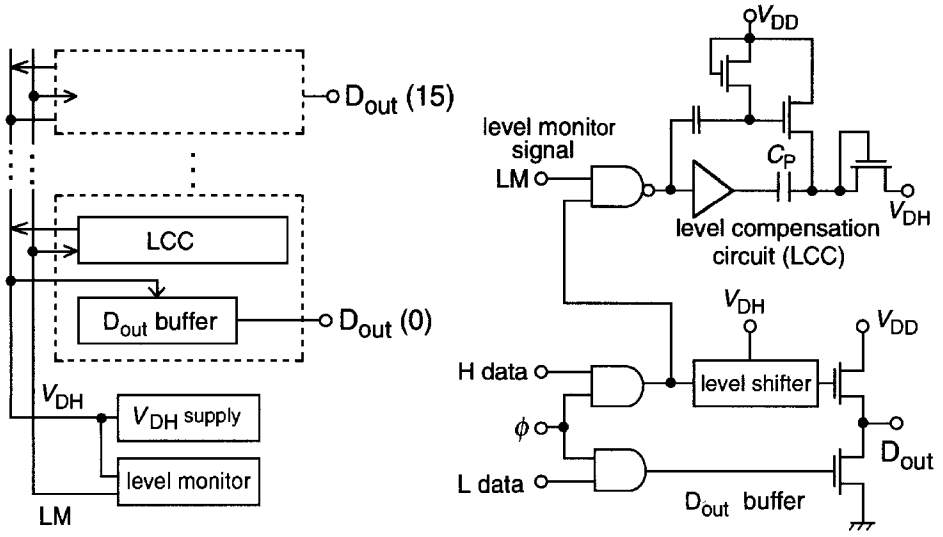


Fig. 5.43. The operation of a D_{out} buffer at V_{DH} [5.65]

5.4 The Voltage Down-Converter

5.4.1 The Roles of the Voltage Down-Converter

An on-chip voltage down-converter has been widely used in commercial DRAMs since the 16 Mb generation, through continuing research [5.33–5.36]. The converter is a key to realizing power-supply standardization and low cost in general-purpose DRAMs, and high speed and the battery operation of LSIs. Here is a summary of the main reasons why the converter is needed [5.1].

Realization of Power-Supply Standardization. The standard supply voltage of a DRAM is dictated by the system supply, which is not scaled fast enough to keep up with advances in CMOS device technology. Therefore, a voltage down-converter that can bridge the supply gap between the system and internal core-circuit devices needs to be integrated on the chip [5.35]. This converter can adjust the converted voltage in accordance with the lowering of the breakdown voltage of the ever more miniaturized devices, while keeping the external power-supply voltage (V_{DD}) the same for as long as possible. Power-supply standardization has actually been realized with the help of a converter since the 16 Mb generation. Thus, quadrupling of memory capacity every three years has been possible with the same V_{DD}, despite the ever-lowering device breakdown voltage. Moreover, successive chip-shrinking with a fixed memory capacity (Fig. 3.2) has been realized, to reduce the bit cost with the same V_{DD}.

Realization of Battery Operations. There has been an increasing number of papers [5.37–5.39] aimed at battery operation of LSIs. The unregulated supply voltages from various batteries inevitably require a wide range of operating voltages to be supplied to LSIs. An on-chip voltage down-converter can fix the internal operating voltage well, regardless of such external supply voltages. It also protects the internal core circuit against high voltages, across a wide range of voltage variation.

Realization of Flexible Designs. The voltage-down converter allows the output voltage (V_{DL}, i.e. the internal operating voltage) to be adjusted in accordance with any variation in the design parameters, such as the supply voltage (V_{DD}), the temperature, and the device parameters. This capability offers the prospect of flexible designs for future LSIs. For example, if V_{DL} is fixed well and is sufficiently lower than V_{DD}, the chip speed is almost constant, independent of any V_{DD} variation. Furthermore, any chip performance could be obtained, if V_{DL} was intentionally adjusted by tracking temperature and fabrication-process variations. The converter could even compensate for a speed loss caused by mobility degradation at a high temperature, if V_{DL} was increased with increasing temperature. Similar compensation could be carried out for the longer channel length and higher V_T that result from volume production. The converter could even compensate for the leakage current of a memory cell, which increases at a high temperature due to an increased

cell-stored voltage [5.40, 5.41]. A universal power-supply scheme [5.42], which accepts a wide range of external power-supply voltages, is one application of the voltage down-converter.

The practical voltage down-converters proposed so far can be categorized as the NMOS-output converter [5.67–5.70] and the PMOS-output converter, each of which has a feedback loop, as shown in Figs. 5.44 and 5.45. The NMOS-output converter has fewer stability problems, and does not require a large output capacitor, due to the inherently low-ohmic output impedance of the source-follower structure – unlike the PMOS-output converter. However, the maximum driving current of the NMOSFET, developed at the highest gate voltage, is small. This is because the effective gate–source voltage, expressed as

$$V_{\mathrm{GSmax}} - V_{\mathrm{TN}} = V_{\mathrm{DD}} - V_{\mathrm{DL}} - V_{\mathrm{TN}}, \tag{5.14}$$

is small. In particular, it is prominent for a low dropout voltage (i.e. a small difference between V_{DD} and V_{DL}). Note an increased threshold voltage (V_{TN}) due to the substrate-bias effect (i.e. body effect). For example, the resultant V_{TN} is as large as 1 V, which is not suitable for low dropout applications

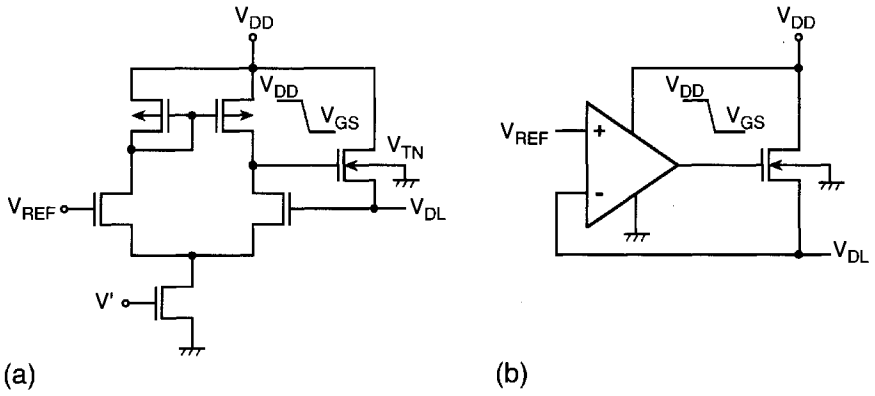


Fig. 5.44. An NMOS-output converter (**a**) and its equivalent circuit (**b**)

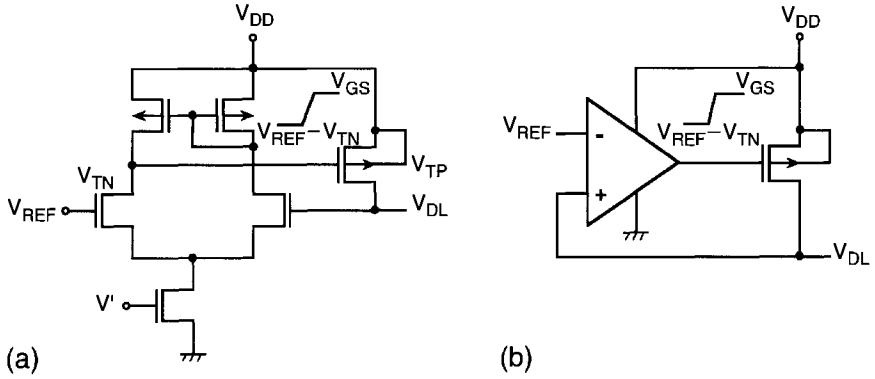


Fig. 5.45. An PMOS-output converter (**a**) and its equivalent circuit (**b**)

such as $V_{DD} = 3.3\,V$ and $V_{DL} = 2.5\,V$. On the other hand, the PMOS-output converter (Fig. 5.45) allows the PMOSFET to turn on more strongly with a higher gate–source voltage and a smaller threshold voltage, due to the abscence of any body effect. In this case the maximum effective gate–source voltage is expressed as

$$V_{GSmax} - |V_{TP}| = V_{DD} - V_{REF} + V_{TN} - |V_{TP}|, \tag{5.15}$$

because the lowest PMOS gate voltages, i.e. the lowest drain voltages of paired NMOSFETs in the saturation condition, must be $V_{REF} - V_{TN}$. Thus, lower V_{DD} and lower dropout applications are acceptable. However, the circuit needs a compensation circuit for feedback-loop stability, because it is a negative feedback circuit consisting of a high-gain amplifier and a high-gain PMOSFET. Moreover, the FET size tends to be large due to the low transconductance of the PMOSFET.

Several attempts have been made to solve the above problems for both circuits. Figure 5.46a shows a boosted gate NMOS-output converter [5.69]. Since the gate voltage can exceed V_{DD}, the converter can have a low dropout voltage. However, the pump and amplifier circuit need to be designed with care, so as not to destroy the low-voltage devices. If a voltage doubler is used as the pump, the devices could undergo a stress as high as $2V_{DD}$. So, longer-channel devices, device stacking, and other circuit techniques are necessary to reduce the stress on each device. The same applies to the amplifier, which typically has stacked devices. In addition, the pump and amplifier consume a large area and power, which originate from the boosted gate voltage. The output voltage division shown in Fig. 5.46b increases the voltage swing at the PMOS gate with a resulting lowered V_{REF}. However, this in turn poses a difficulty, in obtaining a large constant amplifier current, and thus may cause degraded V_{DL} regulation, as discussed later. A buffer circuit inserted between the amplifier and the PMOSFET [5.46, 5.71] may deliver a large voltage swing to the PMOS gate, in spite of a small swing at the amplifier

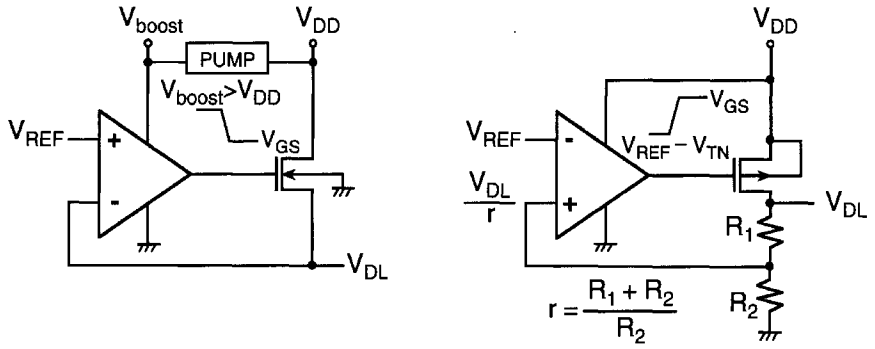


Fig. 5.46. An improved NMOS-output converter (**a**) [5.69] and a PMOS-output converter (**b**)

output. However, the high gain of the buffer tends to make the feedback loop unstable. Eventually, however, the PMOS-output converter has become widely used in commercial DRAMs through the solution of the stability problems. Phase compensation using an on-chip capacitor and resistor is the key to stabilization, as discussed later.

5.4.2 The Negative-Feedback Converter and Design Issues

In the past, a negative-feedback voltage down-converter [5.33] utilizing the V_T of a MOSFET as a reference has been proposed. Eventually, however, the negative-feedback circuit consisting of a PMOSFET and a comparator (differential amplifier) [5.40, 5.43, 5.46] shown in Fig. 5.47 has become standard for commercial DRAM designs. The current-mirror differential amplifier compares the output voltage (V_{DL}) with the reference voltage (V_{REF}). The resulting output signal controls the gate of PMOSFET(Q), so that V_{DL} is

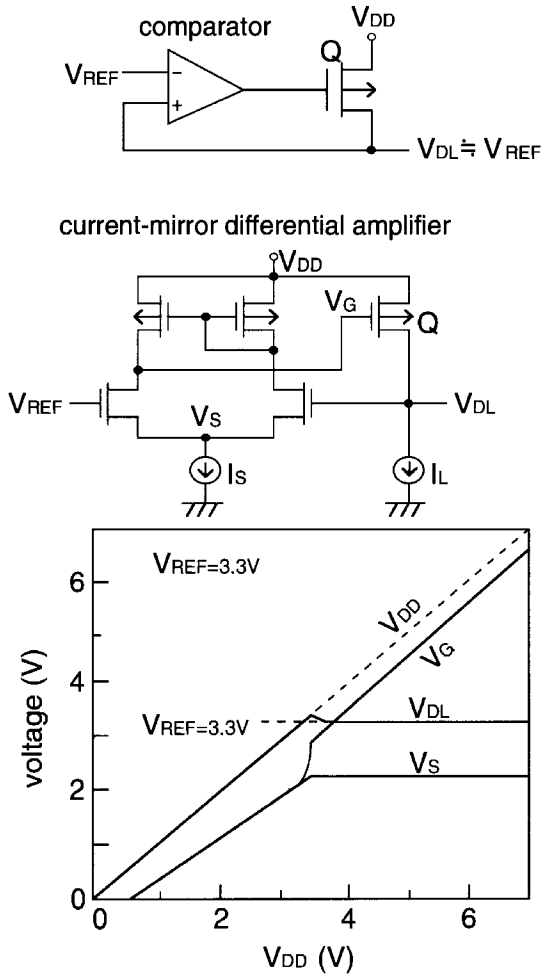


Fig. 5.47. A typical voltage down-converter [5.43, 5.44, 5.46]

stabilized. In other words, when the load current (I_L) flows to ground, the V_{DL} starts to drop, since the output FET(Q) works as an impedance. As soon as the resultant V_{DL} becomes lower than V_{REF}, Q starts to turn on, so that it charges up the load. When the resultant V_{DL} exceeds V_{REF}, however, the Q gate voltage is raised, so as to stop the charge up. As a result, the circuit compensates for the drop in V_{DL}. A larger drop in V_{DL} allows the output to be charged up more quickly, because the feedback loop responds more quickly with a larger amplified voltage at the Q gate. A larger size of Q (i.e. a larger W/L) also charges up the load more quickly. The output voltage is stabilized in this manner. The dc characteristics of the converter [5.46] are shown in the figure.

The converter must offer a well-regulated output voltage that is immune to any variation of the operating conditions. V_{DL} must be regulated well even under a dynamic load, where the heavy load capacitance is dynamically changed and various current pulses flow as a result of the load (i.e. internal core-circuit) operation. The voltage regularity of V_{DL} is determined by the circuit characteristics, such as the response time, the current-driving capability, and the feedback-loop stability of the converter, and by the load characteristics. Thus, to design the converter, a deep understanding of both the circuit and the load characteristics is required. Otherwise, V_{DL} may result in oscillation or ringing waveforms.

The Characteristics of the Converter Circuit. The setting accuracy of the output voltage (V_{DL}) is governed by that of the reference voltage (V_{REF}). For example, V_{DL} is almost equal to V_{REF} for the circuit shown in Fig. 5.46. However, if V_{REF} is quite a high voltage, such as 3–4 V, the accuracy of V_{REF} (i.e. V_{DL}) against parameter variations is not satisfactory, as described later. A high V_{REF} also lowers the voltage swing at the PMOS gate, as discussed previously. Hence, a V_{DL}-division circuit [5.40, 5.45], shown in Fig. 5.48a, has also been widely used. In this circuit, a high V_{DL} is divided using resistors (or MOSFETs), R_1 and R_2, so that the resultant reduced voltage (V_{DL}/r) is compared with a more accurate V_{REF} of about 1 V. However, the circuit suffers from the following drawbacks. First, the response time of the feedback loop is slow, which causes a possible loop instability, discussed later. In addition to the difficulty in obtaining a large amplifier current (I_S), a large RC delay caused by the large input capacitance of the comparator and the high resistivity of R_1 and R_2 is responsible for the slow response. Unfortunately, the capacitance is quite large since the comparator must drive the large MOSFET (Q), and the resistivity also needs to be high enough to reduce the stand-by current. Second, V_{DL} division in a noisy location where large load-current pulses exist, requires careful design regarding phase compensation to avoid loop oscillation, which is discussed later, and careful layout to reduce the noise coupled to the input of the highly sensitive amplifier from its noisy output. Third, the resistors for V_{DL} division need a large layout area to achieve an extremely high resistivity of MΩ with a poly-Si layer, whose sheet

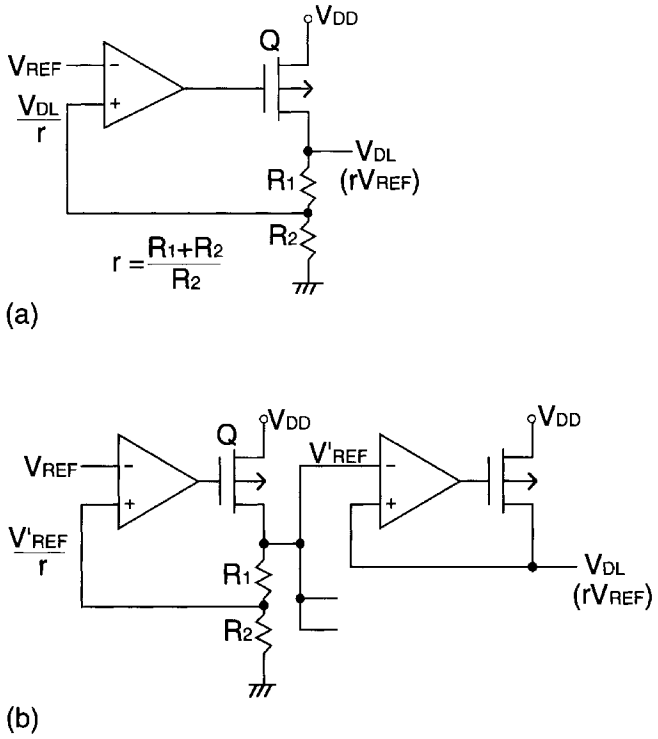


Fig. 5.48. Practical converters for DRAMs [5.1]. (**a**) A voltage-division converter; (**b**) a hybrid converter

resistivity is about 100 Ω/□. Multiple uses of the converter – for example, one for the memory array and another for the peripheral circuit – further increase the chip area. Figure 5.48b shows a hybrid converter [5.46–5.49] consisting of two-stage converters; the first stage is a V_{DL} division converter, and the second stage is a conventional converter. The features of the first stage – less load capacitance and no load-current pulse (in fact, a pure dc circuit) – enable a simpler circuit and layout design even with maintenance of the high resistances for R_1 and R_2. In addition, the first stage could be common to a number of second-stage converters, causing reductions in the chip area and stand-by current, if multiple use of the converter is necessary.

The Load Characteristics. The DRAM load features a dynamically changing capacitance and current. For example, in the memory array of a 0.5–0.6 μm CMOS 16 Mb chip a spike current as large as 150 mA can flow during a short period of 20 ns when many data lines, whose total capacitance is as large as 1300 pF, are simultaneously charged up at cell-signal amplification, as shown in Fig. 5.49. This is just a dynamic load, since the current and the capacitance that the converter can see are as small as almost 0 μA and 100 pF until charging takes place. Such is the case for the peripheral circuit with a peak current of 100 mA, as a result of the charging up of internal nodes such as the many address signal lines.

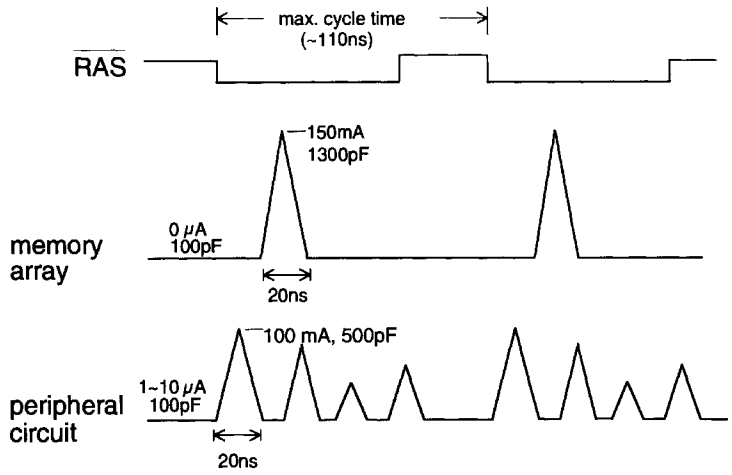


Fig. 5.49. The transient currents of a CMOS 16 Mb DRAM [5.1]. 0.5–0.6 μm 16 Mbit, $V_{DD} = 5\,V$, $V_{DL} = 3.3\,V$, $V_{DL}/2$ precharge

There are specific circuit-blocks that must operate at the external supply-voltage (V_{DD}), while the remaining blocks can operate at the down-converted supply voltage (V_{DL}). In addition to the converter itself, the data output (D_{out}) buffers must operate at V_{DD}. The requirement for the buffers comes from the chip-to-chip interface specification. Even if there is no limitation from the specification, it is difficult for the V_{DL} converter to manage the current overloading when many (1 to 32) buffers quickly drive their heavy off-chip capacitances (usually 100 pF per buffer). To protect the MOSFETs from a higher stress voltage of V_{DD}, the V_{DD}-based circuits must use longer-channel-length MOSFETs or another MOSFET must be inserted (see Q_2 in Fig. 5.23). The chip-input buffers on the addresses, data inputs, write enable, and clocks could be designed with V_{DL} or with both V_{DL} and V_{DD}. The operating voltages of the remaining circuit-blocks depend on the differences in degree between V_{DD} and the breakdown voltage of the MOSFETs in the blocks. If there is a large difference, both the peripheral circuit and the memory array must operate at V_{DL} [5.30, 5.43–5.45]. In the actual design, two kinds of converter are used separately instead of one converter, as in Fig. 5.50a. This separation provides a quieter V_{DL} to the peripheral circuit, as a result of eliminating interference from the noisy memory array where larger pulse currents flow. In the case of a small difference, only the memory array can operate at V_{DL} [5.33, 5.50, 5.51], as shown in Fig. 5.50b, to reduce the memory-array power dissipation that dominates the chip power. V_{DD} operation for MOSFETs in the peripheral circuit can be managed through slight adjustments of the MOSFET sizes. If the difference is marginal, specific circuits such as word drivers, which operate at the highest voltage in the chip, must operate based on V_{DL} [5.39, 5.52], as shown in Fig. 5.50c. The word voltage is created by boosting V_{DL}, instead of V_{DD}, so as to protect

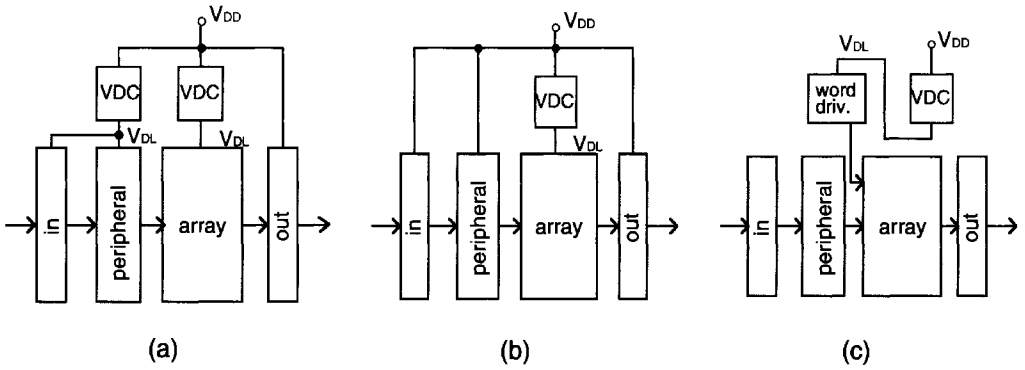


Fig. 5.50. The application of a V_{DL} generator to various blocks in a DRAM chip [5.1]. VDC, voltage down-converter. (**a**) To all blocks except the D_{out} buffers; (**b**) to the memory array only; (**c**) to the word driver only

the relevant MOSFETs from an excessive stress voltage caused by boosting unregulated V_{DD}.

In addition to the need for well-regulated, stable, low-power, and small-sized designs, the burn-in test capability is a key to converter design. Otherwise, a fixed V_{DL} from the converter prevents a high enough stress voltage from being applied to the internal core circuit, even if the external supply voltage (V_{DD}) is increased during the burn-in test.

In the following sections, the design of PMOS-output negative-feedback converters is described in detail. First, an optimum design for the converter is discussed, followed by the phase compensation for loop stabilization. Then, reference-voltage generators, burn-in test circuits, and the voltage trimming technique are explained. Finally, low-power circuits are discussed.

5.4.3 Optimum Design

In this section, the guiding design principles are given, based on the circuit analysis for the conventional voltage down-converter shown in Fig. 5.51a. The converter is a negative-feedback amplifier that consists of two-stage amplifiers; a current-mirror differential amplifier for the first stage, and a PMOS-output transistor (Q_6) for the second stage, which also works as an amplifier. The output voltage – that is, the down-converted voltage (V_{DL}) – is the internal supply voltage, and thus the common source voltage of the PMOSFETs in the CMOS core circuit. The total current of the core circuit is just the load current of the converter. Although V_{DL} is lower than V_{REF} when the load current starts to flow, it can be quickly suppressed by the negative feedback. However, once V_{DL} is higher than V_{REF}, it is no longer suppressed: it is then held at the V_{DL} level for a long time, because the increased Q_6 gate voltage cuts off Q_6. The condition of V_{DL} being higher than V_{REF} could be developed when a positive coupling from other wiring to the V_{DL} line occurs. Thus, a small I_B is needed to discharge V_{DL} down to V_{REF}. Here is

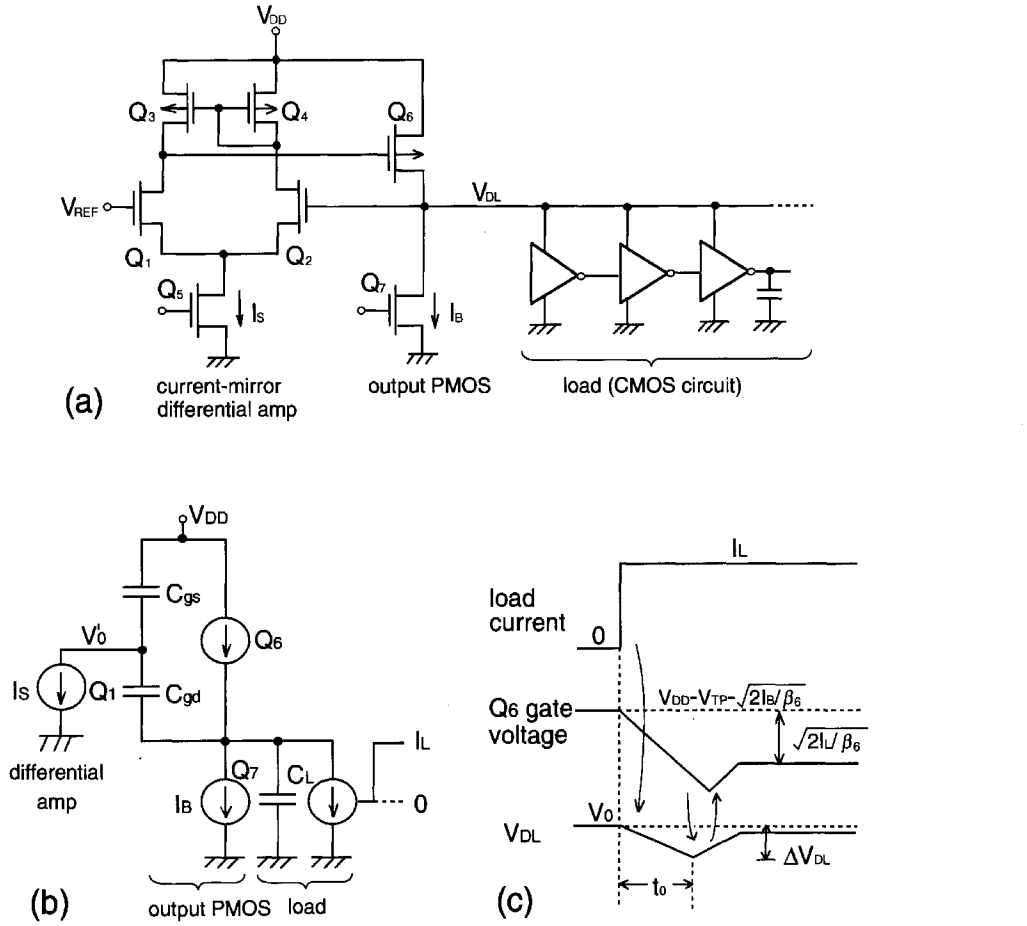


Fig. 5.51. A voltage down-converter connected to the load (**a**), the large-signal equivalent circuit (**b**), and the pulse response (**c**) [5.53]

an analysis [5.53] to obtain the optimum design parameters of the converter, assuming the same channel length for all of the MOSFETs and the same channel width for the paired MOSFETs.

When the load current (I_L) is at a maximum, the source–drain voltage (V_{DS6}) of Q_6 in the saturation region is given by

$$V_{DS6} = V_{DD} - V_{DL} \geq V_{GS6} - V_{TP} = \sqrt{\frac{2I_L}{\left(\frac{W_6}{L}\beta_{P0}\right)}}; \tag{5.16}$$

$$\therefore \frac{W_6}{L} \geq \frac{2I_L}{V_{DS6}^2 \beta_{P0}}, \tag{5.17}$$

where W_6, V_{TP}, and β_{P0} are the channel width, threshold voltage, and conductance of Q_6, respectively, and a small $I_B (\ll I_L)$ is assumed. Here, the voltages are all expressed using absolute values. On the other hand, when I_L is zero, a small I_B makes V_{DL} equal to V_{REF}, allowing a half-I_S to flow through Q_3 in the saturation region. Hence,

$$V_{DS3} = V_{GS6} = \sqrt{\frac{2I_B}{\beta_6}} + V_{TP}, \tag{5.18}$$

$$V_{DS3} \geq V_{GS3} - V_{TP} = \sqrt{I_S \Big/ \left(\frac{W_3}{L}\beta_{P0}\right)};$$

$$\therefore \frac{W_3}{L} = \frac{W_4}{L} \geq \frac{I_S}{V_{DS3}^2 \beta_{P0}}. \tag{5.19}$$

Here, the constant current (I_S) is given by using a large signal equivalent circuit in Fig. 5.51b, as follows. When the load current steps up from 0 to I_L, V_{DL} instantaneously falls, so that more current flows through Q_1 although an equal current of a half-I_S flows into Q_1 and Q_2. Instead, the current flowing through Q_2 and Q_4 is reduced so as to also reduce the current of Q_3 in the current-mirror circuit. Eventually, the current difference between Q_1 and Q_3 discharges the Q_6 gate. Therefore, when V_{DL} falls to the extent that I_S fully flows through Q_1, I_S discharges the Q_6 gate. Since Q_6 is in the saturation region and the gate–drain capacitance (C_{gd}) is thus negligible,

$$C_{gs}\frac{dV_0'}{dt} + I_S = 0 \tag{5.20}$$

is obtained, where V_0' is the Q_6 gate voltage. By using the initial V_{GS6} expressed by (5.18) and (5.20),

$$\int_{V_{DD}-V_{TP}-\sqrt{\frac{2I_B}{\beta_6}}}^{V_0'} dV_0' = \int_0^t -\frac{I_S}{C_{gs}}\,dt, \tag{5.21}$$

$$\therefore V_0' = -V_{TP} + V_{DD} - \sqrt{\frac{2I_B}{\beta_6}} - \frac{I_S}{C_{gs}}t \tag{5.22}$$

are given. Moreover, V_{DL} is expressed as

$$C_L\frac{dV_{DL}}{dt} + (I_L + I_B) - \frac{\beta_6}{2}(V_{DD} - V_0' - V_{TP})^2 = 0. \tag{5.23}$$

By substituting (5.22), we can obtain V_{DL} as

$$V_{DL} = \frac{1}{6}\frac{I_S^2\beta_6}{C_L C_{gs}^2}t^3 + \frac{1}{2}\frac{\sqrt{2I_B\beta_6}I_S}{C_L C_{gs}}t^2 - \frac{I_L}{C_L}t + V_0, \tag{5.24}$$

where V_0 is the initial value of V_{DL}. The time (t_0) when V_{DL} falls to the lowest value is obtained by setting $dV_{DL}/dt = 0$, as

$$t_0 = -\frac{C_{gs}}{I_S}\sqrt{\frac{2I_S}{\beta_6}} + \frac{C_{gs}}{I_S}\sqrt{\frac{2(I_L + I_B)}{\beta_6}}. \tag{5.25}$$

V_{DL} at this time is derived from substituting (5.25) for (5.24), as follows:

$$V_{DL} = \frac{1}{3}\frac{\sqrt{2}\{-I_B^{1.5} - 2I_L^{1.5} + 3\sqrt{I_B}(I_L + I_B)\}C_{gs}}{C_L I_S \sqrt{\beta_6}} + V_0. \tag{5.26}$$

The first term on the right-hand side denotes the output-voltage variation (ΔV_{DL}). Assuming that $I_L \gg I_B$, and that C_{gs} is two-thirds of the total gate capacitance, I_S is given as

$$V_{DL} = \Delta V_{DL} + V_0, \tag{5.27}$$

$$\Delta V_{DL} = -\frac{4}{9}\sqrt{\frac{2}{\beta_6}}\frac{C_{OX}LW_6}{C_L}\frac{I_L^{1.5}}{I_S}, \tag{5.28}$$

$$\therefore I_S = -\frac{4}{9}\sqrt{\frac{2}{\beta_6}}\frac{C_{OX}LW_6}{C_L}\frac{I_L^{1.5}}{\Delta V_{DL}}, \tag{5.29}$$

where C_{OX} is the gate capacitance per unit area. It is obvious that a smaller I_L, a larger β_6, and a larger I_S are effective in reducing ΔV_{DL}.

The sizes of the input NMOSFETs (Q_1, Q_2) are obtained by using the gain (G_{01}) of the amplifier, as follows. The output voltage of the current-mirror amplifier, when the input voltages are equal (i.e. $V_{DL} = V_{REF}$), is $V_{DD} - V_{GS4}$ because $V_{GS3} = V_{GS4}$. When V_{DL} differs from V_{REF}, a variation, $G_{01}(V_{DL} - V_{REF})$, is added to the above output voltage. Since the resulting voltage is the Q_6 gate voltage, the following steady state condition is established:

$$(V_{DD} - V_{GS4}) + G_{01}(V_{DL} - V_{REF}) = V_{DD} - V_{GS6}. \tag{5.30}$$

Hence, by using (5.16) and $V_{GS4} = \sqrt{I_S/\beta_4} + V_{TP}$,

$$G_{01} = \frac{\sqrt{I_S/\beta_4} - \sqrt{2I_L/\beta_6}}{\varepsilon V_{REF}}, \tag{5.31}$$

$$\varepsilon = (V_{DL} - V_{REF})/V_{REF},$$

where ε is the setting accuracy of V_{DL} for V_{REF}. As discussed in Chap. 2 the MOSFET parameters are expressed as follows:

$$G_{01} = r_1 g_{m1}, \tag{5.32}$$

$$r_1 = \frac{1}{(\lambda_P + \lambda_N)\frac{1}{2}I_S}, \tag{5.33}$$

$$g_{m1} = \sqrt{I_S \frac{W_1}{L}\beta_{N0}}, \tag{5.34}$$

where λ_p and λ_N are channel-length modulation constants for the PMOSFETs (Q_3, Q_4) and the NMOSFETs (Q_1, Q_2), respectively. Consequently, the sizes of Q_1 and Q_2 are obtained as

$$\frac{W_1}{L} = \frac{W_2}{L} = \left\{\frac{G_{01}(\lambda_P + \lambda_N)}{2}\right\}^2 \frac{I_S}{\beta_{N0}}. \tag{5.35}$$

Here, both Q_5 and Q_7, which must always operate in the saturation region to give a constant current, have wide voltage margins. Therefore, their sizes are easily obtained, if their gate voltages are determined so as to achieve the necessary I_S and I_B (usually in μA).

Here is an example of converter design [5.53]. MOSFET parameters such as $W_6 \geq 2000\,\mu m$, $W_3 = W_4 \geq 250\,\mu m$, $W_1 = W_2 \geq 290\,\mu m$, $I_S = 1.6\,mA$, and $r_1 = 8.9\,K\Omega$ were obtained for the design conditions of $V_{DD} = 5\,V$, $V_{DL} = 3\,V$, $V_{DS6} = 2\,V$, $V_{REF} = 3\,V$, $I_L \leq 100\,mA$, $I_B = 2.5\,\mu A$, $\Delta V_{DL}/V_{DL} \leq 10\%$, $(V_{DL} - V_{REF})/V_{REF} \leq 1\%$, $C_L = 1\,nF$, $L = 1.2\,\mu m$, $\beta_{P0} = 3 \times 10^{-5}\,S/V$, $\beta_{N0} = 8 \times 10^{-5}\,S/V$, $g_{m1} = 4.9\,mS$, $C_{OX} = 2.3\,fF/\mu m^2$, $V_{TP} = 0.5\,V$, $\lambda_P = 0.1\,/V$, and $\lambda_N = 0.04\,/V$. In fact, an actual $0.5\,\mu m$ 16 Mb DRAM with a converter [5.53] that used similar parameters yielded experimental results of $\Delta V_{DL}/V_{DL} = 6\%$ and $(V_{DL} - V_{REF})/V_{REF} = 2\%$.

5.4.4 Phase Compensation

Stable Conditions. The key to designing the negative feedback amplifier shown in Fig. 5.52 is to cope with its inherent instabilities, such as ringing or oscillation. It is well known that an amplifier starts to oscillate if it has a phase difference between the input (v_i) and the output (v_o) of over 180° and a gain of over one [5.1, 5.54]. For example, when only a half-cycle of a sinewave ($v_i = V \sin \omega t$) is inputted to an amplifier whose gain (G) is one and whose delay time (τ) equals a half of the cycle time (t_c) of the input waveform, the output continues to oscillate as a result of feedback, as shown in the figure. If $G > 1$ the amplitude of the output is amplified with time, while if $G < 1$ it is attenuated. Thus, the parameter set of $G = 1$ and a phase delay of 180° denotes the critical conditions for oscillation. The stabilization

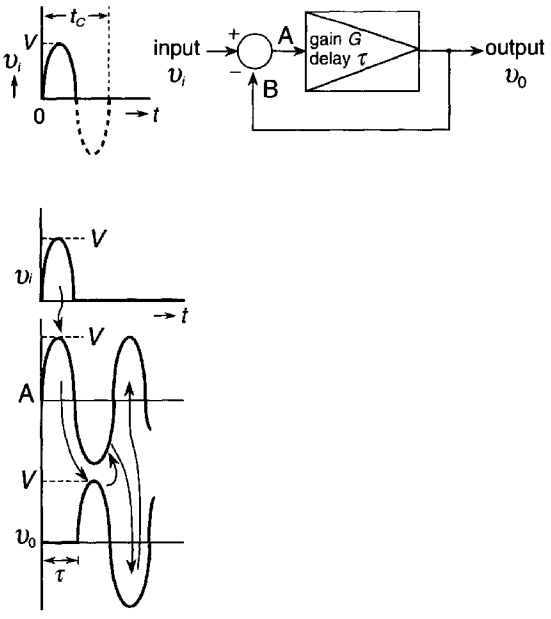


Fig. 5.52. Oscillation in a negative feedback amplifier with $G = 1$ and $\tau = t_c/2$ [5.1]

of such an amplifier is realized by adjusting the circuit parameters so that the phase at the frequency corresponding to $G = 1$ is smaller than 180°. Although an increase in the difference from 180° (i.e. a phase margin) stabilizes the amplifier more, the necessary phase margin must usually be larger than 45° [5.54]. Increasing the phase margin is called phase compensation. The stability of an amplifier whose transfer (gain) function, $G(j\omega)$, is expressed using the complex frequency characterized by $j\omega(= 2\pi f)$, is analyzed by examining both the gain and the phase characteristics of the function. The gain characteristics are examined on the plane of $\log\omega$ (rad/time) and $20\log_{10}|G(j\omega)|$ (decibels, dB) while the phase characteristics are examined on the plane of $\log\omega$ and phase Φ. Before going into detail, we investigate here how the gain and phase behave on the planes, citing examples of typical circuits.

There are some circuits, such as a one-stage amplifier, whose transfer function is generally expressed as

$$G(j\omega) = \frac{G_0}{\left(1 + j\frac{\omega}{\omega_P}\right)} \tag{5.36}$$

where ω_P is the pole and G_0 is the gain at a low frequency. Hence, the gain and the phase are expressed as follows:

$$\begin{aligned} |G(j\omega)| &= \left| G_0 \left(1 - j\frac{\omega}{\omega_P}\right) \Big/ \left[1 + \left(\frac{\omega}{\omega_P}\right)^2\right] \right| \\ &= G_0 \Big/ \sqrt{\left(1 + \frac{\omega}{\omega_P}\right)^2}, \\ 20\log|G(j\omega)| &= 20\log G_0 - 10\log\left[1 + \left(\frac{\omega}{\omega_P}\right)^2\right]; \\ \therefore \Phi &= -\tan^{-1}\left(\frac{\omega}{\omega_P}\right). \end{aligned}$$

Figure 5.53 shows the gain and phase characteristics. When ω is increased, the gain and phase are about $20\log G_0$ and 0° at $\omega \ll \omega_P$, respectively. However, they are $20\log G_0 - 3$ (dB) and −45° at $\omega = \omega_P$, and $20\log G_0 + 20\log\omega_P - 20\log\omega$ (dB) and −90° at $\omega \gg \omega_P$. In other words, the gain curve is flat at $\omega < \omega_P$, but it decreases by 3 dB at $\omega = \omega_P$, and then decreases by 20 dB per decade (dec.) in ω, approaching zero (i.e. $G = 1$). As for the phase, it is 0° at $\omega = 0$. However, it is delayed with increasing ω, reaching −45° at $\omega = \omega_P$, and −90° at $\omega = \infty$. Therefore, the phase margin is 90° at $G = 1$, which is larger than 45°, implying that a one-stage amplifier is inherently stable.

There are other circuits, such as phase-compensation circuits, whose transfer functions are expressed as

$$G(j\omega) = G_0\left(1 + j\frac{\omega}{\omega_Z}\right),$$

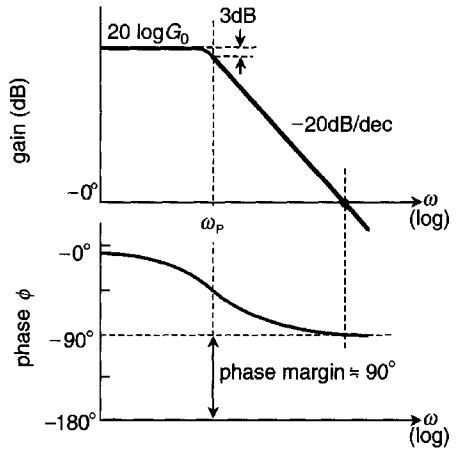


Fig. 5.53. The gain and phase characteristics of a one-stage amplifier [5.1]

where ω_Z is the zero. The role of ω_Z on the planes is opposite to that of ω_P, as analyzed in the same manner as above. The gain increases by 3 dB at $\omega = \omega_Z$, and increases at a rate of 20 dB/dec. at $\omega = \omega_Z$. The phase advances to the positive with an increasing ω, reaching $+45°$ at $\omega = \omega_Z$, and $+90°$ at $\omega = \infty$.

There is another type of circuit, such as the converter, whose transfer functions are characterized by two series-connected transfer functions, $G_1(j\omega) \cdot G_2(j\omega)$. The total transfer function, $G(j\omega)$, is as follows:

$$G(j\omega) = G_1(j\omega) \cdot G_2(j\omega) = \frac{G_{01}G_{02}}{\left(1 + j\frac{\omega}{\omega_{P1}}\right)\left(1 + j\frac{\omega}{\omega_{P2}}\right)}. \tag{5.37}$$

Thus, the total gain, $|G(j\omega)|$, is the sum of $|G_1(j\omega)|$ and $|G_2(j\omega)|$ on the $20 \log G$–ω plane. The total phase, Φ, is also the sum of Φ_1 and Φ_2. Figure 5.54 shows the gain and phase characteristics. The total gain has two poles, ω_{P1} and ω_{P2}, and it decreases at a rate of -20 dB/dec. at $\omega_{P2} < \omega < \omega_{P1}$, and then at -40 dB/dec. at $\omega > \omega_{P1}$. The total phase is $-45°$ at $\omega = \omega_{P2}$, and $-135°$ at $\omega = \omega_{P1}$, reaching $-180°$ at $\omega = \infty$. When Fig. 5.54 is compared with Fig. 5.53, it is obvious that the series connection of the transfer functions creates a phase delay and reduces the phase margin.

The Voltage Down-Converter and Phase Compensation. In this section, a voltage down-converter is analyzed with its ac small-signal equivalent circuit shown in Fig. 5.55a, and then the stable condition is given. Here, a load resistor R_L is added, because it affects the total characteristics of the converter.

When a small signal is inputted, the converter shown in Fig. 5.55a is expressed as a superposition of a dc-component circuit (i.e. a fixed dc-bias circuit) without any input signal and a small signal ac-component circuit. Therefore, ac analysis can be carried out by extracting only the small signal

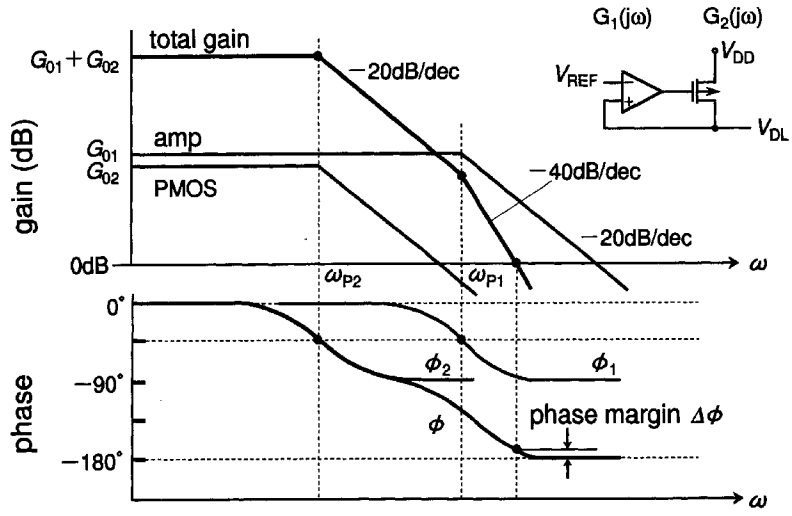


Fig. 5.54. The gain and phase characteristics of the voltage down-converter [5.1]

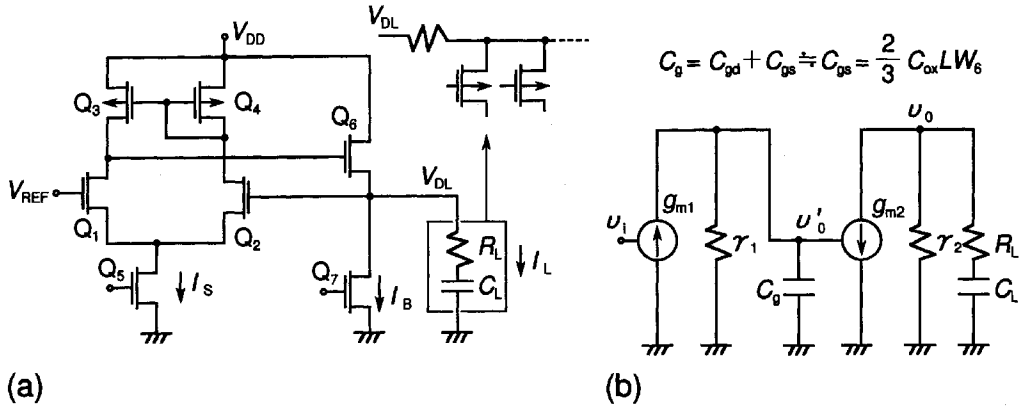


Fig. 5.55. A voltage down-converter (**a**) and its open-loop small-signal equivalent circuit (**b**) [5.53]

ac-component circuit shown in Fig. 5.55b. Here, to simplify the analysis of feedback loop stability, the loop is made open. Through investigating the whole gain and phase characteristics of the open-loop circuit, we can estimate the characteristics of the actual closed-loop circuit. In the following, we will study how to obtain an equivalent circuit, shown in Fig. 5.55b.

Figure 5.56 shows the differential amplifier and its equivalent circuit. Capital letters are used for dc biasings, while small letters are used for small signals. The sizes of the paired MOSFETs (Q_3 and Q_4, Q_1 and Q_2) are the same, and the output resistance and the transconductances of the NMOS and the PMOS are r_n and r_p, and g_{mn} and g_{mp}, respectively. All of the MOSFETs are in saturation mode, and the r_p of Q_4 is negligible because of the forward-biased diode connection. Thus, the current equations at nodes A, B, and C in Fig. 5.56b are as follows:

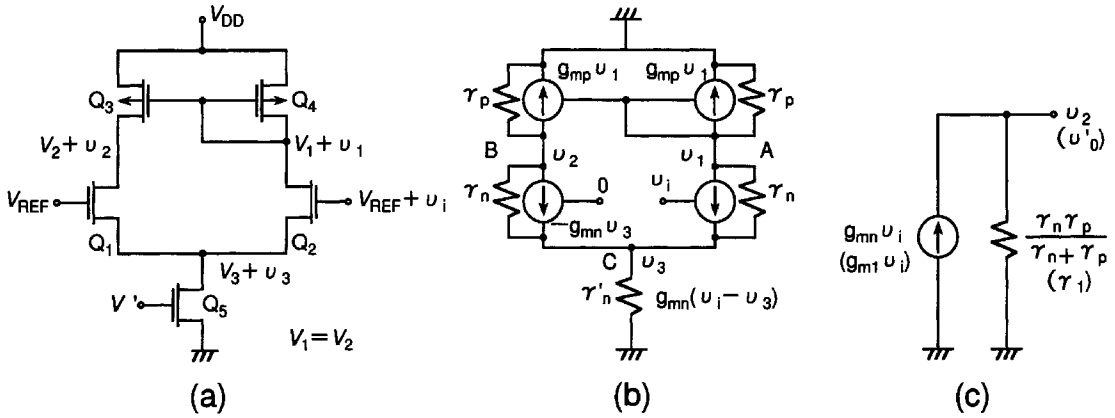


Fig. 5.56. A differential amplifier (**a**), its small-signal equivalent circuit (**b**), and the simplified equivalent circuit (**c**) [5.53]

$$g_{\mathrm{mp}}v_1+\frac{v_1-v_3}{r_{\mathrm{n}}}+g_{\mathrm{mn}}(v_{\mathrm{i}}-v_3)=0,$$

$$g_{\mathrm{mp}}v_1+\frac{v_2}{r_{\mathrm{p}}}-g_{\mathrm{mn}}v_3+\frac{v_2-v_3}{r_{\mathrm{n}}}=0,$$

$$-\frac{v_1-v_3}{r_{\mathrm{n}}}-g_{\mathrm{mn}}(v_{\mathrm{i}}-v_3)-\frac{v_2-v_3}{r_{\mathrm{n}}}+g_{\mathrm{mn}}v_3+\frac{v_3}{r'_{\mathrm{n}}}=0.$$

Assuming that $g_{\mathrm{mn}}r_{\mathrm{n}}\gg 1$, $g_{\mathrm{mp}}r_{\mathrm{p}}\gg 1$, and $g_{\mathrm{mn}}r'_{\mathrm{n}}\gg 1$, v_2 is obtained as follows:

$$v_2=\frac{r_{\mathrm{n}}r_{\mathrm{p}}}{r_{\mathrm{n}}+r_{\mathrm{p}}}g_{\mathrm{mn}}v_{\mathrm{i}}.$$

Thus, the amplifier is eventually simplified to the circuit shown in Fig. 5.56c. Therefore, the whole converter is expressed with the equivalent circuit shown in Fig. 5.55b, where g_{m2} and r_2 are the transconductance and output resistance of the output PMOSFET (Q_6), respectively, C_{g} is the sum of the gate–source capacitance and the gate–drain capacitance of Q_6, R_{L} is the sum of the wiring resistance of the V_{DL} line and the on-resistances of the PMOSFETs in the CMOS core circuit, and C_{L} is the sum of the wiring capacitance of the V_{DL} line and the capacitances seen through PMOSFETs in the core circuit.

The circuit shown in Fig. 5.55b is analyzed with the Laplace transformation as

$$-g_{\mathrm{m1}}v_{\mathrm{i}}+\frac{v'_{\mathrm{o}}}{r_1}+sC_{\mathrm{g}}v'_{\mathrm{o}}=0, \tag{5.38}$$

$$g_{\mathrm{m2}}v'_{\mathrm{o}}+\frac{v_{\mathrm{o}}}{r_2}+\frac{v_{\mathrm{o}}}{R_{\mathrm{L}}+1/sC_{\mathrm{L}}}=0; \tag{5.39}$$

$$\therefore\ \frac{v_{\mathrm{o}}}{v_{\mathrm{i}}}=G(j\omega)=\frac{-g_{\mathrm{m1}}r_1}{sC_{\mathrm{g}}r_1+1}\cdot\frac{g_{\mathrm{m2}}r_2(sC_{\mathrm{L}}R_{\mathrm{L}}+1)}{sC_{\mathrm{L}}(R_{\mathrm{L}}+r_2)+1}.$$

By transforming into the complex-frequency plane ($s = j\omega$), we obtain

$$G(j\omega) = -g_{m1}r_1 g_{m2}r_2 \left(1 + j\frac{\omega}{\omega_z}\right) \Big/ \left[\left(1 + j\frac{\omega}{\omega_{p1}}\right)\left(1 + j\frac{\omega}{\omega_{p1}}\right)\right], \quad (5.40)$$

$$\omega_{p1} = \frac{1}{C_g r_1}, \quad \omega_{p2} = \frac{1}{C_L(R_L + r_2)}, \quad \omega_z = \frac{1}{R_L C_L}, \quad (5.41)$$

which has two poles (ω_{p1}, ω_{p2}) and one zero (ω_z). In DRAM design ω_z is much higher than ω_{p1} and ω_{p2} because R_L (several Ω) $\ll r_1$ and r_2, and $C_L > C_G$, and thus it is located at the frequency where the gain is sufficiently reduced on the ω–gain plane. Therefore, ω_z has no substantial effect on the gain and phase characteristics, making (5.40) equal to (5.37). Moreover, the gain and phase characteristics are similar to those shown in Fig. 5.54, because $\omega_{p1} > \omega_{p2}$, posing a possible oscillation problem caused by a reduced phase margin.

Figure 5.57 shows a simulated V_{DH} waveform. The output (V_{DL}) starts to ring when a large load current of 51 mA is applied, although such a ringing never happens for a small load current of 1 mA. The long decay time after turning off the large current pulse is due to the cutting off of the output PMOSFET. A constant current as small as 1 mA makes the heavily capacitive load, which has been raised to over V_{REF} as a result of ringing, discharge down to V_{REF}.

To stabilize the converter, the phase at a total gain of 0 dB must be lower than 135° (i.e. a phase margin larger than 45°), as explained previously. This is realized by shifting ω_{p1} to a sufficiently higher frequency (see Fig. 5.54) or to a frequency that is sufficiently lower than ω_{p2} for a given ω_{p2}. This is also achieved by shifting ω_{p2} for a fixed ω_{p1}. In any case, the key to stabilization is a sufficient separation of two poles.

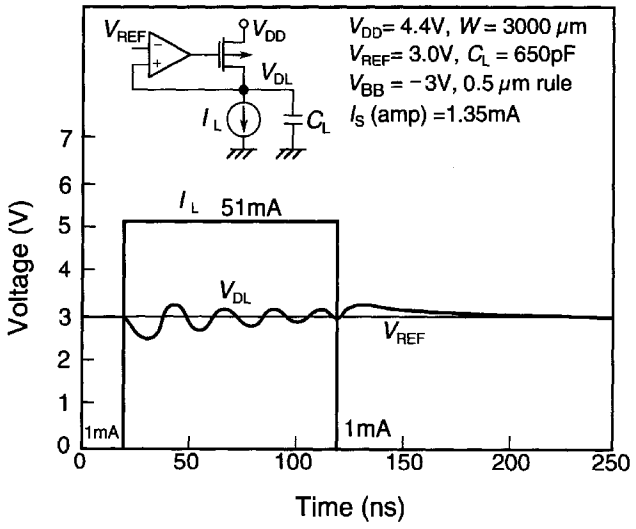


Fig. 5.57. A simulated oscillatory V_{DL} waveform without phase compensation [5.1]

Pole-Zero Compensation. This section describes pole-zero compensation, which is the most promising for DRAM design. In this compensation, the phase compensation devices, R_C and C_C, are added at the output of the circuit shown in Fig. 5.55. Figure 5.58 shows the small-signal equivalent circuit. Here, $C_g = C_{gs}$ – that is, $\frac{2}{3}C_{OX}LW_6$ (when C_{OX} is the gate capacitance per unit area) – because Q_6 is in the saturation state. The current equations at the nodes, expressed using the Laplace transformation, are as follows:

$$-g_{m1}v_i + \frac{v'_o}{r_1} + sC_g v'_o = 0, \tag{5.42}$$

$$g_{m2}v'_o + \frac{v_o}{r_2} + v_o \Big/ \left(R_C + \frac{1}{sC_C}\right) + v_o \Big/ \left(R_L + \frac{1}{sC_L}\right) = 0. \tag{5.43}$$

Hence, the gain function is expressed as

$$\begin{aligned} \frac{v_o}{v_i} &= G(j\omega) \qquad (5.44)\\ &= -G\left(1 + j\frac{\omega}{\omega_{z1}}\right)\left(1 + j\frac{\omega}{\omega_{z2}}\right) \Big/ \left(1 + j\frac{\omega}{\omega_{p1}}\right)\left(1 + j\frac{\omega}{\omega_{p2}}\right)\left(1 + j\frac{\omega}{\omega_{p3}}\right), \end{aligned}$$

where $s = j\omega$, and the poles, zeros, and gain are defined as follows:

$$\omega_{p1} = \frac{1}{C_g r_1}, \tag{5.45}$$

$$\begin{aligned} \omega_{p2} &= \frac{1}{C_L(R_L + r_2) + C_C(R_C + r_2)} \qquad (5.46)\\ &\simeq \frac{1}{r_2(C_C + C_L)} \quad (R_C, R_L \ll r_2) \end{aligned}$$

$$\omega_{p3} = \frac{C_L(R_L + r_2) + C_C(R_C + r_2)}{C_C C_L(R_L R_C + r_2 R_L + r_2 R_C)} \simeq \frac{C_C + C_L}{C_C C_L(R_L + R_C)}, \tag{5.47}$$

$$\omega_{z1} = \frac{1}{R_L C_L}, \tag{5.48}$$

$$\omega_{z2} = \frac{1}{R_C C_C}, \tag{5.49}$$

$$G_0 = g_{m1} r_1 g_{m2} r_2.$$

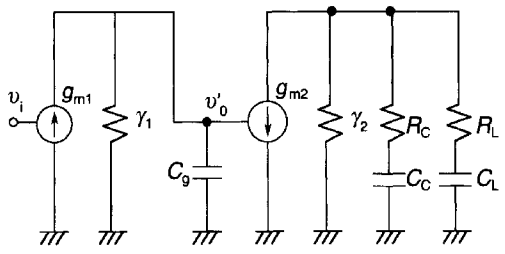


Fig. 5.58. A small-signal equivalent circuit with pole-zero compensation devices R_C and C_C [5.53]

Therefore, in pole-zero compensation, three poles (ω_{p1}, ω_{p2}, ω_{p3}) and two zeros (ω_{z1}, ω_{z2}) are generated. When comparing with (5.40) it is obvious that ω_{p3} and ω_{z2} are newly generated, and ω_{p2} shifts to a lower frequency because of an additional component of $C_C(R_C + r_2)$. Note that ω_{z2} plays an important role, since ω_{p1} can be canceled by ω_{z2} if

$$\omega_{z2} = \omega_{p1}; \tag{5.50}$$

$$\therefore R_C C_C = C_g r_1. \tag{5.51}$$

This implies that a decrease in gain of $-20\,\mathrm{dB/dec.}$ and a phase delay of over $-45°$ at $\omega > \omega_{p1}$ are canceled by an increase in gain of $+20\,\mathrm{dB/dec.}$ and a phase advance of over $+45°$ at $\omega > \omega_{z2}$. Under these conditions, stabilizing conditions for the converter are obtained for two cases of $R_L = 0$ and $R_L > 0$, as follows.

Case of $R_L = 0$. In this case only two poles (ω_{p2}, ω_{p3}) remain in $G(j\omega)$ because $\omega_{z1} = \infty$, and thus

$$G(j\omega) = -G_0 \Big/ \left(1 + j\frac{\omega}{\omega_{p2}}\right)\left(1 + j\frac{\omega}{\omega_{p3}}\right). \tag{5.52}$$

This is same as (5.37) because $\omega_{p3} \gg \omega_{p2}$, if we regard ω_{p3} as ω_{p1} in (5.37). The stabilizing condition, where the gain at $\omega = \omega_{p3}$ is smaller than 0 dB and the phase is $-135°$ (Fig. 5.59), is obtained as

$$|G(j\omega_{p3})| = G_0 \Big/ \sqrt{2}\frac{\omega_{p3}}{\omega_{p2}} \leq 1. \tag{5.53}$$

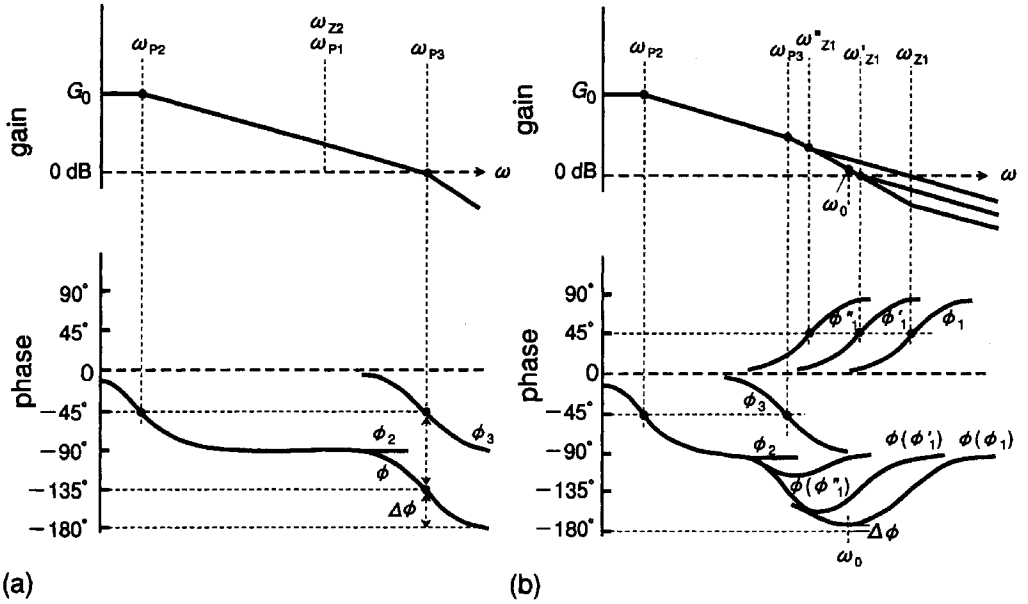


Fig. 5.59. The gain and phase characteristics due to pole-zero compensation for $R_L = 0$ (**a**) and $R_L > 0$ (**b**)

Thus, by using the condition expressed by (5.51),

$$C_{\mathrm{C}} \geq \sqrt{\frac{G_0 C_{\mathrm{L}} C_{\mathrm{g}} r_1}{\sqrt{2}\, r_2}} - C_{\mathrm{L}}, \tag{5.54}$$

$$R_{\mathrm{C}} \leq C_{\mathrm{g}} r_1 \Bigg/ \left(\sqrt{\frac{G_0 C_{\mathrm{L}} C_{\mathrm{g}} r_1}{\sqrt{2}\, r_2}} - C_{\mathrm{L}} \right). \tag{5.55}$$

Fortunately, the phase is equal to $-135°$, since ω_{p2} and ω_{p3} are sufficiently far apart (i.e. $\omega_{\mathrm{p3}} \gg \omega_{\mathrm{p2}}$), as shown in the figure.

Case of $R_{\mathrm{L}} > 0$. The gain is expressed as

$$G(j\omega) = -G_0 \left(1 + j\frac{\omega}{\omega_{\mathrm{z1}}}\right) \Bigg/ \left(1 + j\frac{\omega}{\omega_{\mathrm{p2}}}\right)\left(1 + j\frac{\omega}{\omega_{\mathrm{p3}}}\right). \tag{5.56}$$

When R_{L} is increased from $0\,\Omega$, ω_{z1} decreases as $\omega_{\mathrm{z1}} \rightarrow \omega'_{\mathrm{z1}} \rightarrow \omega''_{\mathrm{z1}}$, as shown in Fig. 5.59b. Due to the increasing R_{L}, the gain is more increased, and the phase is more advanced with the help of ω_{z1}, enabling the phase margin to be increased. When ω_{z1} reaches a certain value of $\omega_{\mathrm{z1}} > \omega_{\mathrm{p3}} \gg \omega_{\mathrm{p2}}$, the phase is expressed as

$$\Phi = -\tan^{-1}\left(\frac{\omega}{\omega_{\mathrm{p2}}}\right) - \tan^{-1}\left(\frac{\omega}{\omega_{\mathrm{p3}}}\right) + \tan^{-1}\left(\frac{\omega}{\omega_{\mathrm{z1}}}\right). \tag{5.57}$$

If we deal with $\omega \gg \omega_{\mathrm{p2}}$, because $\omega_{\mathrm{p2}} \ll \omega_{\mathrm{p3}}, \omega_{\mathrm{z1}}$,

$$\Phi + 90° = -\tan^{-1}\left(\frac{\omega}{\omega_{\mathrm{p3}}}\right) + \tan^{-1}\left(\frac{\omega}{\omega_{\mathrm{z1}}}\right) \tag{5.58}$$

Hence, by using the well-known formula, $\tan(\tan^{-1}a + \tan^{-1}b) = (a+b)/(1-ab)$,

$$\tan(\Phi + 90°) = \left(-\frac{1}{\omega_{\mathrm{p3}}} + \frac{1}{\omega_{\mathrm{z1}}}\right) \Bigg/ \left(\frac{1}{\omega} + \frac{\omega}{\omega_{\mathrm{p3}}\omega_{\mathrm{z1}}}\right). \tag{5.59}$$

The right-hand side of the equation – that is, the phase – is a maximum at

$$\omega = \omega_0 = \sqrt{\omega_{\mathrm{p3}}\omega_{\mathrm{z1}}}. \tag{5.60}$$

Since the stabilizing condition is $\Phi \geq -135°$ at $\omega = \omega_0$, and thus $\tan(\Phi + 90°) \geq -1$, (5.59) is expressed as

$$4\omega_{\mathrm{p3}}\omega_{\mathrm{z1}} \geq (\omega_{\mathrm{z1}} - \omega_{\mathrm{p3}})^2. \tag{5.61}$$

Thus, by using (5.51), the limitations for ω_{z1}, C_{C}, and R_{C} are obtained as follows:

$$\frac{\omega_{z1}}{\omega_{p3}} \leq 5.8; \tag{5.62}$$

$$\therefore C_C \geq \frac{1}{4.8}\left(\frac{r_1 C_g}{R_L} - 5.8C_L\right); \tag{5.63}$$

$$\therefore R_C \leq \frac{4.8 r_1 C_g}{(r_1 C_g / R_L) - 5.8C_L}. \tag{5.64}$$

On the other hand, the gain at $\omega = \omega_0$ is derived as follows:

$$|G(j\omega_0)| = G\sqrt{1+\left(\frac{\omega_0}{\omega_{z1}}\right)^2} \Big/ \left[\sqrt{1+\left(\frac{\omega_0}{\omega_{p2}}\right)^2}\sqrt{1+\left(\frac{\omega_0}{\omega_{p3}}\right)^2}\right] \tag{5.65}$$

$$\simeq G_0\left(\frac{\omega_{p2}}{\omega_{z1}}\right).$$

Note that as long as there are zeros and the phase is less than $-135°$ (i.e. the phase margin $\Delta\Phi \geq 45°$), even a gain of larger than 1 realizes a stable converter. The condition $|G(j\omega)| < 1$ would result in an over-margin, as follows.

In the case shown in Fig. 5.59a, where the converter is substantially characterized only by the poles, the phase increases from $-135°$ to $-180°$ when ω is increased in the frequency region beyond ω_{p3}. The phase increase also increases the degree of instability that just starts to occur due to positive feedback at $-180°$. Thus, to ensure stability, the gain must be strictly limited to less than one. In the case shown in Fig. 5.59b, however, where there is the zero that is influential to performance, a phase margin ($\Delta\Phi$) larger than 45° is ensured at any ω as long as the minimum $\Delta\Phi$ is larger than 45° at $\omega = \omega_0$. Fortunately, the limitation on the gain is not necessarily needed, because the gain at $\omega = \omega_0$, which is $G_0(\omega_{p2}/\omega_{z1})$, is low enough. Therefore, (5.63) and (5.64) are stable conditions.

The converter must be stable against wide changes of load parameters (R_L, C_L, I_L) caused by internal DRAM-core operation. Figure 5.60 shows the relationship between the parameter changes and the phase margin for pole-zero compensation [5.53]. Here, the phase margin is defined as the margin at $\omega_c (= \omega_{c1})$ where the gain is one (i.e. 0 dB) when the margin is fixed to 45° at $\omega = \omega_0$, as shown by $\Delta\Phi'_1$ in Fig. 5.60b. $\Delta\Phi'$ is derived from analyzing (5.56), assumimg $C_C = 1.92\,\text{nF}$ and $R_C = 23.4\,\Omega$, which are obtained by substituting normal values of $R_{L0} = 3\,\Omega$, $C_{L0} = 1\,\text{nF}$, $I_{L0} = 10\,\text{mA}$, $C_g = 7.2\,\text{pF}$, $r_1 = 6.25\,\text{k}\Omega$, $\lambda_P = 0.1\,/\text{V}$, and $\lambda_n = 0.04\,/\text{V}$ in (5.63) and (5.64). The parameters are all for a 0.5 μm 16 Mb DRAM. $\Delta\Phi'$ increases with increasing R_L or C_L because ω_{z1} approaches ω_{p3}, as shown in Fig. 5.59b. However, it is almost a minimum at the normal value of I_{L0} when I_L is increased. This can be explained using Fig. 5.60b: ω_{p2} increases with I_L because $r_2 = (\lambda_p I_L)^{-1}$, which makes the gain curve shifted to the upper with a fixed location of ω_0 that is independent of I_L (i.e. a thick dashed line in the figure). Thus, $\Delta\Phi'$ (i.e. $\Delta\Phi'_2$) at ω_c (i.e. ω_{c2}) on the shifted gain curve increases. On the contrary,

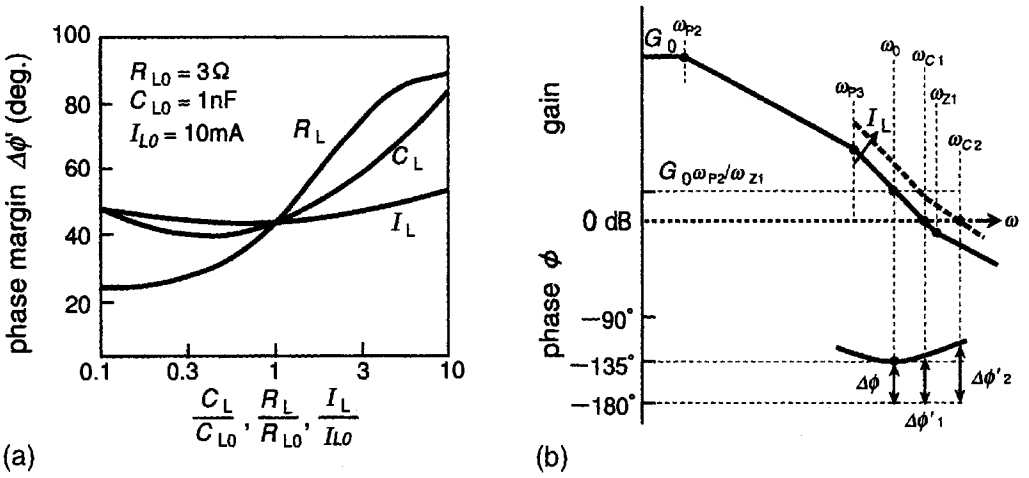


Fig. 5.60. A sensitivity analysis for pole-zero compensation [5.53]. $r_1 = 6.2\,\mathrm{k\Omega}$, $r_2 = (\lambda_P I_L)^{-1}$, $\lambda_P = 0.1/\mathrm{V}$, $\lambda_n = 0.04/\mathrm{V}$, $g_{m1} = 4.9\,\mathrm{mS}$, $g_{m2} = [2I_L(W_6/L)\beta_{P0}]^{-1}$, $\beta_{P0} = 3 \times 10^{-5}\,\mathrm{S/V}$, $W_6 = 3912\,\mu\mathrm{m}$, $L = 1.2\,\mu\mathrm{m}$, $C_{OX} = 2.3\,\mathrm{fF}/\mu\mathrm{m}^2$, $I_S = 2.3\,\mathrm{mA}$, $C_g = C_{gs} = 7.2\,\mathrm{pF}$. (**a**) Phase margin versus load parameters; (**b**) gain and phase characteristics

when I_L is decreased, $\Delta\Phi'$ continues to decrease until it reaches 45°, and then it increases. In any case, the phase margin is not sensitive to variations in I_L. There is an important implication here, that the converter is stable against drastic changes in the load current. Careful attention should be paid to the reduction of R_L because $\Delta\Phi'$ decreases with the reduction. In this case, C_C and R_C must again be obtained by substituting the minimum R_L for (5.63) and (5.64). Thus the converter is stable because a phase margin larger than 45° is always ensured for R_L larger than the minimum R_L, which corresponds to $R_L/R_{L0} = 1$ in Fig. 5.60.

Comparisons between Various Compensations. In addition to pole-zero compensation, two other kinds of compensation (Miller compensation and dominant-pole compensations) have been proposed for application to DRAMs, as shown in Table 5.1 [5.53, 5.55]. The dotted curves in the figure represent the gain and phase curves before compensation, while the solid curves are after compensation. Our major concerns are not only the stability, but also the sizes of the compensation devices, because the devices must be incorporated in a chip.

Miller Compensation. This adds a compensation capacitance (C_C) at the amplifier output. The necessary C_C (usually, $C_C \gg C_G$) can be reduced by utilizing the Miller effect of the PMOS output stage. In DRAM design both R_L and $1/g_{m2}$ are small, and thus ω_{z1} and ω_{z2} are negligible, because they are located at a ω that is sufficiently higher than ω_{p1} and ω_{p2}, although two poles and two zeros are generated in the gain function. Note that the pole without compensation, ω'_{p1}, shifts drastically to a lower frequency, ω_{p1}, due to the Miller effect. This is because the gate capacitance C_g is increased

Table 5.1. A comparison between various phase compensation [5.53, 5.55]

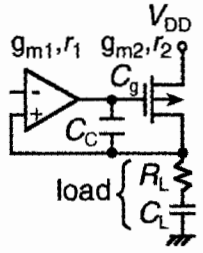
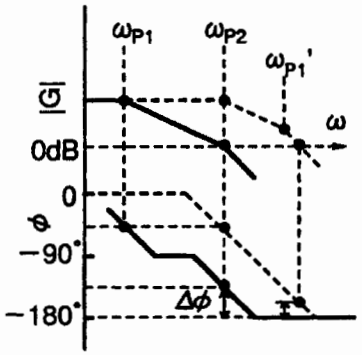
Circuit	Gain/phase characteristics	Gain function	Stability condition
		$G(j\omega) = \frac{-G_0\left(1+j\frac{\omega}{\omega_{Z1}}\right)\left(1-j\frac{\omega}{\omega_{Z2}}\right)}{\left(1+j\frac{\omega}{\omega_{P1}}\right)\left(1+j\frac{\omega}{\omega_{P2}}\right)}$	(a) $C_g \ll C_C C_L$ $C_C \geqslant \frac{C_L g_{m1}}{\sqrt{2} g_{m2}}$
		$\omega_{P1} \fallingdotseq \frac{1}{C_C g_{m2} r_1 r_2}$	(b) $C_g \fallingdotseq C_C C_L$
		$\omega_{P2} \fallingdotseq \frac{g_{m2}}{C_L}$	$C_C \geqslant \frac{1}{2}\sqrt{\frac{A^2}{2}+B}+\frac{\sqrt{2}}{4}A$
		$\omega_{Z1} = \frac{1}{R_L C_L}$	$A = \frac{(C_L + C_g) g_{m1}}{g_{m2}}$
		$\omega_{Z2} \fallingdotseq \frac{g_{m2}}{C_C}$	$B = \frac{2\sqrt{2} C_L C_g g_{m1}}{g_{m2}}$
Miller compensation	$\omega_{Z1}, \omega_{Z2} \gg \omega_{P1}, \omega_{P2}$	$G_0 = g_{m1} r_1 g_{m2} r_2$	

Table 5.1. Continued

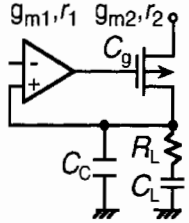
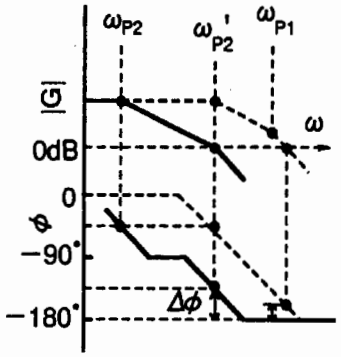
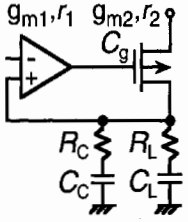
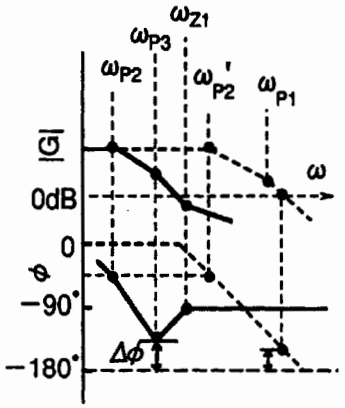
Circuit	Gain/phase characteristics	Gain function	Stability condition
		$G(j\omega) = \dfrac{-G_0\left(1+j\frac{\omega}{\omega_{Z1}}\right)}{\left(1+j\frac{\omega}{\omega_{P1}}\right)\left(1+j\frac{\omega}{\omega_{P2}}\right)\left(1+j\frac{\omega}{\omega_{P3}}\right)}$	(a) $R_L = 0$
		$\omega_{P1} = \dfrac{1}{C_g r_1}$	$C_c \geqslant \dfrac{G_0 C_g r_1}{\sqrt{2} r_2} - C_L$
		$\omega_{P2} \fallingdotseq \dfrac{1}{r_2(C_C + C_L)}$	(b) $R_L \geqslant \dfrac{C_g r_1}{5.8 C_L}$
		$\omega_{P3} \fallingdotseq \dfrac{C_C + C_L}{C_C C_L R_L}$	$C_C = 0$
Dominant-pole compensation	$\omega_{Z1}, \omega_{P3}, \gg \omega_{P1}, \omega_{P2}$	$\omega_{Z1} \fallingdotseq \dfrac{1}{R_L C_L}$	
		$G_0 = g_{m1} r_1 g_{m2} r_2$	

Table 5.1. Continued

Circuit	Gain/phase characteristics	Gain function	Stability condition
		$G(j\omega) = \dfrac{-G_0\left(1+j\frac{\omega}{\omega_{Z1}}\right)\left(1+j\frac{\omega}{\omega_{Z2}}\right)}{\left(1+j\frac{\omega}{\omega_{P1}}\right)\left(1+j\frac{\omega}{\omega_{P2}}\right)\left(1+j\frac{\omega}{\omega_{P3}}\right)}$	(a) $R_L \doteqdot 0, R_C C_C = C_g r_1$ $C_C \geqslant \sqrt{\dfrac{G_0 C_L C_g r_1}{\sqrt{2} r_2}} - C_L$
		$\omega_{P1} = \dfrac{1}{C_g r_1}$	(b) $R_L > 0, R_C C_C = C_g r_1$ $C_C \geqslant \dfrac{1}{4.8}\left(\dfrac{r_1 C_g}{R_L} - 5.8 C_L\right)$
		$\omega_{P2} \doteqdot \dfrac{1}{r_2(C_C + C_L)}$	
		$\omega_{P3} \doteqdot \dfrac{C_C + C_L}{C_C C_L (R_L + R_C)}$	
Pole-zero compensation	$\omega_{P1} = \omega_{Z2}$	$\omega_{Z1} = \dfrac{1}{R_L C_L}, \omega_{Z2} = \dfrac{1}{R_C C_C}$	
		$G_0 = g_{m1} r_1 g_{m2} r_2$	

to $g_{m2}r_2C_C$, as seen when comparing ω_{p1} in the table with that in (5.41). Therefore, the phase margin is increased. A small C_C of 50 pF is enough for stable operation of a 0.5 μm 16 Mb DRAM with $C_L = 500\,\text{pF}$, $0 \leq R_L \leq 10\,\Omega$, and $I_L = 100\,\text{mA}$ [5.55]. It has been reported [5.56] that a V_{DL} oscillation of a 4 Mb BiCMOS SRAM using an on-chip voltage down-converter with $W_6 = 3000\,\mu\text{m}$ and a C_L of thousands of pF was stopped by adding a C_C as small as 30 pF. The drawback, however, is a poor power-supply rejection ratio (PSRR) [5.55]. If the power supply (V_{DD}) has a high-frequency noise component, as in the case of a V_{DD} bump with rise and fall times of less than 10 ns, a noise couples to the V_{DL} output node through C_{gs} and C_C.

Dominant-Pole Compensation. This adds a compensation capacitance (C_C) at the output, which generates three poles (ω_{p1}, ω_{p2}, and ω_{p3}) and one zero (ω_{z1}). A small R_L and a large C_L ($\gg C_g$) allow $\omega_{p3}, \omega_{z1} \gg \omega_{p1}, \omega_{p2}$, and $\omega_{p1} \gg \omega_{p2}$. Therefore, ω_{p2} shifts from a high frequency (ω'_{p2}) of $1/[C_L(R_L + r_2)]$, expressed in (5.41), to a low frequency (ω_{p2}) of $1/[r_2(C_C+C_L)]$, enabling a wider phase margin. However, the compensation requires a quite large C_C, which can be expressed as

$$C_C \geq \frac{G_0 C_g r_1}{\sqrt{2}\, r_2} - C_L, \qquad r_2 = \frac{1}{\lambda_P I_L}. \tag{5.66}$$

The large gain (about 2000) is responsible for the large C_C. For example, typical parameters for 0.5 μm process technology result in $C_C \geq 100\,\text{nF}$ [5.53]. If C_C is fabricated with a thin (15 nm) oxide sandwiched between the n-well and the poly-Si, it consumes $50\,\text{mm}^2$, which is almost half of the area of a 16 Mb chip. Dominant-pole compensation is therefore not suitable for an on-chip voltage converter.

Pole-Zero Compensation. In this type of compensation, the insertion of ω_{z1} and ω_{z2} extends the phase margin, as discussed previously. As for the necessary C_C, it is rather smaller than that for dominant-pole compensation, as seen in (5.54) for $R_L = 0$, where it is proportional to $\sqrt{G_0}$, instead of G_0 for dominant-pole compensation. When R_L is increased, C_C is further reduced, as seen in (5.63). For example, a R_L of $3\,\Omega$ needs a 1.9 nF C_C, which is one-fourth that for $R_L = 0$, consuming an area of $1\,\text{mm}^2$, which is less than 1% of the area of a 16 Mb. Although a compensation resistance (R_C) of about $10\,\Omega$ is needed, a small resistor is realized by adjusting the W/L value of a poly-Si resistor with a sheet resistivity of about $50\,\Omega/\square$. It has been reported that a 16 Mb DRAM, whose parameters are $R_L = 5\,\Omega$, $C_L = 650\,\text{pF}$, and $I_L = 100\,\text{mA}$, has realized a sufficient phase margin of 53° through a pole-zero compensation of $C_C = 650\,\text{pF}$ and $R_C = 7\,\Omega$. Moreover, a noise as small as 50 mV appeared at the V_{DL} output node even when a peak-to-peak noise as large as 1.5 V was superposed on V_{DD}. Thus, pole-zero compensation is the best option for application to DRAMs [5.53].

5.4.5 Reference-Voltage Generators

In this section, the reference-voltage (V_{REF}) generators that govern the characteristics of the voltage down-converter are discussed. The V_{REF} generators for application to DRAMs are categorized as the V_T referenced V_{REF} generator, the V_T difference (ΔV_T) V_{REF} generator, and the band-gap V_{REF} (BGR) generator. A low V_{REF} is for the V_{DL} in Fig. 5.48a, while a high V_{REF} is for that in Fig. 5.47.

The V_T-Referenced V_{REF} Generator. Figure 5.61a shows a V_{REF} generator that consists of series-connected MOS diodes for obtaining quite a high V_{REF}, such as 4 V [5.44]. If there is no substrate-bias effect, V_{REF} is equal to the product of the number of MOSFETs and V_T. Although the generator is simple, it suffers from drawbacks such as V_{REF} setting inaccuracy: the variation in V_{REF} depends strongly on that in V_T, caused by variations in the fabrication process. The strong dependence on temperature is also serious. If V_{REF} is generated by a series connection of six MOSFETs, the variation in

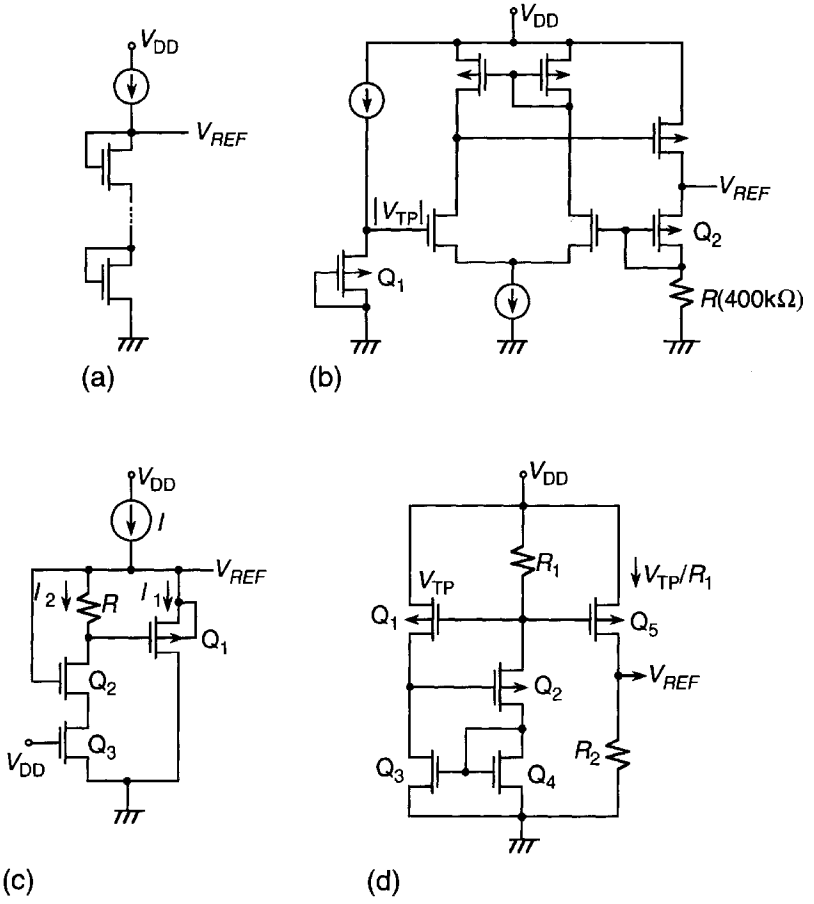


Fig. 5.61. V_T-referenced V_{REF} generators using series-connected MOS diodes [5.44] (**a**), a differential amplifier [5.49] (**b**), automatic adjustment of gate–source biasing [5.59] (**c**), and a current mirror [5.40] (**d**)

V_{REF} is six times larger than the variation in V_T. For example, a temperature variation of 100 °C causes a variation in V_{REF} as large as 1.2 V for a V_T temperature coefficient of about −2 mV/°C [5.57].

Figure 5.61b shows another V_T-referenced generator [5.49] to obtain a low V_{REF}. The circuit configuration is similar to that of the V_{DL} circuit shown in Fig. 5.48, in which V_{REF} corresponds to the threshold voltage ($|V_{TP}|$) of PMOS (Q_1), while R_1 corresponds to PMOS (Q_2). The temperature dependence of V_{REF} can be canceled if poly-Si is used for R. The cancellation is explained by the following equation:

$$V_{REF} = |V_{TP}| \left(1 + \frac{1}{Rg_m}\right). \tag{5.67}$$

Here, the poly-Si resistance features no temperature dependence [5.1] if the sheet resistivity is around 50 Ω/□, while the Q_2 transconductance (g_m) and $|V_{TP}|$ are both lowered with increasing temperature. Thus, the temperature dependence is canceled. The whole voltage down-converter implementing the generator was designed for a 16 Mb DRAM [5.49]. The resulting performance is as follows: $\Delta V_{DL}/\Delta_T = 0.9\,\text{mV/°C}$, $\Delta V_{DL}/\Delta R = 18.6\,\text{mV/Ω/□}$, $\Delta V_{DL}/\Delta V_{TP} = 1.5$, and $\Delta V_{DL}/\Delta V_{DD} = 0.07$ for $V_{DD} = 5$ V and $V_{DL} = 4$ V. The stand-by current of the converter is 35 μA, and the layout area of 400 kΩ with 50 Ω/□ sheet resistivity is less than 0.1% of 16 Mb chip area of 128 mm^2. The drawback of the generator is a large variation of V_{REF} caused by variations in $|V_{TP}|$ and R.

Figure 5.61c is a generator [5.59] that is suitable for obtaining a low V_{REF} through automatically adjusting the gate–source bias voltage of PMOS (Q_1). If the parameters of R, Q_1, and Q_3 are set so that the voltage drop (I_2R) is larger than the Q_1 threshold voltage ($|V_{TP}|$), the constant current (I) separates into I_1 and I_2, as a result of Q_1 being turned on. Any change in V_{REF} is eventually suppressed by the negative feedback of the circuit. For example, a positive change in V_{REF} is detected by Q_2, enabling the Q_2 gate voltage to be reduced and thus I_1 to be increased. I_2 is increased instead, and the resulting reduced drop in I_2R prevents Q_2 from delivering more current. The generator can reduce the temperature dependence of V_{REF} by utilizing the canceling effect of temperature coefficients of $|V_{TP}|$ and an equivalent resistance (R_{eq}) formed by Q_2 and Q_3, as explained previously. This is because V_{REF} is expressed as

$$V_{REF} = |V_{TP}| \left(1 + \frac{R_{eq}}{R}\right). \tag{5.68}$$

The V_{REF} generator provides excellent performance, producing of an almost constant V_{REF} of 1.3 V over a wide range of V_{DD} of 2–6 V, $\Delta V_{REF}/\Delta T = 0.2\,\text{mV/°C}$, and 2 μA at $V_{DD} = 2.5$ V. The drawback, however, is large V_{REF} variations caused by variations in $|V_{TP}|$ and R_{eq}/R.

Figure 5.61d shows a current-mirror V_{REF} generator [5.40, 5.54]. A current mirror consisting of Q_3 and Q_4 allows the same bias current (I) to flow to

Q_1 and R_1. If the W/L value of Q_1 is large enough and I is small enough, the drop IR_1 is given by

$$IR_1 = V_{GS}(Q_1) = V_{TP} + \sqrt{\frac{2I}{\beta_1}} \simeq V_{TP};$$

$$\therefore I = V_{TP}/R_1.$$

Since the same current (I) flows to Q_5, which is of the same size as Q_1, V_{REF} is expressed as

$$V_{REF} = \frac{R_2}{R_1} V_{TP}. \tag{5.69}$$

Variations in V_{REF} due to variations in resistivity are suppressed by using the same width of poly-Si for R_1 and R_2. Variation due to variation in V_T is also suppressed by means of the laser-trimming of both resistors. In addition, if the data-line polycide and the doped poly-Si of the cell-capacitor plate are utilized as the materials for R_1 and R_2, respectively, V_{REF} is less sensitive to temperature variations because the temperature dependence of V_{TP} is canceled by that of R_2/R_1. In fact, a generator for a 16 Mb DRAM incorporating such a trimming technique and the above materials has yielded excellent experimental results [5.40]: $V_{REF} = 1.935\,\text{V}$ at $V_{DD} = 5\,\text{V}$ and 25 °C, $\Delta V_{REF}/\Delta T = +0.15\,\text{mV/°C}$, and $\Delta V_{REF}/\Delta V_{DD} = +10\,\text{mV/V}$.

In summary, the V_T-referenced V_{REF} generator suffers from variations in V_{REF} caused by variations in V_T, although additional process steps are not needed.

The V_T-Difference (ΔV_T) V_{REF} Generator. Figure 5.62a shows a generator [5.47] that utilizes the difference in V_T (ΔV_T) between two MOSFETs. V_{REF} is less sensitive to variations in V_{DD} and temperature because it is determined only by ΔV_T. In addition, V_{REF} is immune to voltage fluctuations of the p-substrate and V_{DD}, since the PMOSFETs in the n-well are used, and their sources are not directly connected to V_{DD}. In the generator, V_{REF} is equal to ΔV_T ($= |V_{TP1}| - |V_{TP2}|$) if two PMOSFETs (Q_1, Q_2) are the same in size and the same current (I) flows. The temperature dependence is expressed as follows:

$$\frac{dV_{REF}}{dT} = \frac{d}{dT}(|V_{TP1}| - |V_{TP2}|) = \frac{k}{q} \ln \frac{N_1}{N_2}, \tag{5.70}$$

where N_1 and N_2 are the impurity concentrations of Q_1 and Q_2. To reduce the temperature dependence a smaller N_1/N_2 (i.e. a smaller ΔV_T) is preferable, which inevitably causes a small V_{REF}. For example, V_{REF} is as small as 1.1 V for $N_1/N_2 = 100$. Although the temperature dependence of V_{TP} depends on the current density [5.60], the difference between Q_1 and Q_2 (i.e. the temperature dependence of V_{REF}) is a constant as small as $0.4\,\text{mV/°C}$, independent of the current density, as shown in the figure.

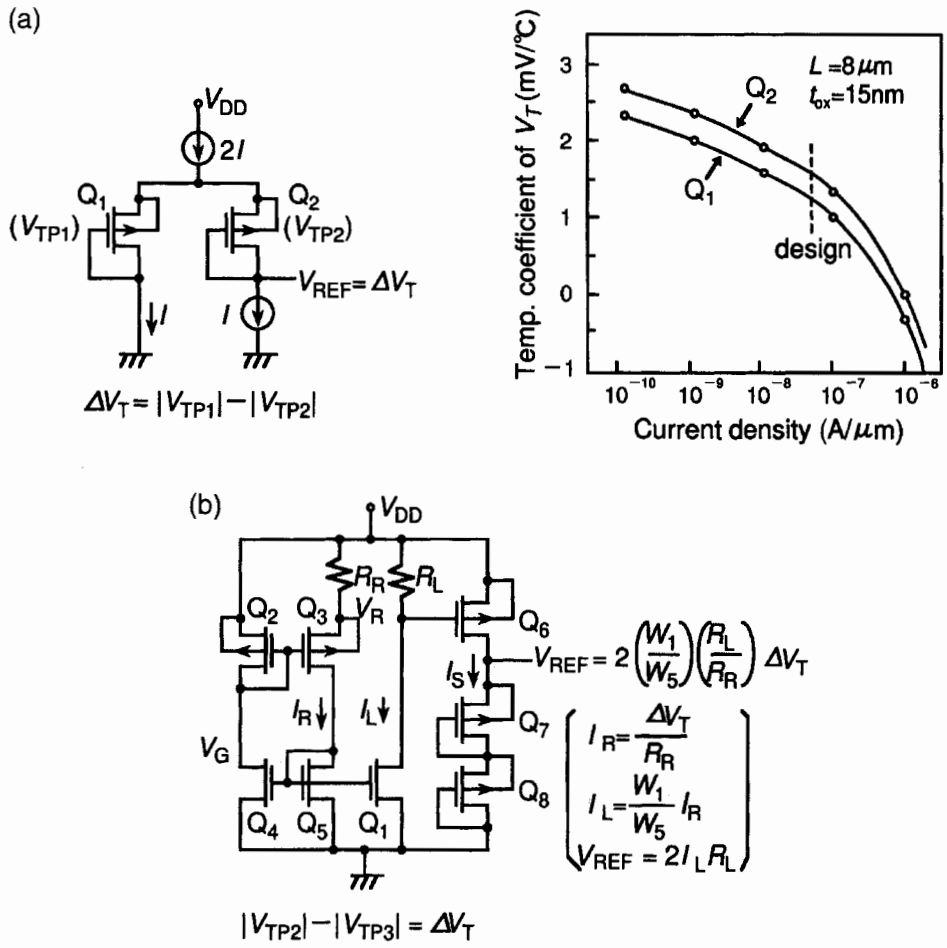


Fig. 5.62. V_T-difference (ΔV_T) V_{REF} generators using constant-current sources [5.47] (**a**) and a current mirror [5.61] (**b**)

Figure 5.62b shows a current-mirror V_{REF} generator [5.61] that utilizes a V_T difference (ΔV_T). It features a direct conversion from a small ΔV_T to a large V_{REF} without adding the first-stage converter shown in Figure 5.48b. The MOSFETs (Q_2–Q_5) generate the constant current (I_R), and Q_1 converts I_R to I_L. Then, Q_6–Q_8 convert I_L to V_{REF}, which is a constant voltage determined only by the size and resistance ratios (W_1/W_5 and R_L/R_R) and ΔV_T. The details are given in the following. If Q_4 and Q_5 have the same size and V_T value, a current mirror comprised of Q_4 and Q_5 allows the current (I_R) flowing from Q_3 to Q_5 to be the same as that flowing from Q_2 to Q_4. Since the saturation current of Q_2 is the same as that of Q_3,

$$\frac{\beta_2}{2}(V_{DD}-V_G-|V_{TP2}|)^2=\frac{\beta_3}{2}(V_R-V_G-|V_{TP3}|)^2. \tag{5.71}$$

Hence, by using $|V_{TP2}| - |V_{TP3}| = \Delta V_T$ and $\beta_2 = \beta_3$,

$$V_R = V_{DD} - \Delta V_T;$$
$$\therefore I_R = \frac{V_{DD} - V_R}{R_R} = \frac{\Delta V_T}{R_R}. \tag{5.72}$$

Another current mirror composed of Q_1 and Q_5 converts I_R to I_L ($= I_R \cdot W_1/W_5$, $W_1 > W_5$). Consequently, a voltage across R, $I_L R_L$, is applied to the gate–source of Q_6, and a saturation current (I_S) flows from Q_6, and is expressed as

$$I_S = \frac{\beta_6}{2}(I_L R_L - |V_{TP6}|)^2. \tag{5.73}$$

Since Q_7 and Q_8 are in the saturation condition and their currents are the same,

$$I_S = \frac{\beta_7}{2}(V_{GS7} - |V_{TP7}|)^2 = \frac{\beta_8}{2}(V_{GS8} - |V_{TP8}|)^2. \tag{5.74}$$

If Q_6–Q_8 are the same in size, and the n-well of each MOSFET, which is separated from the others, is connected to the source to eliminate the substrate-bias effect and thus to obtain the same value of V_{TP}, (5.73) and (5.74) give the following equation:

$$V_{GS7} = V_{GS8} = I_L R_L. \tag{5.75}$$

Thus, the same gate–source voltage of $I_L R_L$ is supplied to Q_6–Q_8. Hence, V_{REF} is obtained as

$$V_{REF} = 2I_L R_L = 2\left(\frac{W_1}{W_5}\right)\left(\frac{R_L}{R_R}\right)\Delta V_T. \tag{5.76}$$

If R_L and R_R are designed with the same width of poly-Si, the ratio R_L/R_R is almost independent of V_{DD}, temperature, and fabrication-process variations. W_1/W_5 and ΔV_T can also be constant against parameter variations. Thus, a stabilized V_{REF} is obtained, exemplified by $\Delta V_{REF}/\Delta T = 0.89\,\mathrm{mV/°C}$ for $V_{DD} = 3.3\,\mathrm{V}$, $V_{REF} = 2.4\,\mathrm{V}$, $\Delta V_T = 0.3\,\mathrm{V}$, $W_1/W_5 = 1$, $R_R = 18\,\mathrm{k\Omega}$ and $R_L = 72\,\mathrm{k\Omega}$.

In summary, a quite high V_{REF} is available from a small ΔV_T without a special conversion circuit. However, V_{REF}-trimming might be still required, although the variations in V_{REF} are much smaller than in the V_T-referenced V_{REF} generator. An additional process step to produce a V_T-difference is also needed.

The Band-Gap V_{REF} Generator. Figure 5.63a shows the well-known band-gap V_{REF} (BGR) generator [5.45] proposed for application to DRAMs, favoring the generation of a low V_{REF} of 1.2–1.6 V. If Q_1 and Q_2 have the same emitter size, the Q_2-emitter voltage is given as

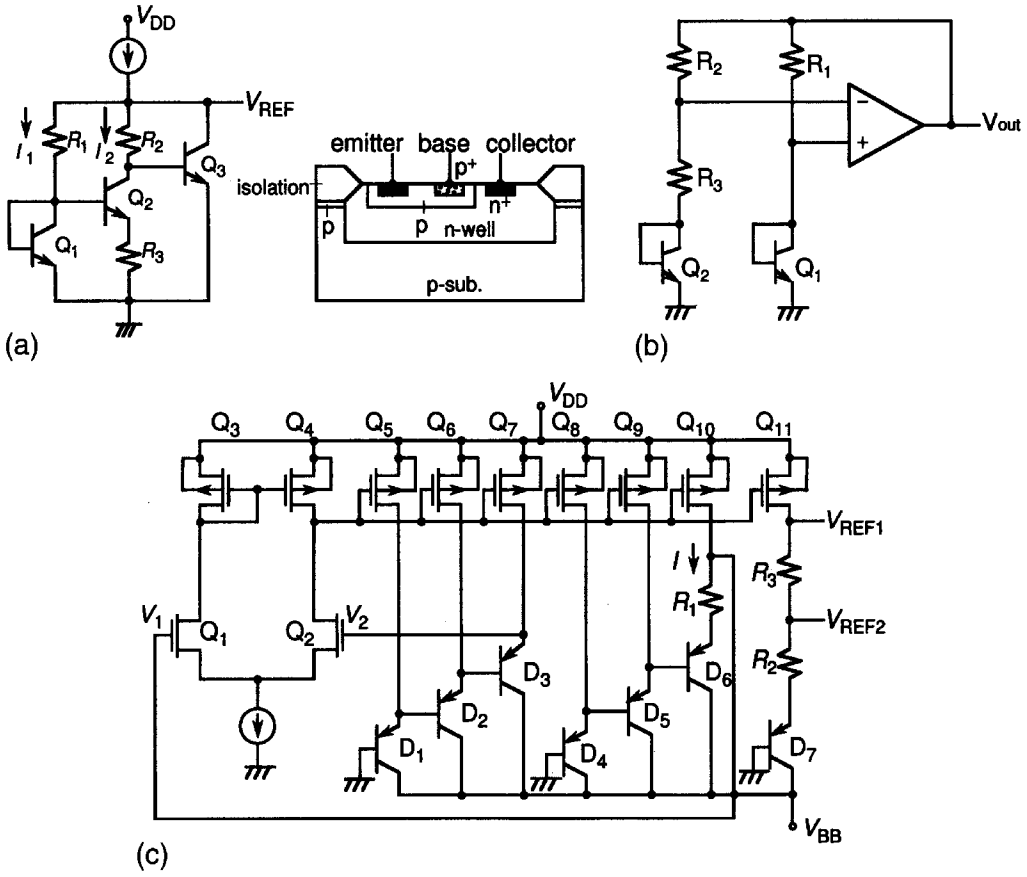


Fig. 5.63. Band-gap V_{REF} generators using npn transistors [5.45] (**a**), a comparator [5.54] (**b**), and pnp transistors [5.41] (**c**)

$$V_{\mathrm{BE}}(Q_1) - V_{\mathrm{BE}}(Q_2) \simeq \frac{kT}{q} \ln \frac{I_1}{I_2};$$

$$\therefore V_{\mathrm{REF}} \simeq V_{\mathrm{BE}}(Q_3) + \frac{R_2}{R_3} \frac{kT}{q} \ln \frac{I_1}{I_2}. \tag{5.77}$$

Because of the negative temperature coefficient ($-1.6\,\mathrm{mV/°C}$) of V_{BE}, the temperature coefficient of V_{REF} is zero if R_2/R_3 is appropriately chosen for $I_1 > I_2$. It has been reported [5.57] that the V_{DL} of a 1 Mb BiCMOS DRAM incorporating a BGR generator using high-performance bipolar transistors has revealed a temperature coefficient of $-0.39\,\mathrm{mV/°C}$. Another BGR generator using cheap but low-performance bipolar transistors, constructed in a CMOS structure, has been also proposed for the voltage down-converter of a 16 Mb DRAM [5.45]. Bipolar transistors are fabricated as follows. After forming the n-wells, p-regions for the bases of the bipolar transistors and field-isolation are formed with a high-energy implantation, followed by simultaneous formation of n^+-regions for emitters and collectors, and for the drains and sources of the NMOSFETs. Then, the base connections (p^+), and the drains and sources of the PMOSFETs are concurrently fabricated. The

resultant bipolar transistor has a current gain (h_{fe}) of 60, and the temperature coefficient of V_{DL} is $+1.6\,mV/°C$ for $R_2/R_3 = 17$ and the same size of transistors (Q_1–Q_3). The drawback is a non-negligible V_{DD}-dependence of V_{REF} due to a high collector resistance [5.47, 5.49].

Figure 5.63b shows another BGR generator [5.54]. Since the differential input voltage of the amplifier (op amp) must be zero, and resistors R_1 and R_2 have equal voltages across them, the output voltage V_{out} is

$$V_{out} = V_{BE}(1) + \frac{R_2}{R_3}\frac{kT}{q}\ln\frac{R_2}{R_1}\frac{I_{S2}}{I_{S1}}, \tag{5.78}$$

where I_{S1} and I_{S2} are the saturation currents of transistors Q_1 and Q_2, which are proportional to their emitter–base junction areas and temperature. V_{out} becomes independent of temperature at $V_{out} = 1.26\,V$.

Figure 5.63c shows a variation [5.41] of the basic BGR shown in Fig. 5.63b. The input voltage (V_1) of the differential amplifier consisting of Q_1–Q_4 is the sum of $3V_{BE1}$ and IR_1. Here, V_{BE1} is the base–emitter voltage, V_{BE}, of the pnp-transistors (D_4–D_6) with the same emitter area. The other input voltage (V_2) is $3V_{BE2}$ (V_{BE2} is the V_{BE} of D_1–D_3 with the same emitter area). If the emitter area of D_4–D_6 is chosen to be larger than that of D_1–D_3, V_{BE1} is lower than V_{BE2} because the same current flows to each pnp transistor, allowing the emitter voltage of D_6 to be lower than that of D_3. Thus, the current (I) is increased by decreasing the output voltage (i.e. the Q_{10} gate voltage) of the amplifier, so that V_1 becomes equal to V_2. In this circuit V_{REF1} and V_{REF2} are given as follows. If Q_5–Q_{11} are the same in size and same saturation current (I) flows,

$$V_1 = 3V_{BE1} + IR_1 = \frac{3kT}{q}\ln\frac{I}{I_{S1}} + IR_1,$$

$$V_2 = 3V_{BE2} = \frac{3kT}{q}\ln\frac{I}{I_{S2}}.$$

By using the condition that $V_1 = V_2$,

$$IR_1 = \frac{3kT}{q}\ln\frac{I_{S1}}{I_{S2}};$$

$$\therefore V_{REF2} = V_{BE}(D_7) + IR_2 = V_{BE}(D_7) + \frac{R_2}{R_1}\frac{3kT}{q}\ln\frac{I_{S1}}{I_{S2}}; \tag{5.79}$$

$$\therefore V_{REF1} = V_{BEF2} + IR_3 = V_{BE}(D_7) + \frac{R_2 + R_3}{R_1}\frac{3kT}{q}\ln\frac{I_{S1}}{I_{S2}}. \tag{5.80}$$

Therefore, the temperature coefficients of V_{REF1} and V_{REF2} can be changed by adjusting the ratios (R_2/R_1 and R_3/R_1). The pnp transistors can be fabricated by utilizing a CMOS device structure in which the p-substrate, n-well, and p-diffused layer are used for the collector, base, and emitter, respectively. A negative V_{BB} can be produced by an on-chip V_{BB} generator if the current flowing to V_{BB} is small.

In summary, the band-gap V_{REF} generator is the best option with the smallest variation in V_{REF}, providing a wide variety of circuit configurations.

5.4.6 Burn-In Test Circuits

The Burn-In Test Condition. The burn-in test is a kind of reliability test of LSIs to quickly get rid of potential defects. As a result of acceleration through applying a high stress voltage and high temperature, potential defects show up quickly. The test must also be applied to DRAM chips that incorporate an on-chip voltage down-converter, although special care should be taken so that the converter does not prevent the application of a stress voltage to the internal circuit during the burn-in test. Figure 5.64 shows the concept of the burn-in test for DRAMs using an on-chip voltage down-converter. The DRAM core circuit operates at V_{DL}, while the I/O circuit (L_2) and voltage converter (VL) operate at V_{DD}. The voltages during normal operation (point N) are $V_{DL} = V_{LN}$, and $V_{DD} = V_N$, while $V_{DL} = V_{LB} = k_1 V_{LN}$ and $V_{DD} = V_B = k_2 V_N$ during the burn-in test. Here, k_1 and k_2 are the acceleration coefficients of V_{DL} and V_{DD}, respectively. To successfully perform the burn-in test, the following conditions must be satisfied.

Uniform Acceleration Throughout Chip. Each node voltage during normal operation must be raised with the same acceleration coefficient. Thus, equation

$$\frac{V_{LB}}{V_{LN}} = \frac{V_B}{V_N} = k_1 = k_2 \tag{5.81}$$

must be established. Thus, points N and B are on line a in the figure.

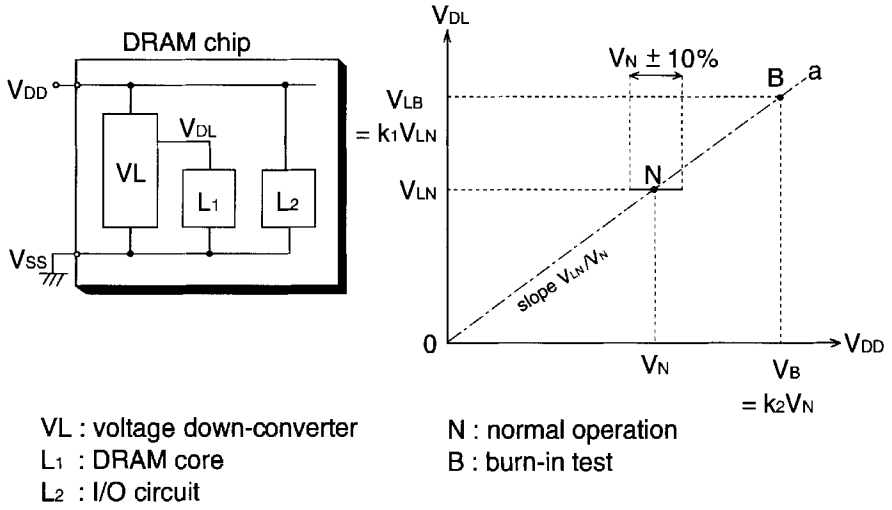


Fig. 5.64. The concept of the burn-in test for DRAMs, using an on-chip voltage down-converter [5.48]

Constant V_{DL} at Around Normal-Operation V_{DD}. To obtain stable operation, V_{DL} must be almost constant even for the maximum variation in V_{DD} (at least $\pm 10\%$) that is guaranteed at the normal-operation V_{DD}. Therefore,

$$V_{DL} = V_{LN} \qquad (0.9V_N \leq V_{DD} \leq 1.1V_N), \tag{5.82}$$

Burn-In Voltage Generation. Figure 5.65 illustrates a V_{DL} generator [5.48] that meets the above requirements. Note that since the generator was implemented for the first stage of the series-connected converter shown in Fig. 5.48b, V_{DL} in the figure corresponds to V'_{REF} in Fig. 5.48b. The converter features the switching of a line V_{RN} via a flat line V_{RN} to a line V_{RB} at a certain V_{DD} (around 6 V in the figure). The line V_{RN} governs normal-operation characteristics, while the line V_{RB} governs the burn-in (i.e. V_{LB}) characteristics. When V_{DD} is low enough, $V_{DL} = V_{DD}$ because Q_1 is always turned on (see Fig. 5.47). When V_{DD} is higher than $r_1 V_{REFN}$, however, the output V_{DL} is fixed at V_{LN} (3.3 V). On the other hand, the V_{RB} characteristics are obtained by using two comparators (CP_2 and CP_3), a PMOSFET (Q_2), and an NMOSFET (Q_3). The reference voltage, $V_{DD} - V_{REFB}$ ($V_{REFB} \simeq 2$ V), is generated based on V_{DD}. When V_{DD} is low enough, the resulting high voltage

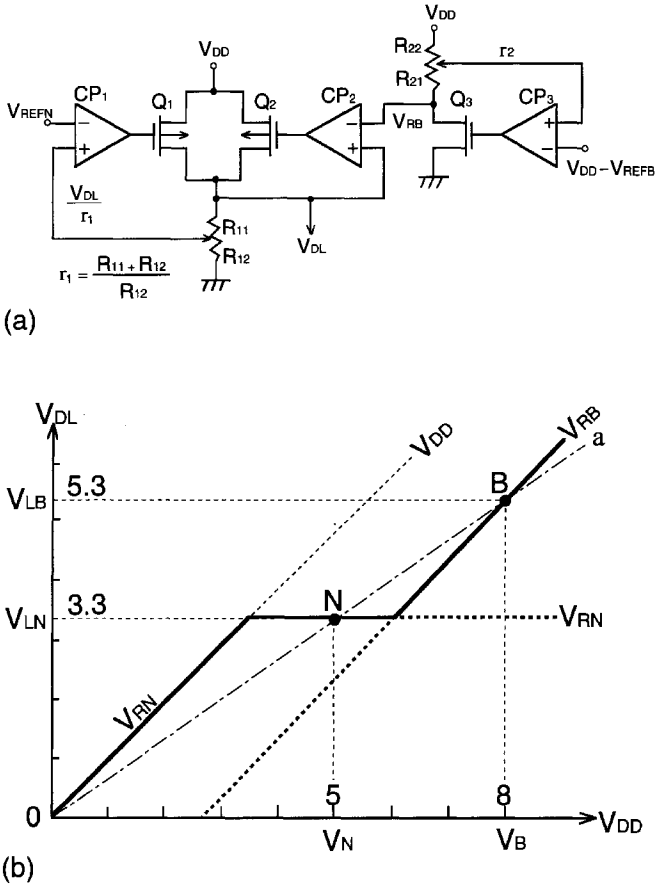


Fig. 5.65. A burn-in test circuit (**a**) and its characteristics (**b**) [5.48]

at the CP_3 output makes the CP_2 input (V_{RB}) low. Since another CP_2 input voltage is V_{DD} as a result of CP_1 and Q_1 operation, the CP_2 output voltage is high enough to cut Q_2 off. When V_{DD} is higher than $V_{DD} - V_{REFB}$, however, V_{RB} becomes $V_{DD} - r_2 V_{REFB}$. Here, $r_2 = (R_{21} + R_{22})/R_{22}$. If the resulting V_{RB} voltage is higher than V_{DL}, Q_2 is turned on while Q_1 is turned off. Thus, V_{DL} is in turn determined by V_{RB}, eventually causing the characteristics shown in Fig. 5.65. Here, the condition to ensure the flatness of V_{DL} against variations in V_{DD} at $V_N \pm 10\%$ is expressed as

$$1.1V_N - V_{LN} < V_B - V_{LB}. \tag{5.83}$$

Thus, by using (5.81), the acceleration coefficient is given as

$$k_1 = k_2 > \frac{1.1V_N - V_{LN}}{V_N - V_{LN}}. \tag{5.84}$$

For example, $k_1 = k_2 > 1.29$ is obtained for $V_N = 5\,V$ and $V_{LN} = 3.3\,V$. Since the burn-in test is usually performed with an acceleration coefficient that is much larger than this value, the V_{DL} flatness is ensured sufficiently.

Figure 5.66 shows another burn-in test circuit and set of V_{DL} characteristics [5.41]. It features an excellent V_{DL} flatness against wide variations in V_{DD} at around V_N, and a flexible V_B-setting to meet users' requirements. In

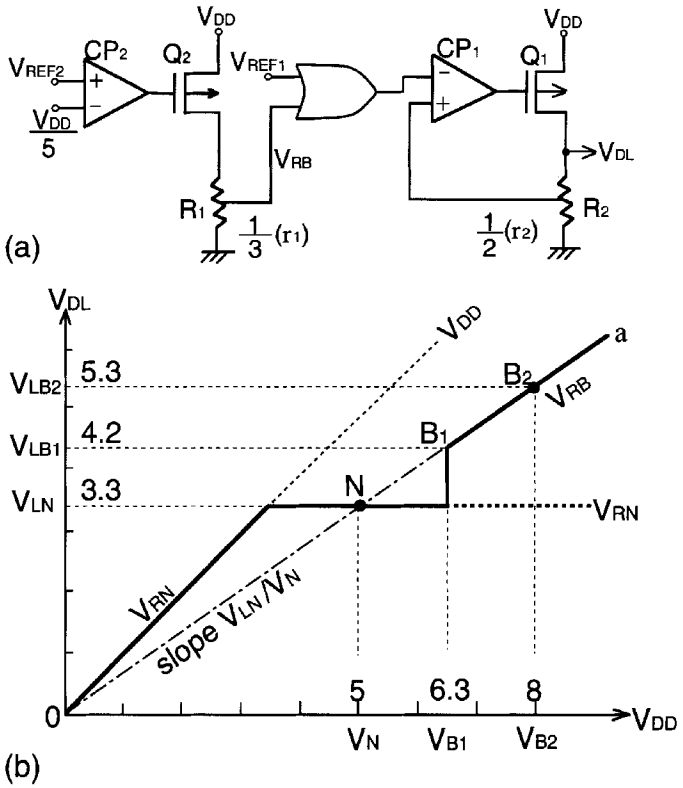


Fig. 5.66. A burn-in test circuit (**a**) and its characteristics (**b**) [5.41]

order to perform the burn-in test more quickly, a high V_B is better, as long as V_B is higher than V_N and lies on line a. In addition, the flatness of V_{DL} must be ensured, with a wider V_{DD} margin at V_N. Both requirements are satisfied with the thick solid line in Fig. 5.66, in which V_B is assumed to be higher than 6.3 V when $V_N = 5\,V$ and $V_{LN} = 3.3\,V$ (i.e. the slope of line $a = V_{LN}/V_N = 2/3$). These characteristics are realized by using the circuit shown in Fig. 5.66a and the V_{REF} generator shown in Fig. 5.63c. When V_{DD} is low enough, Q_2 is off, enabling $V_{RB} = 0$. Thus, V_{DL} is under the control of a fixed V_{REF1} (1.65 V), so that V_{LN} is fixed to 3.3 V at $V_{DD} > 2V_{REF1}$ (3.3 V) because of a voltage-division ratio of 1/2 at R_2. When V_{DD} is increased further to the extent that $V_{DD}/5$ is higher than V_{REF2} (1.26 V) – that is, $V_{DD} > 5V_{REF2}$ (6.3 V) – V_{RB} is $V_{DD}/3$ as a result of turning Q_2 on. V_{RB} is higher than V_{REF1}, allowing the V_{DL} characteristics to be governed by V_{RB} at a V_{DD} higher than 6.3 V. Eventually, V_{DL} changes at a rate of $(2/3)V_{DD}$, lying on line a.

Figure 5.67 shows another example of a burn-in test for a 16 Mb DRAM [5.40]. In this scheme the reduced voltage, V_{DL}, during normal operation is switched to V_{DD} during the burn-in test, with $V_B = V_{DD}$. The drawback is that during the burn-in test the appropriate stress voltage for the large devices in the L_2 block in Fig. 5.64 may destroy the small devices in the L_1 block. In other words, the appropriate stress voltage for the L_1 block becomes less of a stress voltage for the L_2 block. Thus, the scheme is acceptable only when the difference between V_N and V_{LN} is small, or the scale of integration of the L_1 block is much larger than that of the L_2 block, so that the reliability of the whole chip is determined by that of the L_1 block. The burn-in test starts with a burn-in test control signal, Φ_B, which is generated by a timing sequence of $\overline{WE}$ before $\overline{CAS}$ before $\overline{RAS}$, followed by n-time applications of $\overline{CAS}$. Q_2, which is turned on by the application of Φ_B, generates V_{DD} at the output. The application of $\overline{CAS}$ inhibits the start of the burn-in test, with a similar timing sequence that happens when powered on.

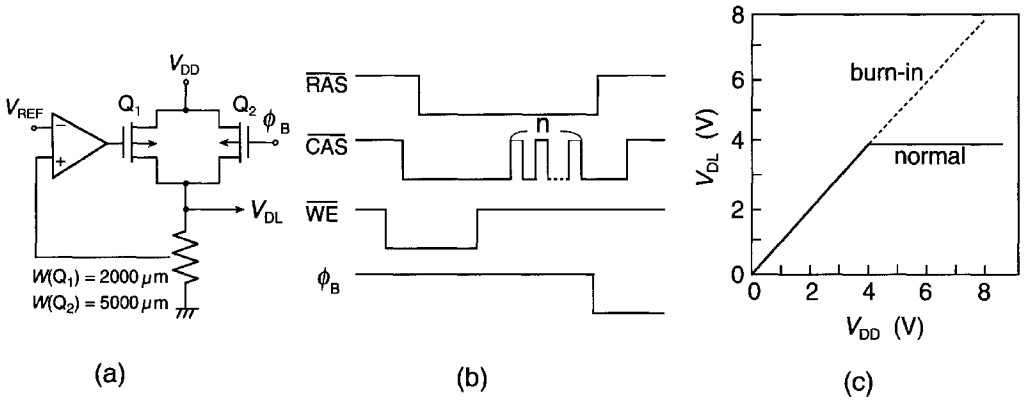


Fig. 5.67. A burn-in test circuit (**a**), its timing diagram (**b**), and its characteristics (**c**) [5.40]

5.4.7 Voltage Trimming

The variation in V_{REF} (ΔV_{REF}) caused by variations in V_{DD}, temperature, and the fabrication process must be reduced as much as possible, because the dc level of V_{DL} is directly affected by ΔV_{REF}. To minimize ΔV_{REF}, if any, V_{REF} trimming in accordance with the level of V_{DL} is thus important. Laser trimming is usually used, through the cutting of poly-Si fuses with a laser beam. Since the fuses are the same as those for redundancy, which has been widely used in commercial DRAMs, as discussed in Chap. 3, no additional photo masks and process steps are needed for the ΔV_{REF} trimming. In the following, trimming techniques for two voltages (V_{LN} and V_{LB}) are explained, using the V_{DL} generator shown in Fig. 5.65 as an example.

Figure 5.68a shows a V_{DL} generator with a trimmer [5.48]. The V_{REF} generators shown in Fig. 5.68b use the V_T difference (ΔV_T) scheme in Fig. 5.62. A bias-current circuit composed of a mirror circuit gives both the V_{REFN} and V_{REFB} generator bias-currents (I and $2I$), enabling the same current (I) to flow to Q_1 and Q_2. Here, V_{REF} is equal to twice ΔV_T because of the series connection of two identical MOSFETs. The doubled ΔV_T is convenient in the design of the trimmer. Q_4 in the V_{REFN} generator improves the constant-current characteristics of Q_3 by suppressing the dependence of the current on the drain–source voltage [5.54]. Here, variations in the fabrication process change V_{REFN} and V_{REFB}, eventually causing V_{RN} and V_{RB} variations as large as $\pm 20\%$ for the gate-oxide and dopant variations of $\pm 10\%$ [5.47]. Even if such variations occur, V_{RN} ($= r_1 V_{REFN}$) can be adjusted to any level through the trimming of R_{11}/R_{12} (i.e. r_1). The variation of V_{RB} can also be adjusted by the trimming of r_2. Figure 5.68c shows a decoded trimmer using fuses [5.48]. The trimmer allows the input voltages (V_{FN}, V_{FB}) to the comparators (CP_1, CP_2) to be precisely controlled by trimming of r_1 and r_2 through programming of fuse-ROMs. One of voltages divided by resistor is selected by a decoder as a result of the programming of three fuse-ROMs. When the voltage V_i is selected, V_{RN} is expressed, by using the condition $V_{FN} = V_{REFN}$, as

$$V_{RN} = \frac{R_0 + R_1 + \cdots + R_8}{R_0 + R_1 + \cdots + Ri} V_{REFN}. \tag{5.85}$$

V_{RB} is also given by using the condition $V_{FB} = V_{DD} - V_{REFB}$, as

$$V_{DD} - V_{RB} = \frac{R_0 + R_1 + \cdots + R_8}{R_0 + R_1 + \cdots + Ri} (V_{DD} - V_{REFB}). \tag{5.86}$$

The circuit enables the minimum number of fuses, which are due to the use of the decoder: three fuses for normal operation and another three fuses for the burn-in test. Hence, the number of fuses cut is six at most, enabling a negligible additional cost for trimming.

A V_{DL} generator with the above trimmers has actually been implemented on a 16 Mb DRAM [5.48]. The experimental results were as follows:

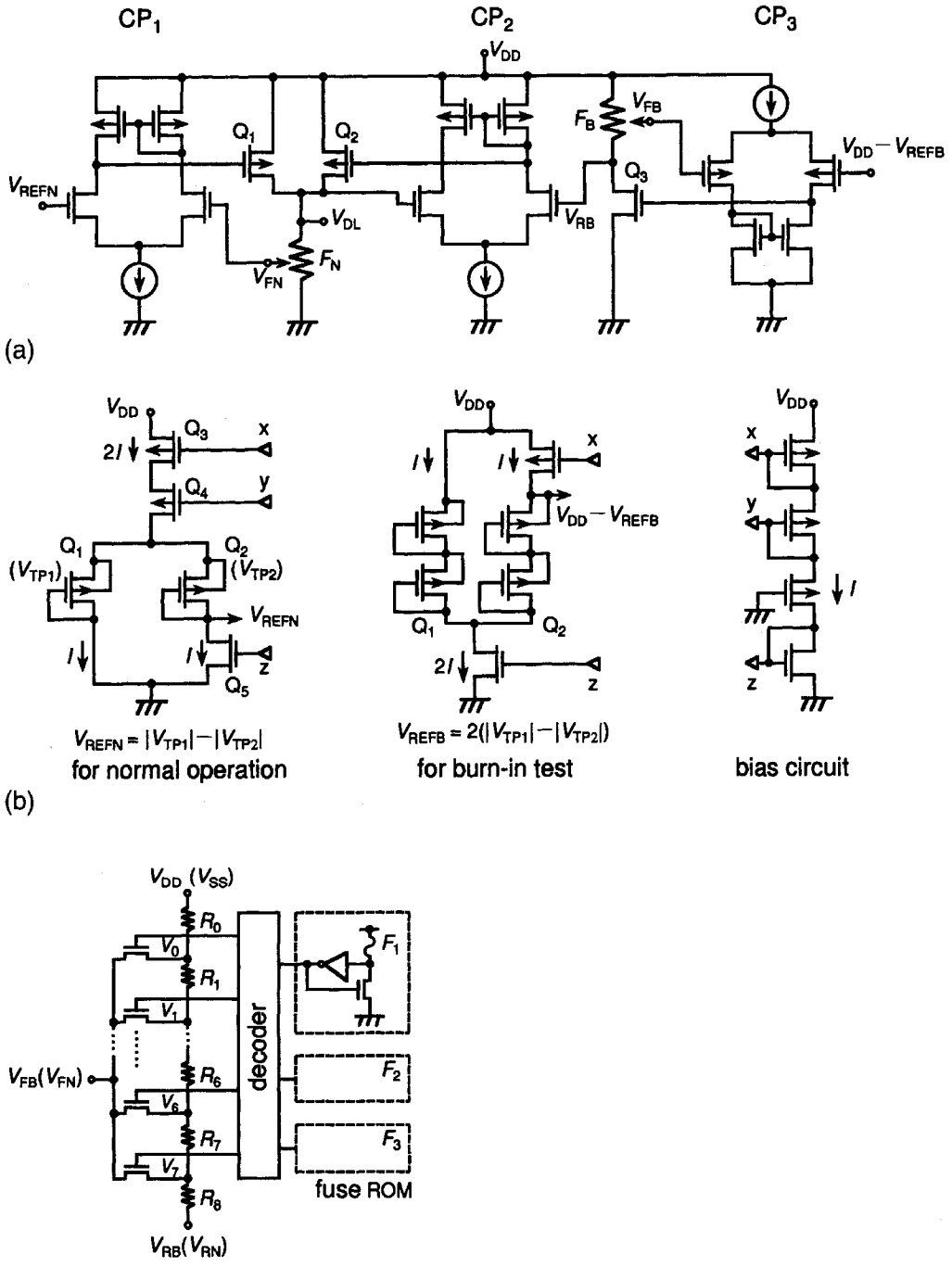


Fig. 5.68. Trimming circuits for V_{DL} (**a**) and reference voltages (**b**), and a decoded trimmer (**c**) [5.48]

$V_{LN} = 3.3\,V$ at $V_{DD} = 5\,V$, $V_{LB} = 5.3\,V$ at $V_{DD} = 8\,V$, $\Delta V_{LN} = \pm 9\,mV$ at $\Delta V_{DD}(4.5\text{–}5.5\,V) = 1\,V$, $\Delta(V_{DD} - V_{RB}) = \pm 50\,mV$ at $\Delta V_{DD}(7\text{–}8\,V) = 1\,V$, $\Delta V_{LN} = \pm 84\,mV$ and $\Delta V_{LB} = \pm 70\,mV$ at $\Delta T(20\text{–}120\,°C) = 100\,°C$, and $\Delta V_{LN} = \Delta V_{LB} = \pm 130\,mV$ at $\Delta V_T = \pm 0.3\,V$. The generator occupied only 0.6% of the chip area of $127\,mm^2$. Moreover, a 16 Mb DRAM [5.62] with a generator incorporating a V_{REF} generator shown in Fig. 5.61d revealed that laser-trimming allowed the center value of V_{DL} to shift from 4.6 V to 4.0 V, while reducing ΔV_{DL} from 0.7 V to 0.2 V.

5.4.8 Low-Power Circuits

A DRAM chip must maintain its low-power advantages, such as a low data-retention power, even if a V_{DL} generator is incorporated. Thus, the generator must provide a small enough current when the load circuit does not need a large current. However, the generator must deliver a current that is large enough to suppress the voltage dip (see ΔV_{DL} in Fig. 5.51) during the active period. The schemes of low-power circuits to meet these requirements are similar to those of the V_{BB} generators discussed previously.

A typical example is shown in Fig. 5.69, which is similar to the basic generator in Fig. 5.47 if the PMOSFETs and NMOSFETs are exchanged. The constant-current source composed of Q_2 and Q_3 is controlled by a clock, Φ [5.43, 5.57]. For example, the current – that is, I_S in (5.29) – is increased by turning on Q_S synchronously with the charging up of many data lines that involve a large current. After completing the charging up, Q_3 is turned off to save power. Q_2 is always turned on with a small current to maintain the V_{DL} level while the V_{DL} load current is quite small. The values of the power dissipation were 15 mW and 2 mW in the active mode and the stand-by mode, respectively.

Figure 5.70 shows a V_{DL} generator [5.40] that delivers the current only when V_{DL} is changed due to a load operation. Before activation, the comparator and Q_1 are off because CBR is raised to $V_{REF1} + V_{TP}$. Since the Q_2 gate voltage is pumped to $V_{REF1} + V_{TP} + V_{TN}$, the output voltage V_{DL} is

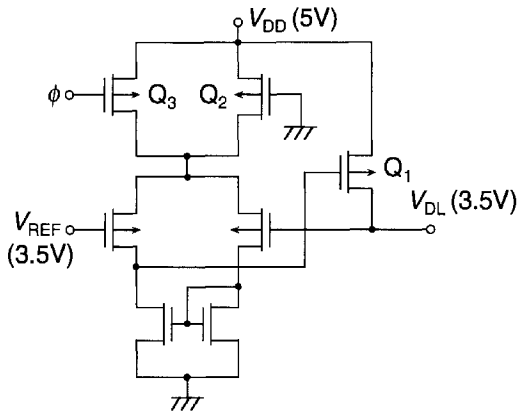


Fig. 5.69. A current-switching V_{DL} generator [5.44]

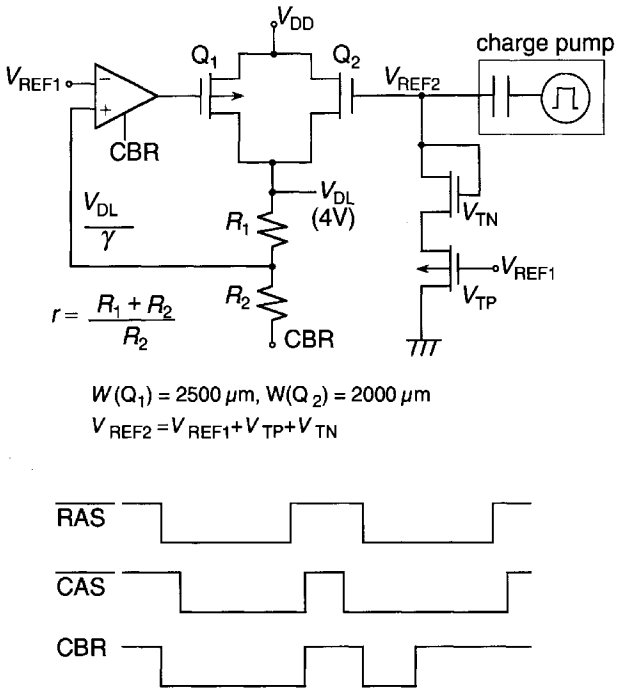


Fig. 5.70. A V_{DL} generator activated by a load operation [5.40]

$V_{REF1} + V_{TP}$, preventing current flow through R_1 and R_2. Even when CBR falls to 0 V, Q_1 remains off although the comparator is activated and V_{DL} stays almost at $V_{REF1} + V_{TP}$, if $(R_1 + R_2)$ is high enough and the Q_1 cutoff condition

$$\frac{V_{REF1} + V_{TP}}{r} \geq V_{REF1}$$

is satisfied. When a change in V_{DL} (ΔV_{DL}) occurs, however, Q_1 is turned on so as to suppress ΔV_{DL}, if ΔV_{DL} satisfies

$$\frac{V_{REF1} + V_{TP} - V_{DL}}{r} \leq V_{REF1}.$$

It has been reported that the generator consumes 4 mA in the active mode and 20 µA in the stand-by mode, with an area penalty of 0.16% for a 16 Mb DRAM.

Figure 5.71 shows the parallel connection of two generators [5.45]: a low-power generator for supplying a small current (0.2 mA at $V_{DD} = 5$ V) during the stand-by mode, and a high-power generator for supplying a large enough current during the active mode, with Φ-activation.

Reduction of the V_{REF}-generator current is also important, since even the V_{REF} circuit in Fig. 5.62a always consumes 10–100 µA. Note that the stand-by current of the whole DRAM chip must be confined to less than 10 µA, exclusive of the refresh-relevant current. Figure 5.72 shows a dynamic V_{REF} generator [5.61] that realizes a current of less than 1 µA, in which the static

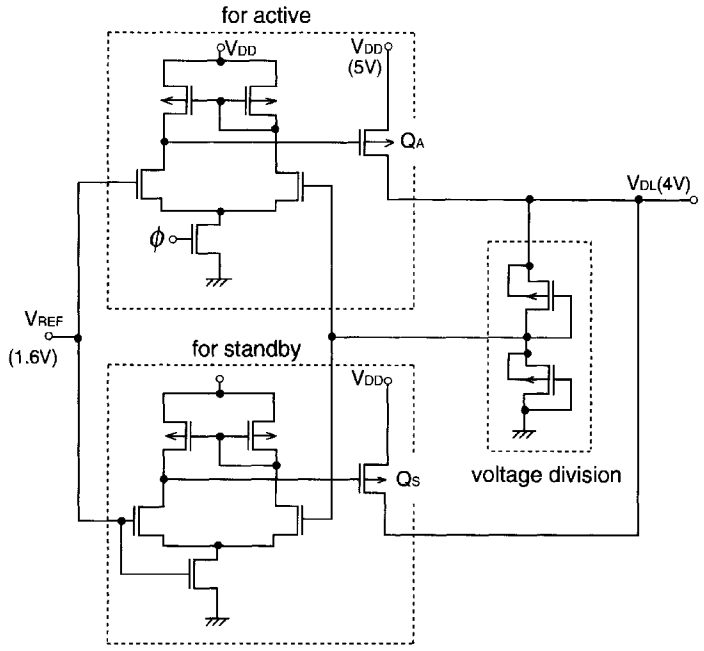


Fig. 5.71. The parallel connection of two generators, one for active mode and the other for stand-by mode [5.45]

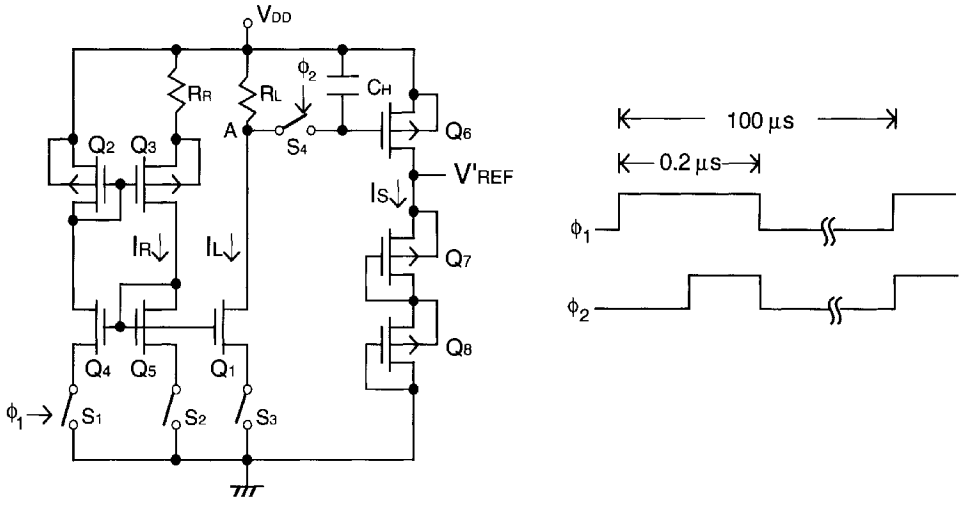


Fig. 5.72. A dynamic V_{REF} generator [5.61]

current mirror shown in Fig. 5.62b is changed to a dynamic one. It features dynamic operation, like a sample and hold technique. The current path is enabled when Φ_1 is applied, and the output voltage is sampled on the hold capacitance C_{H} while Φ_2 is applied. These control pulses are generated from an on-chip self-refresh circuit. For a Φ_1 pulse width of 200 ns and a sampling interval of more than 100 μs, the total current is reduced to only 0.5 μA.

5.5 The Half-V_{DD} Generator

The voltage setting accuracy and transient characteristics for variations in V_{DD} and load operations are the keys to designing a half-V_{DD} generator, because a half-V_{DD} power supply is closely related to the cell signal voltage and its reference voltage of sensing.

Figure 5.73 shows a half-V_{DD} generator [5.66] consisting of a bias circuit and push-pull output circuit. If R_1 and R_2 are high enough resistances and the MOSFETs have the same value of V_T, the voltages at nodes N, N_1, and N_2 are $V_{DD}/2$, $(V_{DD}/2)+V_T$, and $(V_{DD}/2)-V_T$, respectively. Thus, the output is stabilized at $V_{DD}/2$. Since both source–gate voltages of Q_{N2} and Q_{P2} are V_T, they are weakly turned on, allowing a small subthreshold current to flow. The resultant stand-by current is 50 µA at $V_{DD} = 5$ V. Any change in the output voltage is eventually suppressed, because either of the two is strongly turned on, providing several mA. A switching scheme [5.31] of two generators similar to the above circuit, one dedicated to $V_{DD}/2$ for normal operation and the other dedicated to $(V_{DD}-V_T)/2$ for battery back-up mode, increased the cell data-retention time of the back-up mode by 20% by switching the data-line precharge level from $V_{DD}/2$ to $(V_{DD} - V_T)/2$. However, this generator has the following drawbacks when V_{DD} is lowered. The voltage setting accuracy gets worse for a fixed variation in V_T, as exemplified by a ±0.1 V output fluctuation caused by a ±0.1 V independent fluctuation in V_T for PMOS and NMOSFET. In addition, a high-speed response is more difficult to achieve for the ever-increasing load capacitance with increasing memory capacity. For example, for a load (i.e. a cell-capacitor plate) capacitance of a 64 Mb DRAM as heavy as 115 nF, the rise time of the $V_{DD}/2$ waveform during power-on is as long as 160 µs at $V_{DD} = 1.5$ V [5.38]. Such a slow response might be detrimental to normal operation.

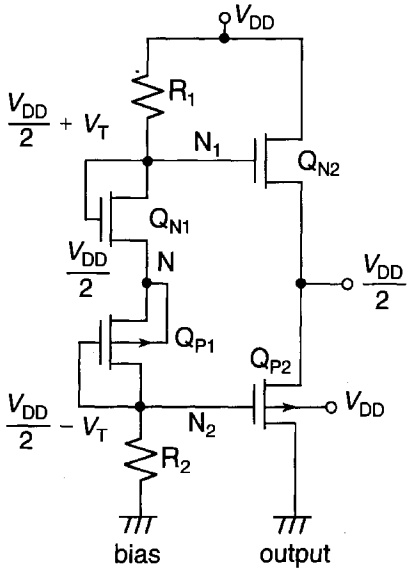


Fig. 5.73. A half-V_{DD} generator [5.66]

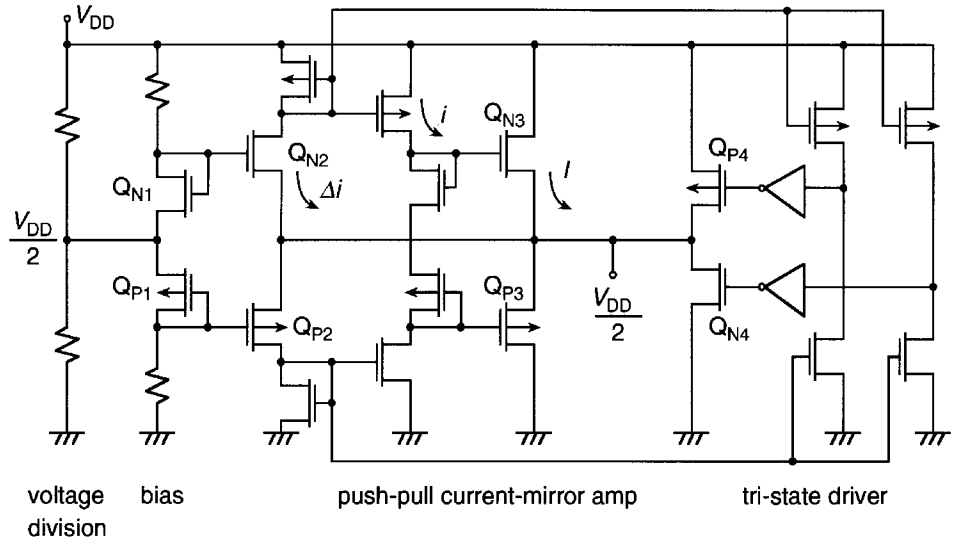


Fig. 5.74. A half-V_{DD} generator for low-voltage operation [5.38]

The generator [5.38] shown in Fig. 5.74 solves the possible problems mentioned above. One feature is the separation of the voltage-division circuit and the bias circuit, to improve the accuracy by about one order. Only the difference in V_T between the MOSFETs with the same conduction type (Q_{N1} and Q_{N2}, or Q_{P1} and Q_{P2}), which is usually small, is related to the accuracy. Another feature is the addition of a push-pull current-mirror amplifier and a tri-state output driver to speed up operation. The amplifier outputs $V_{DD}/2$ at steady state, because the gate voltages of Q_{N2} and Q_{P2} are $V_{DD}/2 + V_T$ and $V_{DD}/2 - V_T$, respectively. When the load voltage falls from $V_{DD}/2$ during a load operation, the charging current Δi flows as a result of more forward biasing of Q_{N2}. Thus, the mirrored current i charges up the Q_{N3} gate to V_{DD}. The resulting large current I charges up the $V_{DD}/2$ load, so that the load recovers to the original $V_{DD}/2$. On the contrary, when the load voltage increases, the lower part of the amplifier suppresses the load-voltage variation in the same manner. The tri-state driver enhances the driving capability. This generator has achieved a 30-fold speed increase at $V_{DD} = 1.5\,V$, compared with that in Fig. 5.73.

5.6 Examples of Advanced On-Chip Voltage Generators

This section describes consistently the most advanced on-chip voltage generators [5.76] applicable to low-voltage gigascale DRAMs. They generate concurrently a negative quiescent word-line voltage (the so-called NWL scheme, described in Chap. 4) and a boosted active word-line voltage. To obtain well-regulated voltages, they fully utilize the circuit features of the on-chip voltage generators explained so far.

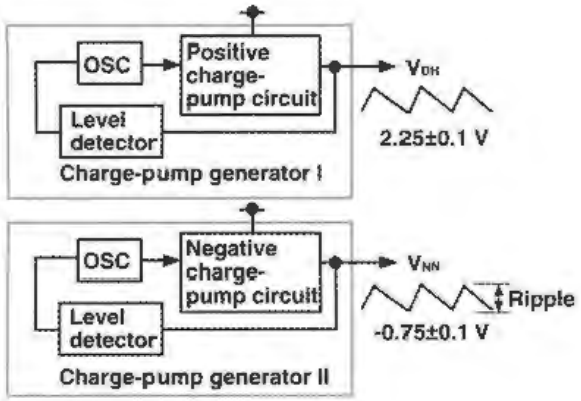


Fig. 5.75. A conventional charge-pump generator [5.76]

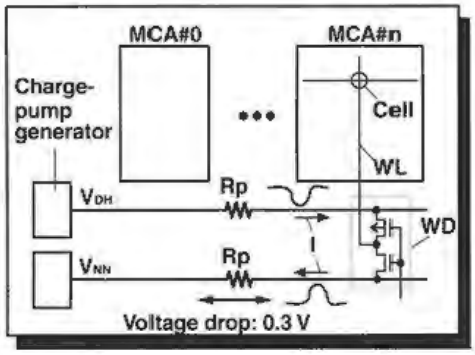


Fig. 5.76. The application of a conventional charge-pump generator to a memory array in a DRAM chip [5.76]

Figure 5.75 shows conventional charge-pump generators for high voltage (V_{DH}) and negative voltage (V_{NN}) in a DRAM chip. There are two problems with this design. One is the voltage drop caused by the wiring resistance. The other is ripple of about 0.2 V peak-to-peak that stems from the hysteresis of the level detector used for stabilization. Since the charge pumps are placed at a local site (Fig. 5.76), due to their large layout area and reduction of the minority carrier injection into the substrate, the wiring resistance from the pump to the word-line driver (WD) can be as high as 100 Ω. If a current of 3 mA passes through this resistance, the voltage difference across the resistance will be about 0.3 V. This causes V_{DH} to drop and V_{NN} to rise. These degradations result in impossible full write operation of the cell and a shortened data-retention time. The ripple accelerates these detrimental effects. To prevent these effects, the absolute values of V_{DH} and V_{NN} must be set higher as a result of the voltage variations than the required minimum value. However, this causes an excessive stress voltage to scaled devices.

Figure 5.77 shows a hybrid generator consisting of a charge pump and a series-pass regulator. Both V_{DH} and V_{NN} are provided through the hybrid generator. Since the regulator is just the same as the voltage down-converter discussed previously in circuit configuration, the ripple is suppressed by its excellent Power Supply Rejection Ratio (PSRR) of more than 30 dB at

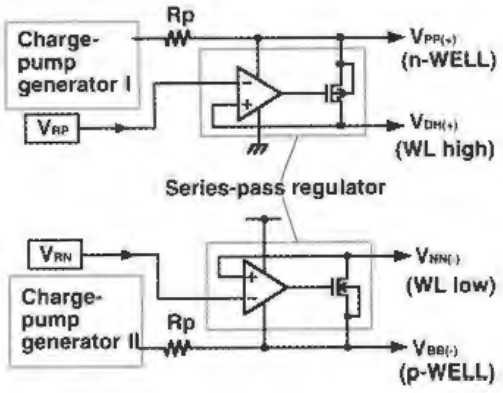


Fig. 5.77. A hybrid generator [5.76]

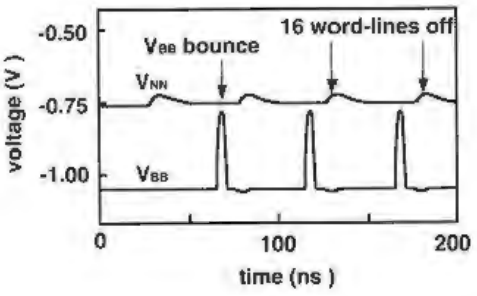


Fig. 5.78. A simulated V_{NN} waveform for word-line switching and V_{BB} bouncing [5.76]

0–10 MHz [5.53]. Also, the regulator can be placed near the word-line driver because of its small layout area. This sufficiently reduces the word-line voltage drop caused by the wiring resistance R_P, even if the charge pumps are placed at a local chip site. Here, the outputs of the charge pumps, V_{PP} and V_{BB}, are connected to the n-well and the p-well in the memory cell array. This reduces the size of the smoothing capacitor because the parasitic capacitance of the n-well and the p-well is more than 1000 pF in gigascale DRAMs. Figure 5.78 shows a simulated V_{NN} waveform for word-line switching and V_{BB} bouncing. The V_{NN} noise is negligible (less than 5 mV) with respect to the 250 mV V_{BB} noise. Also, the word-line switching noise of the output, when the word lines are turned off, is suppressed to as low as 30 mV because of the low output impedance of the generator.

Figure 5.79 shows the requirements for a reference-voltage generator for DRAMs with a negative word-line scheme. The negative reference voltage V_{RN} should be held constant with respect to V_{SS}, and the positive reference voltage V_{RP} should be kept higher than V_{DD} (or V_{DL}) by the threshold voltage of the memory-cell FET, V_{TM}. Since variations in V_{RN} and V_{RP} degrade reliability in a memory-cell FET, they must be highly accurate. In practice, the allowable variation is about ±100 mV. Figure 5.80 shows a reference-voltage generator. This circuit has two main features. One is the low-voltage band-gap generator using a differential amplifier, shown in Fig. 5.63b. The other is the offset voltage generator, which uses a current-mirror circuit. The

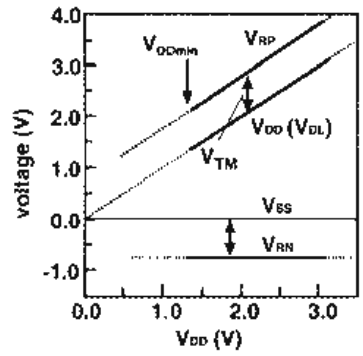


Fig. 5.79. Reference voltage requirements [5.76]

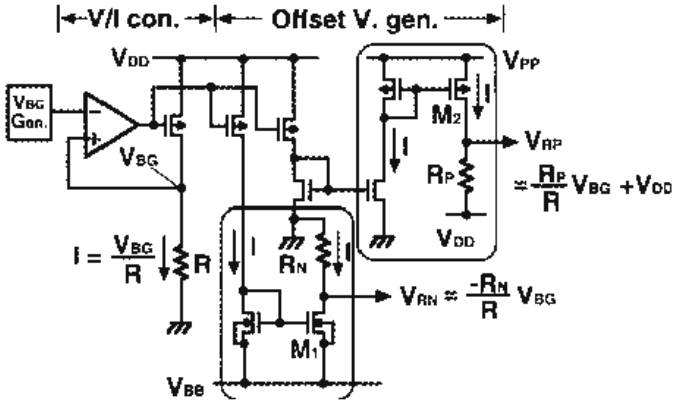


Fig. 5.80. A reference-voltage generator [5.76]

offset voltage generator ensures highly accurate output voltages and I/V conversion resistors with a small layout area. The circuit operates as follows. The V/I converter converts the band-gap voltage V_{BG} into the reference current I, which is expressed as V_{BG}/R. The mirror current I in resistor R_{N} flows from V_{SS} to V_{BB}. Therefore, V_{RN} is expressed as follows:

$$V_{\mathrm{RN}} = -(R_{\mathrm{N}}/R)V_{\mathrm{BG}}. \tag{5.87}$$

In a similar way, the mirror current I in R_{P} flows from V_{PP} to V_{DD}. Thus, V_{RP} is expressed as follows:

$$V_{\mathrm{RP}} = (R_{\mathrm{P}}/R)V_{\mathrm{BG}} + V_{\mathrm{DD}}. \tag{5.88}$$

Because the resistance ratio is independent of temperature and process variations, V_{RN} and V_{RP} remain constant despite those variations. Since a resistance is not needed to create the voltage between V_{DD} and V_{SS} in this configuration, precise offset voltages above V_{DD} and below V_{SS} are created with a small layout area. In addition, since V_{PP} increases with respect to V_{DD}, and V_{BB} is constant with regard to V_{SS}, the drain–source voltages of the current mirror FETs, M_1 and M_2, are held almost constant, to eliminate the error in the output voltages caused by the drain conductance.

The whole generator measures $420\,\mu m \times 506\,\mu m$, with a $0.3\,\mu m$ DRAM process. The experimental data are as follows. The current consumption is $7\,\mu A$ at $V_{DD} = 3.5\,V$, the output voltage error is less than $-100\,mV$ for V_{RP} and less than $+50\,mV$ for the V_{RN} from $1.3\,V$ to $3.5\,V$ of the V_{DD} without trimming, and the temperature dependency is less than $500\,ppm/°C$.

6. High-Performance Subsystem Memories

6.1 Introduction

In the past, the high-speed SRAM cache and the DRAM main memory have contributed to enhancing the performance of hierarchical memory systems (Fig. 6.1) in personal computers (PCs), workstations, servers, and so on. Recently, however, to bridge the ever-increasing speed gap between microprocessor units (MPUs) and main memories, high-speed DRAMs [6.1–6.5] such as the synchronous DRAM (SDRAM) and the Rambus DRAM have become increasingly important. These DRAMs incorporate memory subsystem technologies that memory-system designers have used to enhance not only access and cycle times, but also throughput. Here, throughput is defined as the product of bus width and frequency. The resulting high throughput of the DRAM chips has been indispensable even for PC graphics, image processing,

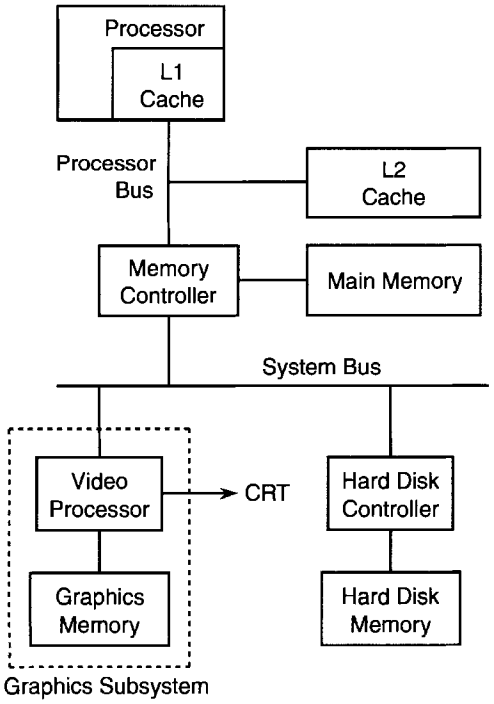


Fig. 6.1. The graphics memory in a personal computer [6.6]

and other multimedia applications [6.6–6.8]. In particular, high throughput is a key to high-performance graphics systems, as shown in Fig. 6.1. There are two missions for a graphics memory. One is to output data to display devices such as the cathode-ray tube (CRT). The other is to transfer drawing data to and from the video processor. Display performance has achieved 200 Mbyte/s, but the perception of the human eye is limited in terms of resolution and after-image, and therefore the high-performance requirement is getting saturated. The need for high-drawing performance, however, is increasing incessantly. Graphics expression was originally text-based, and then the two-dimensional (2-D) still picture, but now it has reached natural and moving pictures, and is progressing toward three-dimensional (3-D) moving pictures, that need a greater improvement in performance. Graphics memories have been designed using DRAM technology. In the past, the dominant form of graphic memory was video RAM (VRAM), which has independent display and drawing terminals (i.e. a dual port), as shown in Fig. 6.2. For example, while outputting serial data from one port to a low-performance display, the other port can be used for high-speed random-access drawing. However, the performance of the serial port is so poor that VRAM cannot deal with high-speed block data processing, which is essential for modern audio/video equipment. This problem is solved by high-performance single-port RAMs such as SDRAM, in which the single port is used for both display and drawing. Recently, even video data is being stored in the main memory for 3-D graphics applications of low-end personal computers, in which the graphics memory is eliminated. In this case, high-performance DRAMs are strongly needed, since both the MPU and the video processor frequently access the same main memory. Games machines that deal with 3-D graphics also need high-performance DRAMs.

Meanwhile, custom memories have been also crucial. In particular, embedded DRAMs are emerging architectures that boost the performance of standard DRAMs. Fortunately, recent developments [6.9, 6.10] imply that state-of-the-art DRAM miniaturization technology has at last progressed to the point at which embedded DRAMs are becoming attractive in terms of cost. The elimination of the large buffer and bonding pad area for many I/Os, and the reduction of I/O noise and power, also favor embedded DRAMs. At present, the question of which approach is better in terms of cost/performance

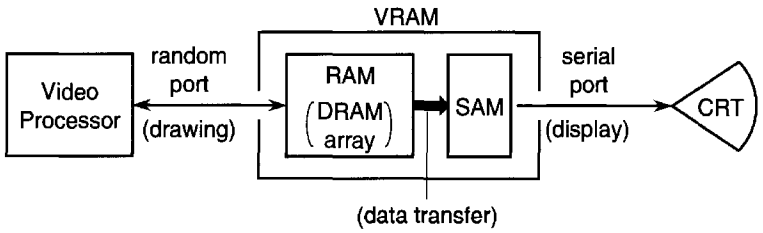


Fig. 6.2. The concept of video RAM (VRAM) [6.8]. SAM, serial access memory

and availability is a major concern: standard DRAMs combined with high-density packaging technology or embedded DRAMs.

This chapter describes high-performance subsystem memories, especially for a hierarchical memory system. First, improvements of memory-subsystem performance in a memory hierarchy consisting of a cache SRAM and a DRAM main memory, and memory-chip performance that plays a critical role in these improvements, are discussed. A comparison of performance between DRAMs and SRAMs is also included here. Second, memory-subsystem technologies that are essential for high-speed DRAM chips are explained. Third, trends in high-speed DRAMs that use the above technologies are reviewed, and then an SDRAM and a Rambus DRAM are described in detail. Finally, embedded memories are explained, with an emphasis on the DRAM.

6.2 Hierarchical Memory Systems

6.2.1 Memory Hierarchy

To realize a substantially unlimited amount of fast memory at a low cost, the memory of a processor is implemented as a memory hierarchy. A memory hierarchy consists of multiple levels of memory with different speeds and size. The upper level (i.e. the one closer to the processor) is smaller and faster than the lower level. The upper level is a subset of any lower level: all data at one level is also found at the level below, all data at that lower level is found at the one below it, and so on, until we reach the bottom of the hierarchy. The minimum unit of information that can be either present or not present in the two-level hierarchy is called a block. If the data requested by the processor appears in some block in the upper level, this is called a "hit". Hit time is the time to access the upper level of the memory hierarchy, which includes the time needed to determine whether the access is a hit or a miss. If the data is not found in the upper level, the request is called a miss. The lower level in the hierarchy is then accessed to retrieve the block containing the requested data. The miss penalty is the time to replace a block in the upper level with the corresponding block from the lower level, plus the time to deliver this block to the processor. Because the upper level is smaller, and is built using a faster memory, the hit time is much smaller than the time to access the next level in the hierarchy, which is the major component of the miss penalty. System performance trends to improve with increases in the memory size (i.e. the capacity) of the upper level and the block size because of the increased hit rate. In addition, it improves with a reducing miss penalty, which is closely related to memory-system organization and memory-chip performance.

Figure 6.3 shows a typical example composed of SRAM caches, DRAM main memory, and hard disk memory. By adding another level of cache between the original cache (i.e. the first-level cache – the L1 cache) and the main memory, the L1 cache can be made small enough to match the

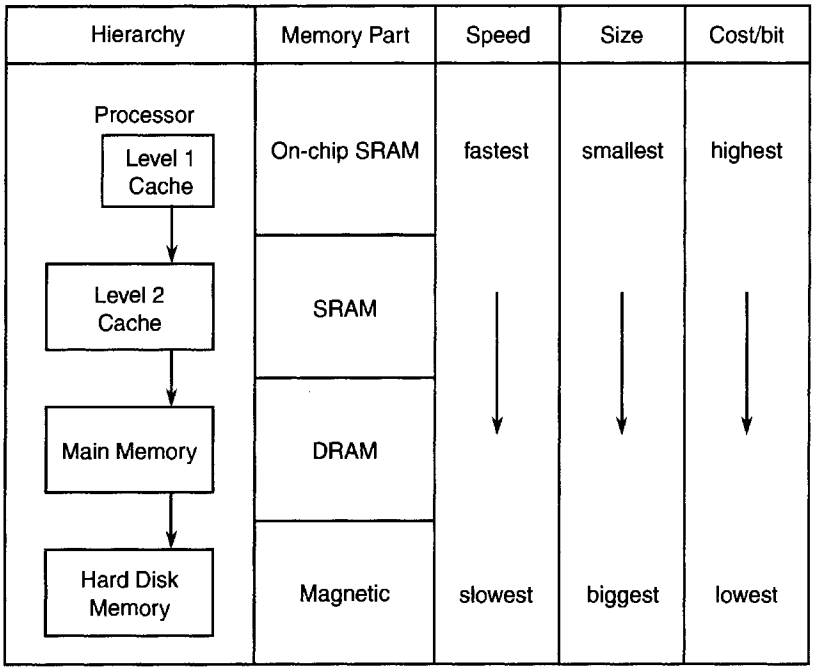


Fig. 6.3. A memory hierarchy of a processor [6.2, 6.3]

clock cycle of the fast MPU, while the second-level cache (the L2 cache) can be made large enough to capture many accesses that would go to the main memory, thereby lessening the effective miss penalty. A processor in search of data or instructions looks first in the L1 cache, which is closest to the processor. If the information is not found there, the request is passed on to the L2 cache. Naturally, while awaiting the needed information, the processor is not fully utilized. Both the performance and the throughput (bandwidth) are improved by the integration of SRAM on to the processor assembly, as an on-chip L1 cache. Traditionally, SRAM cells that are integrated into the processor chip are larger than in stand-alone commodity SRAM cells, and may also have performance-enhancing features such as dual-port capability. They are fabricated with the same process as is used for the processor logic and usually have the highest performance possible. An L1 cache is faster than the L2 off-chip caches which tend to be SRAM as well, though these often rely on specialized SRAM processes that minimize cell area. The relative sizes and performance of memory units are conditioned primarily by economic considerations, with roughly equal costs allocated to the different cache levels of memory. This equipartition of costs to the memory subsystem components leads to an L1 that is smaller than an L2 and so on, ensuring the best possible performance for a given overall system cost.

Figure 6.4 shows a block diagram [6.11] of a typical personal computer system using the Pentium microprocessor, in which the block size of the on-chip L1 cache is 32 bytes (i.e. 256 bits). This block must be replaced by data from the off-chip L2 cache through a 64-bit CPU data bus, when a miss hit in

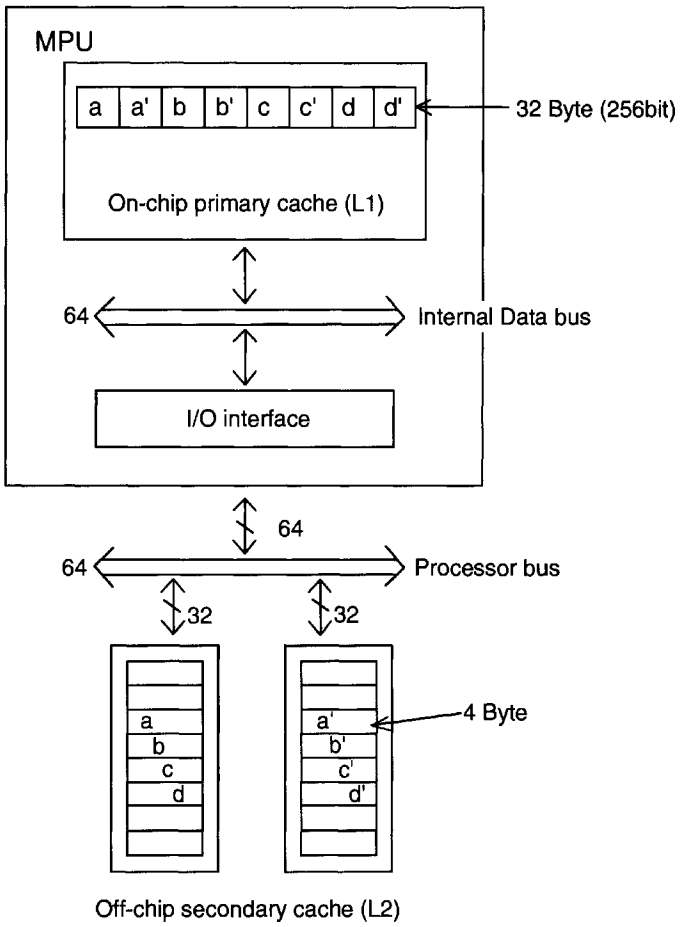


Fig. 6.4. A block diagram of a typical personal computer system, using the Pentium microprocessor [6.11]

the on-chip cache occurs. Thus, the replacement cannot be accomplished at one time, but needs four cycle times. This implies that the replacement time depends not only on the first access time of the SRAM, but also on the cycle time of the subsequent four cycles. Thus, the cycle time (in other words, the throughput), as well as the access time, is an important factor affecting the performance of cache memories. Such is the case between two adjacent levels of L2 cache and DRAM main memory.

As processor performance has increased, the wait time in idle cycles or memory stall cycles has also risen to an unacceptably high level. This latency is aggravated by the fact that the main memory (DRAM) performance, coupled with off-chip access times, has not kept pace with improvements in processor performance, giving rise to the so-called memory-processor performance gap (Fig. 6.5). Thus, in addition to the fast access time, high throughput is becoming increasingly important, as exemplified by demands on DRAMs from personal computers shown in Fig. 6.6 [6.4, 6.17].

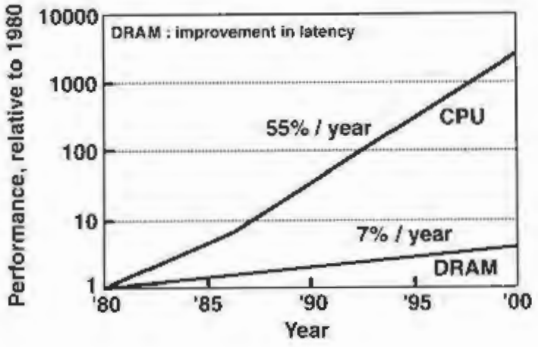


Fig. 6.5. A performance comparison between the processor (CPU) and the DRAM [6.3]

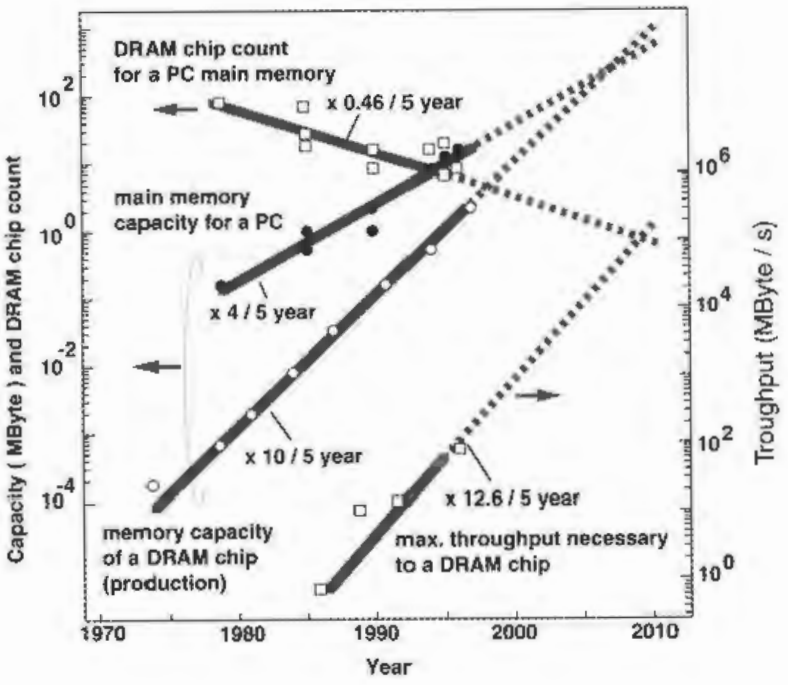


Fig. 6.6. The demands made on DRAM chips by PCs [6.14, 6.17]

6.2.2 Improvements in Memory-Subsystem Performance

The clock frequency of an MPU is rapidly increasing from 500 MHz at present toward 1 GHz, using sophisticated internal architectures and pipelining to process almost all instructions with one clock. In accordance with the ever-increasing MPU performance, the peak throughput of the system bus between the MPU and the memory has been also improved, so that instructions and data are fast and are successively provided from memory to MPU when hitting a miss. For example, the bus frequency of PCs has been enhanced from 66 MHz to 100 MHz, and then to 133 MHz. The bus width has also increased from 16 bit to 32 bit, and then to 64 bit for the Intel 8086, 386,

and Pentium generations, respectively. Engineering workstations (EWS) and even graphics systems need a 128-bit width. Note that a peak throughput as large as 1 Gbyte/s is attainable with a 133 MHz 64-bit bus. Actually, the peak throughput of the bus is not realized because of degradations of the throughput of memory, and the use of the bus by memory. In a memory hierarchy, these degradations are closely related to the cache miss rate and the cache miss penalty, as will be explained with a two-level memory hierarchy composed of a SRAM cache and DRAM main memory.

Reducing Cache Misses. The simplest way to reduce the miss rate is to increase the block size as follows. Since programs access a relatively small portion of their address space at any instant of time, if an item is referenced, other items whose addresses are close by will tend to be referenced soon. To take advantage of this spatial locality, a cache must have a block size larger than one word. Thus, when a miss occurs we can fetch multiple words that are adjacent and carry a high probability of being needed shortly. Obviously, the miss rate falls when the block size is increased. An excessive increase, however, increases the miss penalty as follows. The miss penalty is determined by the time required to fetch the block from the next lower level of the hierarchy and load it into the cache. The time to fetch the block has two components: the latency to the first word, and the total transfer time of the remaining words of the block, as suggested in Fig. 6.4. Clearly, the transfer time increases as the block size grows, and thus the miss penalty also grows.

Reducing the Cache Miss Penalty. The cache miss penalty is closely related to memory organization, bus utility, and memory-chip performance.

Memory Organization [6.2, 6.3]. In the case of the DRAM main memory, it is difficult to reduce the latency to fetch the first word from memory due to the inherently slow access time of the DRAM. Instead, the miss penalty can be reduced if the throughput from the memory to the cache is increased. The primary method of achieving higher memory throughput is to increase the physical or logical width of the memory system. Let us compare the miss penalty for various memory organizations, shown in Fig. 6.7. Here, it is assumed that a hypothetical system with a cache block of four words and four bytes per word needs one clock cycle to send the address to the memory, ten clock cycles for the access time per word, and one clock cycle to send a word of data.

Figure 6.7a shows one-word-wide memory organization, in which all components are one word wide and all accesses are made sequentially. The data-I/O configuration of each memory chip is assumed to be 2 Mwords × 32 bits. The miss penalty of the memory would be $1+4\times10+4\times1 = 45$ clock cycles. Thus, the number of bytes transferred per clock cycle for a single miss would be $4\times4/45 = 0.35$. Figure 6.7b shows the organization widening the bit width of the memory and buses between the MPU and the memory. This allows parallel access to four words of the block. The miss penalty and the number of bytes would be $1+1\times10+1\times1 = 12$ clock cycles and $16\times1/12 = 1.33$,

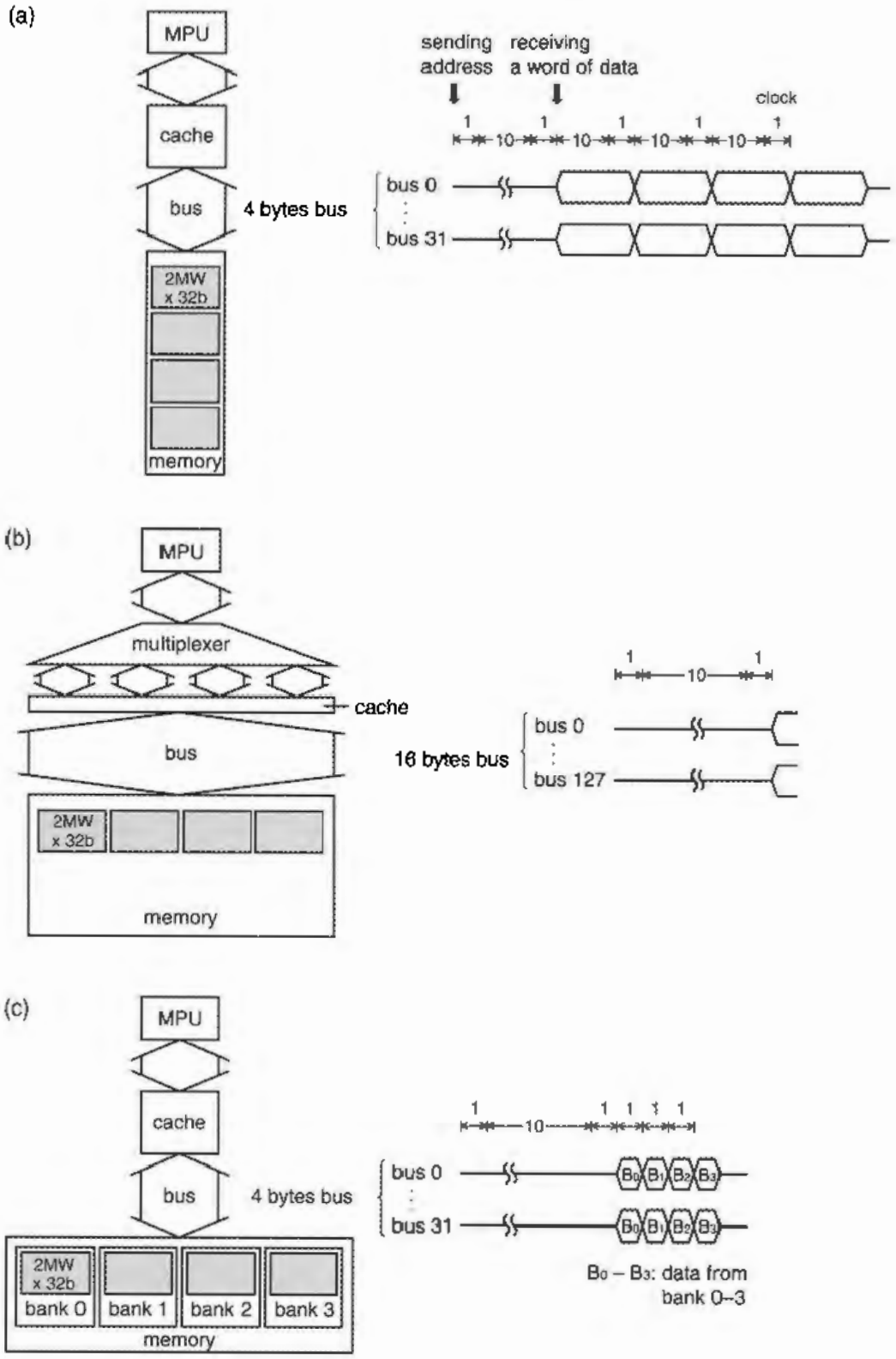


Fig. 6.7. One-word-wide (a), wide (b), and interleaved (c) memory organization [6.2, 6.3], with timing diagrams showing the number of clock cycles needed

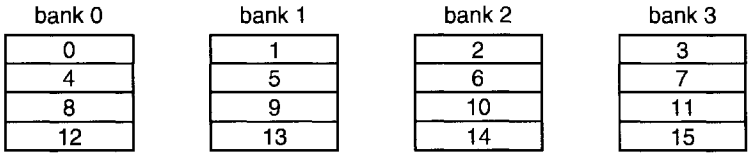


Fig. 6.8. A four-bank interleaved memory [6.3]. Word addresses from 0 to 15 are sequentially mapped from bank 0 to bank 3 so that four words in a block (for example, word addresses 0-3) are read sequentially

respectively. The miss penalty drops from 45 clock cycles to 12 clock cycles despite the drawbacks of increased cost due to the wider bus and additional buffers at the memory. Figure 6.7c shows the organization widening the bit width of the memory, but not the interconnection bus. Instead of making the entire path between the memory and cache wider, the memory chips can be organized in banks to read or write multiple words at a time, rather than a single word. Each bank could be one word wide, so that the width of the bus and the cache need not change, but sending addresses to several banks permits them all to be read simultaneously. This scheme, which is called *interleaving*, retains the advantage of incurring the full memory latency only once. For example, with four banks, the time to get a four-word block would consist of one cycle to transmit the address to the banks, ten cycles for all four banks to access memory, and four cycles to send the four words back to the cache. This yields a miss penalty of $1 + 1 \times 10 + 4 \times 1 = 15$ clock cycles. This is an effective throughput per miss of just over 1 byte per clock, or about three times the throughput for the one-word-wide memory and bus. The mapping of addresses to banks affects the behavior of the memory system. The example shown in Fig. 6.8 assumes that the addresses of the four banks are interleaved at the word level. A cache read miss is an ideal match to word-interleaved memory, as the words in a block are read sequentially. Banks are also valuable on writes. Each bank can write independently, quadrupling the write throughput. Interleaved memory banks need the wide-bit-data-I/O chip configuration as the memory capacity of the chip increases. Otherwise, the minimum number of memory chips necessary for the main memory and the add-on memory capacity is increased, which is not suitable for small systems. For example, for a 1 b I/O configuration the number is as large as 128, while for a 32 b I/O configuration it is reduced to four. The more detailed scheme of multibank interleaving will be explained later.

The column modes (the nibble, page, and static column modes, described in Chap. 3) of traditional DRAM designs are beneficial for memory system designs. The advantage of such modes is that they use the circuitry that is already on the DRAMs, adding little cost to the system while achieving an almost fourfold improvement in throughput, as will be explained. For example, the nibble mode was designed to take advantage of the same program behavior as interleaved memory. The chip reads 4 bits at a time internally, supplying 4 bits externally in the time of four column cycles. If

the nibble mode cycle is fast enough to be equal to the bus clock cycle, the interleaved memory organization in Fig. 6.7 could be realized with a high speed of nibble mode. Wide memory organization is not needed. Moreover, the minimum number of 64 Mb chips is one for a 32 b I/O configuration, while it is 32 for a 1 b I/O configuration, offering flexible system design to designers. The new breed of DRAMs such as DDR and Rambus DRAMs (which will be discussed later) further improve the throughput, although the cost penalty caused by the additional chip area is controversial.

Bus Utility. If the bus could transfer the data from memory at the maximum bus frequency, the peak throughput of the bus would be available. This is true as long as the throughput of the memory can be maximized by interleaving or high-speed column modes so as to keep up with the peak bus throughput, and an infinite length of burst data is transferred. In actual practice, however, it is not realized due to a degraded utility of the bus. The following irregular accessing is responsible for the degradation.

Despite increased utility, an excessive length of burst column data or an excessively large block size might allow unnecessary data to be transferred to the cache in vain. On the other hand, when MPU urgently requests data, it must wait until a string of data has been completely transferred. Thus, the burst length must be appropriate in order to permit frequent interruptions by row-address accesses. Furthermore, when the memory capacity of the cache is increased, the probability of row-address accesses is increased. This is because the cache is more likely to store the neighboring data and thus the randomness of access to the DRAM is increased. These row-address accesses degrade the bus utility because the following dead cycles are involved.

When DRAM is accessed by an address set of a row and column on a cache miss, a word line (a so-called page) is activated, and all the resulting cell data along the word line are amplified and latched on the data lines. These successive operations, from row-address activation to data latching, are simply called the opening of a page. Note that system designers map the address so that multiple words of a cache block can always be found on the same page. Therefore, when the column address corresponding to the first word of the block is selected, the corresponding latched data are outputted after latency, followed by the data resulting from successive selections of the remaining column addresses. However, when another cache miss with a different row address occurs, we cannot utilize the data previously latched on the data lines. Hence, we must turn off the selected word-line pulse and then precharge the data lines to prepare the next row access, which takes a long time, as discussed later. Thus, alternate accessing of different pages (i.e. word lines) in the same bank (i.e. the same memory array) interrupts the sequence of high-speed column modes. Here is a typical example of bus utility. 64-bit bus MPUs such as the Pentium complete a 32-byte data transfer with two clock cycles (i.e. with four bursts), if a double data rate (DDR) DRAM, which needs two clock cycles for each page opening, read latency (i.e. column-

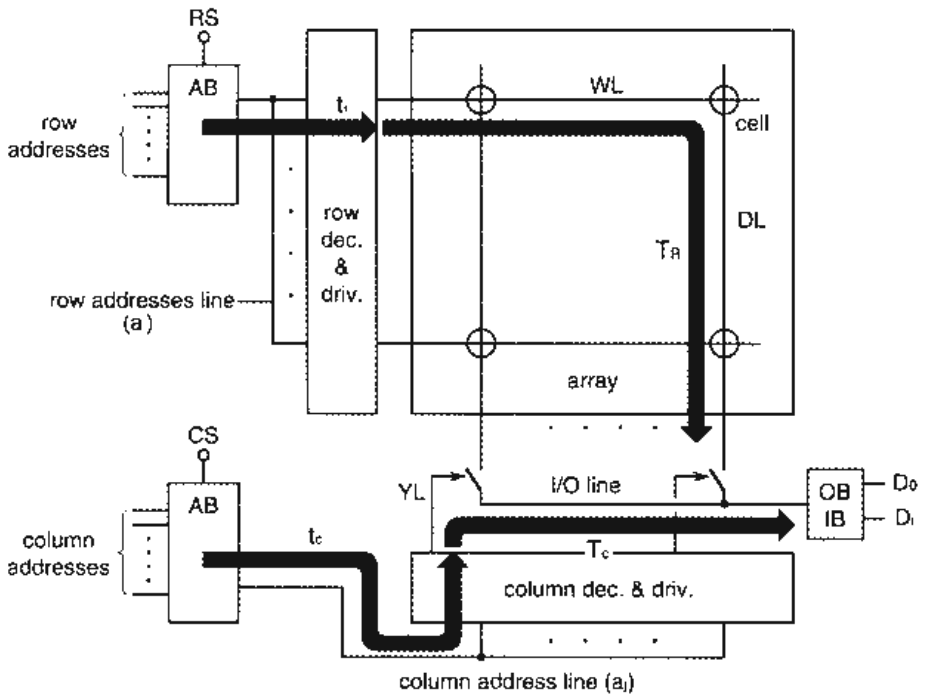


Fig. 6.9. A schematic of memory chip architecture. Only blocks in the critical path are conceptually illustrated. RS, row select; CS, column select

address activation to data output), and precharging operation, is used. Since the clock cycles for opening the page and precharging, all of which are related to row-address accessing, are dead cycles for data transfer, the data bus utility is only 25%. Dead cycles are also needed when mode transitions, such as the read-to-write operation and the selection of a different DRAM chip, occur.

6.2.3 Memory-Chip Performance

Memory-chip performance, which is represented by access and cycle times, has the greatest influence on the cache miss penalty. The performance is almost determined by a memory array that consists of large-RC delay components, by the current-driving capabilities of the circuits that are relevant to the memory array, and by the voltage swings and timing sequences necessary for memory-cell operations (Figs. 6.9 and 6.10, and Table 6.1). Here, the peripheral circuits such as the chip I/O and logic circuits are fast enough, as in logic chips.

The size of a memory array is usually large, and thus the word and data lines are heavily capacitive. Note that, for high-end embedded SRAMs, the area of the memory cell array is 50% or so of the total chip area. For stand-alone SRAMs, it may be 60–70% or so. Standard (commodity) DRAMs usually have array utilizations of 55–70%, while high-performance DRAM and protocol-intensive DRAMs such as Rambus are likely to have lower array

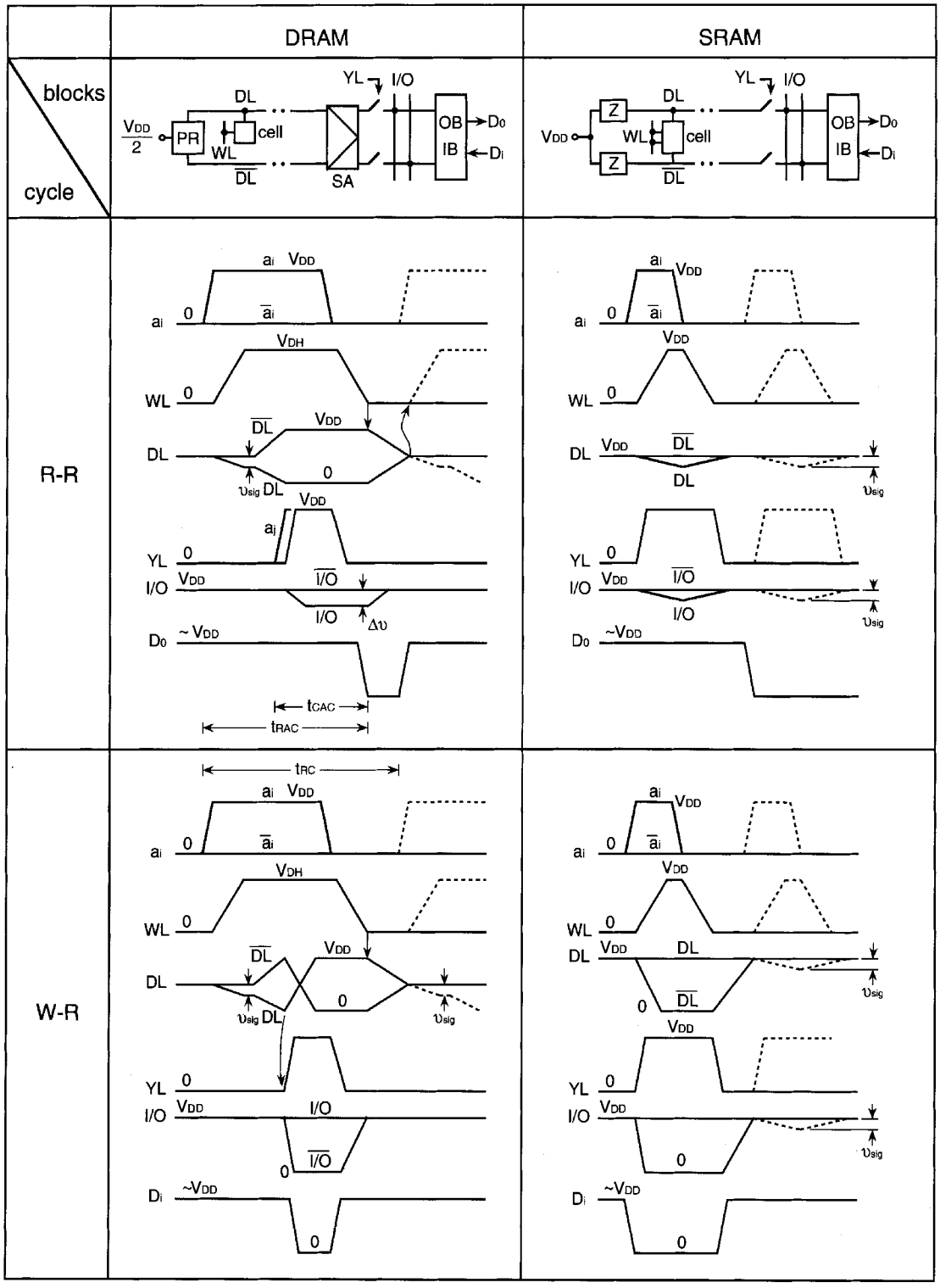


Fig. 6.10. A speed comparison between DRAMs and SRAMs

Table 6.1. Materials and voltage swings of major signal lines

Heavily capacitive lines	DRAM			SRAM		
	Material	Voltage swing		Material	Voltage swing	
		read	write		read	write
Row address line (a_i)	Metal	V_{DD}	V_{DD}	Metal	V_{DD}	V_{DD}
Word line (WL)	Poly-Si Polycide	V_{DH}	V_{DH}	Poly-Si Polycide	V_{DD}	V_{DD}
Data line (DL)	Polycide	$v_{sig}, V_{DD}/2$	V_{DD}	Polycide	v_{sig}	V_{DD}
Column address line (a_j)	Metal	V_{DD}	V_{DD}	Metal	V_{DD}	V_{DD}
Column select line (YL)	Metal	V_{DD}	V_{DD}	Metal	V_{DD}	V_{DD}
I/O line	Metal	$\Delta v(> v_{sig})$	V_{DD}	Metal	v_{sig}	V_{DD}

utilization. In addition, the lines are resistive, because they are formed using refractory metals such as poly-Si, which realize a high-density memory cell with the capabilities of self-aligned processing and fine patterning despite the resistive materials. Thus, a memory array inevitably suffers from large RC line delays. To reduce the line delays, the multidivision of a memory array with full usage of multilevel metal wiring and small-signal transmissions on the lines is useful, as explained later and in Chap. 3.

Row and column decoders/drivers, sense amplifiers, and precharge circuits also suffer from slow speeds caused by poor current-driving capabilities, because they must be small enough to be laid out within the tight pitches of the word and data lines. In particular, the DRAM sense amplifier on each data line is relatively slow because of the small input voltage of a cell signal of 100–200 mV that needs a long time for amplification.

In general, DRAM is slower than SRAM when the same physical size of memory array is assumed. The no-gain destructive read-out characteristics of the DRAM cell are responsible for the slow speed. Thus, SRAMs are suitable for cache applications, while DRAMs are used for main memory applications.

DRAM Speed.

DRAM Cell with no Gain. The absence of gain imposes the following strict timing sequence on DRAM operation. First, a small signal voltage is developed on the floating data line, which is susceptible to noise due to word-line activation. Next, the signal is amplified up to a sufficiently high voltage by a sense amplifier. Then, the amplified signal is outputted to the I/O line by activating a column selection line (YL). If the sequence is out of order, the sensing is unsuccessful due to a shortage of sensing signal voltage. For example, if the word line is activated while keeping YL turned on, the load

capacitance of the memory cell, which is the sum of the data-line and I/O-line capacitances, is increased, and thus the cell signal becomes so small that the sense amplifier cannot amplify the signal correctly. The fact that the cell has no gain is responsible for the small signal. If YL is activated before the signal voltage on the data line is amplified sufficiently, the signal voltage transmitted to the I/O line is so small that the main amplifier on the I/O line fails to discriminate the cell information correctly. This is due to the heavy I/O-line capacitance, compared with the data-line capacitance. This timing sequence, which is needed in the conventional design, is a major reason why the row access time of a DRAM is inherently slow. However, if the direct sensing (see Fig. 3.63) is applied to the DRAM, despite an area penalty, a small cell signal can be quickly transmitted to the I/O line without waiting for amplification by the sense amplifier, thus realizing a faster row access time.

The Destructive Read-Out of the DRAM Cell. The row cycle time of the DRAM is also slow due to the destructive read-out characteristics of the DRAM cell, which necessitate an additional rewrite operation with a large voltage swing of V_{DD}, as exemplified by the R-R cycle in Fig. 6.10, that stands for a read followed by another read. In this case, the word pulse could be turned off just after the data-line voltages were fully amplified to V_{DD} and 0 V for a full rewrite operation, enabling the fastest possible cycle time. However, the cycle time, becomes longer in the case of the W-R cycle in Fig.6.10, which is for a write followed by a read. In particular, our concern is about when data that differs from the cell stored data will be written, because this is the slowest condition. Here, the write data voltage must be inputted from the I/O line to the data line later than the YL activation for the read operation. Otherwise, an earlier YL activation and thus an earlier large data-voltage application to the data line, could couple noise to the adjacent data lines on which signal voltages are still small, causing unstable operations. Obviously, it is difficult for the data-input buffer (IB) to quickly replace an old data voltage latched at the sense amplifier with a new data of the opposite polarity data voltage on the large-RC data line, even with the help of the sense amplifier. In any event, after the data lines have reached V_{DD} or 0 V, the word pulse can be turned off, and then a large differential voltage of V_{DD} is equalized to a half-V_{DD} by a precharged circuit (PR). The equalizing (or precharging) time is also long, because the precharge circuit must continue to be driven until the differential voltage becomes negligible compared with the signal voltage (v_{sig}) of the next read operation. Note that, in practice, the word pulse width of the read operation is the same as that of the write operation.

The row cycle time of a DRAM is always slower than the row access time. The row access time (Fig. 6.9), t_{RAC}, is expressed by

$$t_{RAC} \simeq t_r + T_R + T_C \,, \tag{6.1}$$

where t_r is the delay time from the row address buffer (AB) input to the row (WL) driver output, T_R is the delay time from the word driver output

to latching of the cell read signal to each data line, and T_C is the delay time from column (YL) activation to data output (Do) through the data output buffer (OB). The minimum row W-R cycle time (t_{RC}) is obtained as

$$t_{RC} \simeq t_r + T_R + t_{op} + t_{eq} , \tag{6.2}$$

where t_{op} is the data-replacing time described above, and t_{eq} is the equalizing time of a pair of data lines. Here, T_C is fast enough, compared with $t_{op} + t_{eq}$. This is because a main amplifier on the I/O line and the succeeding logic circuits in the output buffer can be laid out with large sizes so as to provide large drive currents, and their loads are less capacitive and resistive. Thus, the relationship of $t_{RAC} < t_{RC}$ is satisfied.

The column speeds are fast enough compared with the row speeds because large-delay components relevant to the memory array, such as the word line, the data line and the memory cell, are eliminated in this path. Thus, once all the read data of the memory cells along the selected word line are latched at the corresponding data lines, they can be randomly read out to the I/O lines, and then to the data output, at a fast access and cycle time by random selection of column addresses. The high-speed features of column selection can be explained as follows. The column access time (t_{CAC}) and the column cycle time (t_{CC}) are expressed by

$$t_{CAC} \simeq t_c + T_C , \tag{6.3}$$

$$t_{CC} \simeq t_c + T_C + t'_{eq} , \tag{6.4}$$

where t_c is the delay time from the column address buffer input to YL activation, and t'_{eq} is the equalization time of the I/O line.

The ratios of the row speeds to the column speeds are expressed by

$$\frac{t_{RAC}}{t_{CAC}} \simeq 1 + \frac{1+x}{1+(t_c/T_C)} , \tag{6.5}$$

$$\frac{t_{RC}}{t_{CC}} \simeq 1 + \frac{x+(t_{eq}/T_C)}{1+(t_c+t_{eq})/T_C} , \tag{6.6}$$

assuming that $t_r \simeq t_c$, $t_{eq} \simeq t'_{eq} \simeq t_{op}$ and $T_R/T_C = 1 + x$. For example, if $x = 1$ and $T_C = t_c = t_{eq}$, as in a traditional page-made design, the following expressions are derived:

$$t_{RAC}/t_{CAC} \simeq 2 , \tag{6.7}$$

$$t_{RC}/t_{CC} \simeq 1.7 . \tag{6.8}$$

Many column modes that utilize this high-speed feature have contributed to improvements to system performance, as discussed in Chap. 3 and in the following section.

SRAM Speed. Due to the possession of a gain, the SRAM cell allows its small signal voltage to be directly transmitted to the main amplifier on the I/O line without imposing any restriction on YL activation timing. Moreover, the non-destructive read-out characteristics of SRAM cell allow write data to be inputted to the selected data line at the earliest timing, eliminating the rewrite operation that is necessary for DRAM. Even so, the resultant high-voltage swing on the selected data line does not destroy small read voltages on the adjacent data lines, despite capacitive coupling, since the small voltages are static ones and thus are immune to various noises, unlike the floating ones of the DRAM. Thus, the necessary word pulse can be shortened, enabling fast row access and cycle times. If the sum of the data-replacing time involved in a write operation and the equalizing time on the data lines is equal to the sum of the column delay (T_C) and the equalizing time on the I/O lines, as in usual SRAM designs, the row access time is equal to the row cycle time. The column speeds are faster than the row speed, as in the DRAM. If multi-stage pipelining is used, even the cycle time faster than the access time is achievable [6.11].

6.3 Memory-Subsystem Technologies

It is obvious that wide-bit I/O chip configurations combined with parallel operation of multidivided arrays increases the throughput. Moreover, to improve the traditional DRAM performance discussed in Chap. 3 modern DRAMs incorporate memory-subsystem technologies such as multibank interleaving, synchronous operation with a latch function, pipeline/prefetch operations, high-speed clocking schemes, and terminated interfaces combined with high-density packaging. These technologies are supported by command operations, on-chip mode registers, and packet protocols.

6.3.1 Wide-Bit I/O Chip Configurations

These configurations [6.4, 6.17] offer high throughput as well as ease of use, which is realized by reducing the chip count needed by the system and by adding flexible add-on memory capability. Despite the possibility of at least 256 b organization, compared to the 32 b organization for current experimental 1 Gb chips [6.4, 6.17], the number of I/O pins is eventually restricted by the following drawbacks: the chip power increases rapidly with an increase in the pin count, because the number of simultaneously charged and discharged DLs (i.e. m in the logical array) increases. The chip area also increases due to an increase in the I/O relevant circuits. The details are given in Chap. 7.

6.3.2 Parallel Operation of Multidivided Arrays

The concept of the multidivided array [6.4, 6.17], a combination of multidivided data lines (DLs) and multidivided word lines (WLs), shown in Fig. 6.11,

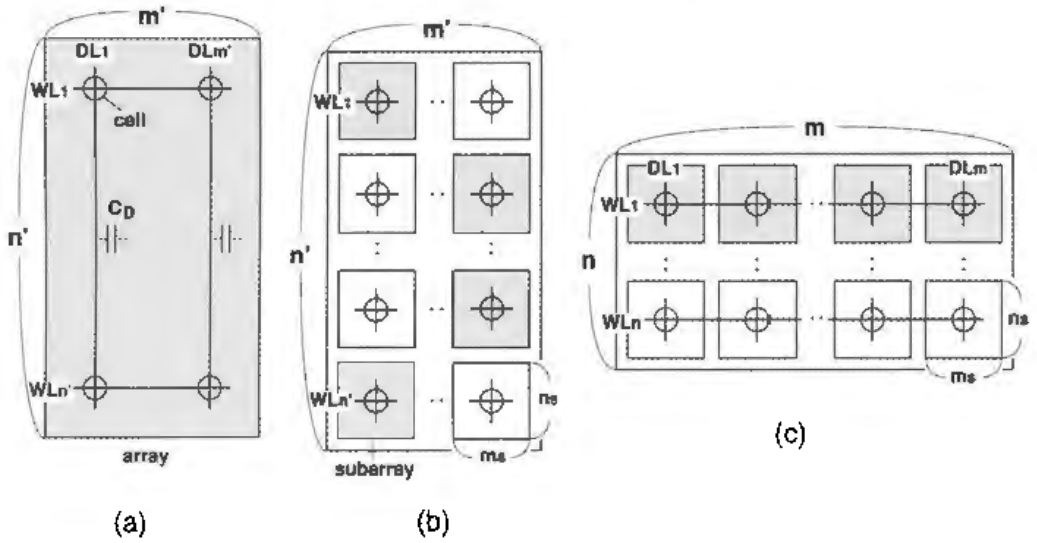


Fig. 6.11. The concept of a multidivided RAM array [6.14, 6.17]: the shading denotes the activated area. $M = n'm' = nm$. (**a**) A non-divided array; (**b**) a multidivided array; (**c**) a logical array

is the key to designing a high-performance RAM. The division of a DL and its partial activation dramatically increases the inherently small signal voltage and reduces the DL power dissipation that dominates the total chip power. The division of a WL is also essential to improve the ever-increasing WL delay with increasing memory capacity. A multidivided array realizes the high performance of a resulting subarray, if low-resistivity multilevel metal wiring and high-speed subarray-selection circuits are adopted. Multilevel metal wiring also minimizes the additional increases in area at the divisions. Any combination of a number of subarrays could be simultaneously activated, since each subarray could be randomly accessed. The parallel operation capability of subarrays enables multibank interleaving, if each bank in a memory system is asserted by each subarray. However, for DRAMs the number of simultaneously activated subarrays is restricted by the DL power dissipation and the maximum refresh time (the data-retention time of the cell), t_{REFmax}, which is specified in the catalog for the chip. The activation for the complicated physical array is simplified by using the logical array comprising n virtual word lines shown in Fig. 6.11c. Here, n is the number of refresh cycles in the catalog specification, which are usually distributed within t_{REFmax}; and m is the number of simultaneously activated DLs, taking m_{s} to M/n_{s}, where M, n_{s}, and m_{s} are the memory capacity of the chip, the number of sub-WLs, and the number of sub-DLs in a subarray, respectively. Here, n_{s} is less than 1k in the megabit era, which is determined by the minimum signal voltage for a successful sensing, while m_{s} is 256 or 512 in terms of the WL delay. Figure 6.12a shows a more detailed multidivided array, widely used in multimegabit DRAM products, featuring the DL orthogonally aligned to the internal I/O line. In an actual design, in addition to the shared Y decoders

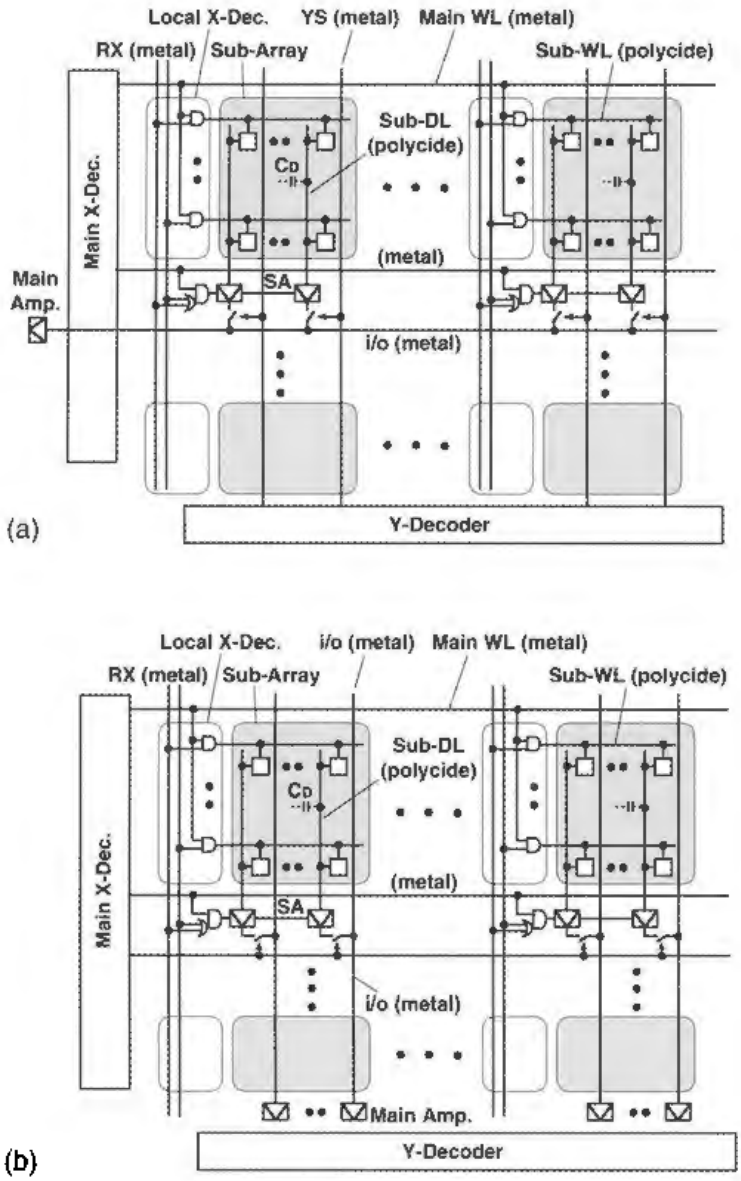


Fig. 6.12. An actual multidivided DRAM array [6.4, 6.17]: the paired-line arrangement is actually applied for each DL and each i/o line. (a) DL orthogonal to i/o line; (b) DL parallel to i/o line

shown in the figure, a combination of a shared sense amplifier (SA) and a shared I/O is used to further reduce the DL charging capacitance, as well as the chip area. The details are discussed in Chap. 3.

The DRAM array is a built-in structure capable of a massively parallel operation at the expense of large DL power. This stems from the refresh operation requirement of the 1 T cell, that needs simultaneous activation of all the cells along the selected WL. A multidivided array further increases the available number of data, with at least one unit of data from each subarray.

If the I/O line is arranged in parallel to the corresponding DL, as shown in Fig. 6.12b, the available number is maximized by one unit of data from each DL. Thus, the DRAM array inherently favors embedded DRAM designs, in which high throughput is the first priority, although increases in the DL and I/O power become serious concerns. To achieve high throughput of multidivided arrays while reducing the area, a large logic-gate block in an embedded DRAM chip will require additional layers of metal wiring.

6.3.3 Multibank Interleaving

Multibank interleaving has been widely used to increase the throughput with a substantially parallel operation of the multibank. In this scheme, a memory system consists of N banks, which are sequentially addressed from bank 1 to bank N, as shown in Fig. 6.13. Each bank is composed of a memory module using many DRAM chips, so that the memory bus has an N_{MB}-bit data-bus (i.e. I/O) width. When N words, each of which comprises N_{MB} bits, are

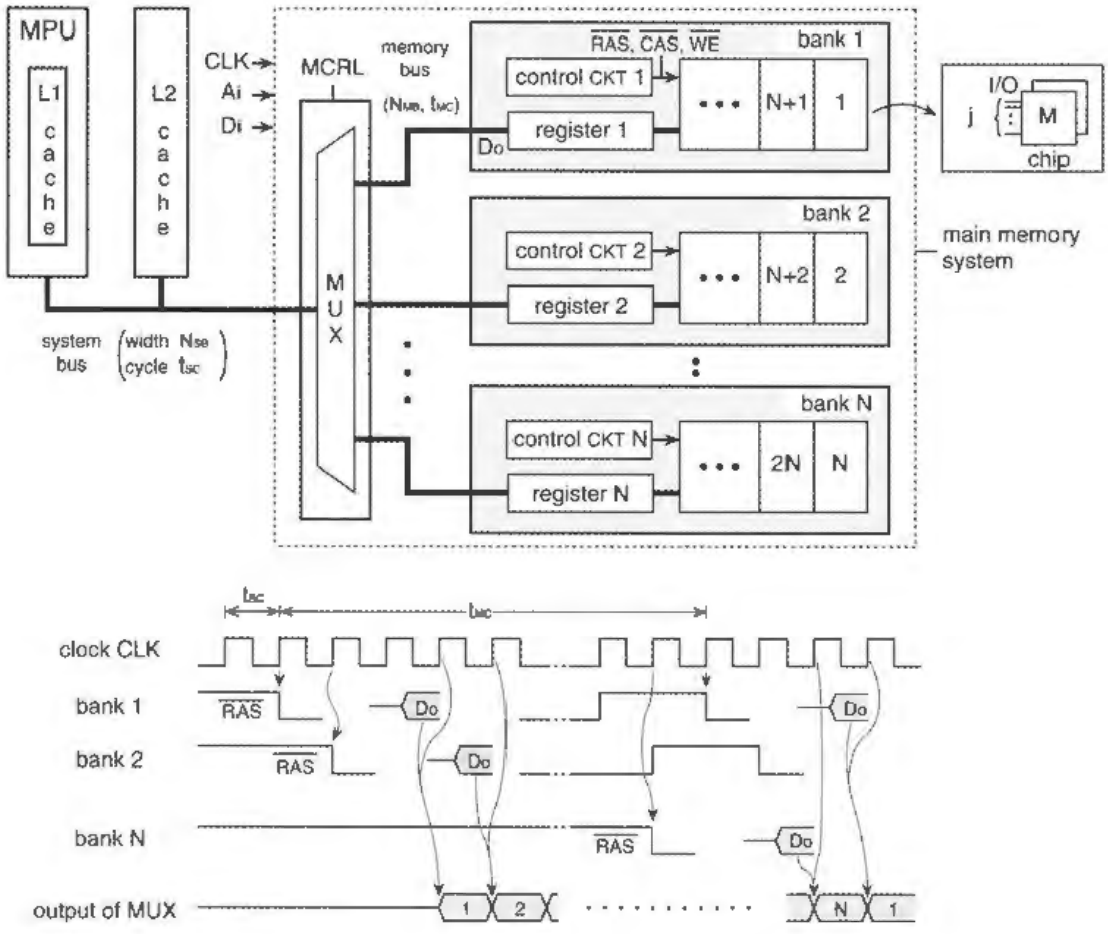


Fig. 6.13. Conventional multibank interleaving [6.8]. MUX, multiplexer; MCRL, memory controller

successively read, the resulting N_{MB}-bit data are outputted on the memory bus of each bank, and stored in each register so as to be able to wait for the next accessing. As a result of multiplexing of data from each bank, N_{MB}-bit data are available every system-clock cycle t_{sc} on the system bus. Since it continues for the memory cycle time (t_{MC}) of bank, the throughput is increased from N_{MB}/t_{MC} to $N_{MB}N/t_{MC}$, with an N-fold increase. However, this approach causes an increase in the minimum add-on memory capacity, which is expressed as $MN_{MB}N/j$, where M and j are the memory capacity of the chip and the I/O pin count of chip, respectively. Moreover, it enables an increase in the number of bus lines and relevant devices with $N_{MB}N$, preventing flexible design, miniaturization, and a low cost for the whole system. Increasing j is beneficial to increasing the throughput for a fixed add-on memory capacity, or to reducing the minimum add-on memory capacity for a fixed throughput. However, an excessively large j causes increased power dissipation, chip area, and package size, as discussed in Chap. 7. It would also degrade the chip speed, with increased noise at the I/O pins.

The multibank interleaving that memory-system designers have taken for granted can be implemented on one chip, if the multidivided array structure is utilized. This is because the subarray shown in Fig. 6.11 can be regarded as a bank, if each subarray equips its own peripheral circuit and each address buffer of the bank, and thus it can latch the input address signal so that different banks can be successively selected at the minimum system clock cycle. When a certain bank is selected, the corresponding address signals are latched at the address buffers of the bank, so that address input lines are ready for the next addressing for a different bank. While the memory operation of the succeeding circuits in the bank proceeds with the latched address signals, the different bank is selected with the different addresses. Thus, the MPU can successively access different banks without having to wait by the memory cycle of the bank. Moreover, while a bank is accessed, other banks could perform precharge or refresh operations, enabling these inherent DRAM operations to be hidden.

6.3.4 Synchronous Operation

In this scheme, all of the input/output signals of chip are latched at the I/O interface circuits of the chip, synchronously with the system clock. This scheme not only allows chip designers to incorporate various high-speed functions, but it also allows system designers to improve the system speed, with an easier timing on the board.

Figures 6.14 and 6.15 show the concepts behind asynchronous- and synchronous operation [6.8], assuming a non-divided data line. Synchronous operation has been used for modern DRAMs, such as the synchronous DRAM (SDRAM), The double-data-rate (DDR) DRAM, and the Rambus DRAM, while asynchronous operation has been used for traditional DRAMs, such as the page mode, nibble mode, static column mode, and extended-data-out

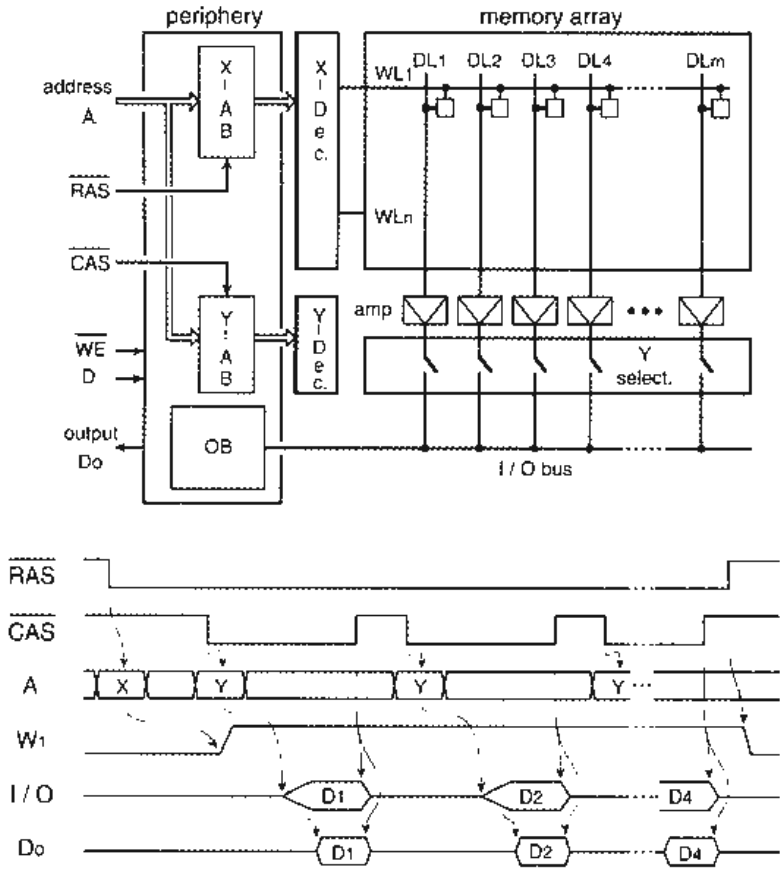


Fig. 6.14. The asynchronous operation of a DRAM chip [6.8]. AB, address buffers; OB, data-output buffer; X, row; Y, column

(EDO) DRAMs. In synchronous operation, a row address strobe signal RS1 is generated from the chip-select signal $\overline{CS}$ at the rising edge of the system clock CLK, so that row addresses are strobed and the corresponding word line (for example, WL_1) is activated. The resulting cell signals on m data lines are amplified in the usual manner. On the other hand, a column address strobe signal CS1 is generated by the next CLK, so that the column addresses are strobed and the amplified signal on the corresponding data line (for example, DL_1) is outputted on the common I/O bus line. Then, the data output Do is available from the chip synchronously with the succeeding CLK. The synchronous operation widens the timing margins between the internal control signals. Easier timing designs due to the use of simple latch circuits that work synchronously with the system clock are responsible to the wider margins. On the other hand, in asynchronous operation the internal timing designs are complicated, because they are closely related to many sophisticated set-up/hold timing specifications between external input/output signals. Moreover, the latch function achieves high throughput with the resultant pipeline operation (discussed later) and multibank interleaving. The

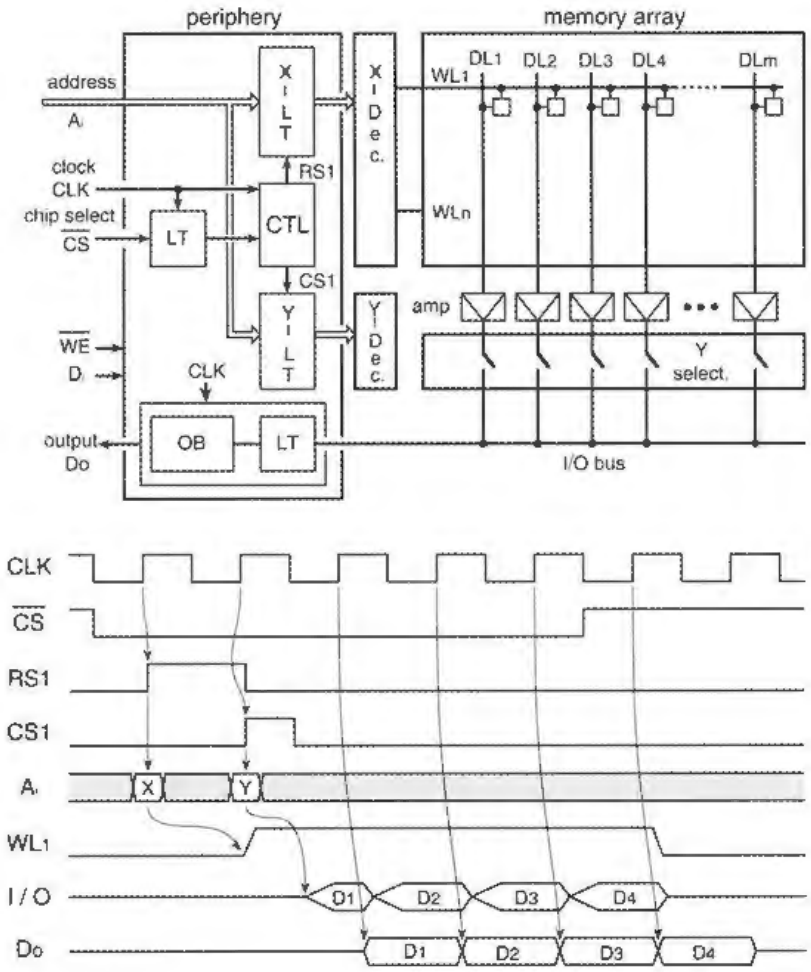


Fig. 6.15. The synchronous operation of a DRAM chip [6.8]. LT, latch; CTL, control circuit; OB, data-output buffer

synchronous operation can eliminate the $\overline{\text{RAS}}$ and $\overline{\text{CAS}}$ functions that are familiar in asynchronous operation, since the same functions are carried out by the internal signals RS1 and CS1. The operation also eliminates long hold-signals such as $\overline{\text{CS}}$ in Fig. 6.15 if the command instruction scheme is adopted. Various operation modes are set by the combination of command signals whose pulse widths are almost equal to that of the system clock, as seen in SDRAMs (see Fig. 6.32).

The synchronous operation of DRAM chips also increases the system speed. The asynchronous operation in Fig. 6.16, in which many control signals are generated at a timing generator (similar to control circuits in Fig. 6.13) by using the system clock, makes it complicated to synchronize address signals with the control signals and/or the system clock. For example, although various signal skews exist as a result of running on the memory board of the bank, the minimum timing specifications of the DRAM chip regarding the row-address set-up time t_{ASR}, the row-address hold time t_{RAH}, the column-

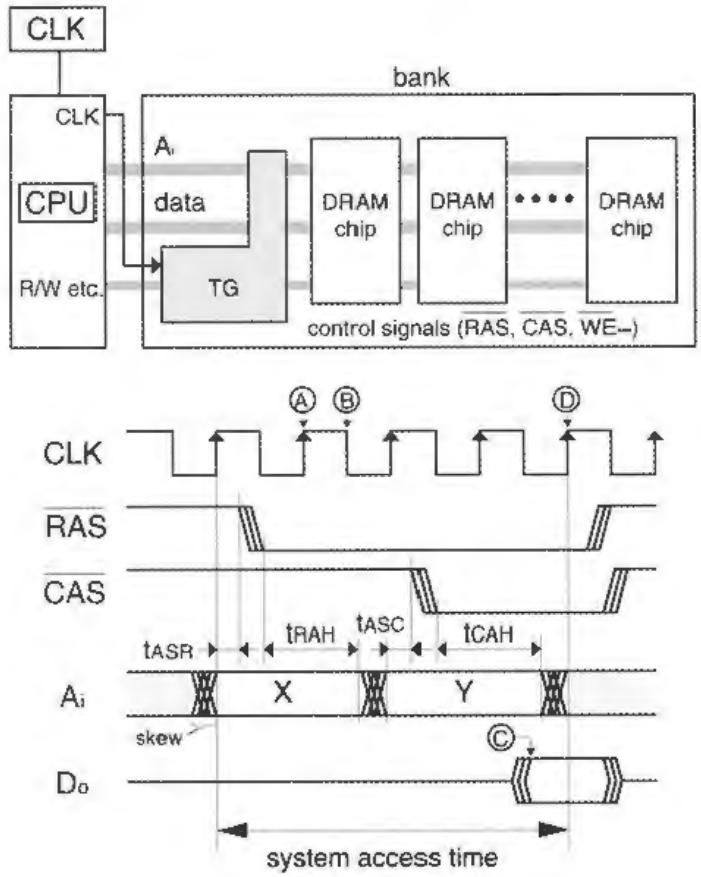


Fig. 6.16. The asynchronous operation of the memory system [6.8]. TG, timing generator; X, row; Y, column

address set-up time t_{ASC}, and the column-address hold time t_{CAH} must be ensured. Obviously, four kinds of skews are added to the minimum cycle time achieved by the chip itself, degrading the memory-bank cycle time. A timing mismatch between the control signals and the system clock further degrades the cycle time. For example, even if t_{RAH} exists between timings Ⓐ and Ⓑ of the clock and thus addresses could be switched to the column ones, the address switching must take place at the falling edge Ⓑ. In addition, even if the chip can output the data at timing Ⓒ, the bank actually outputs it at timing Ⓓ. Such redundancies in timing fail to fully bring out the high speed of the chip. On the contrary, the synchronous operation shown in Fig. 6.17 allows the memory access time and cycle time to be expressed as an integer multiple of the system clock cycle, and all the relevant signals to be synchronized with only one system clock. Thus, it is enough for system designers to pay attention only to the set-up and hold time for one clock, without paying any attention to the phase relations between various signals. System designers are thus released from a complicated timing design caused

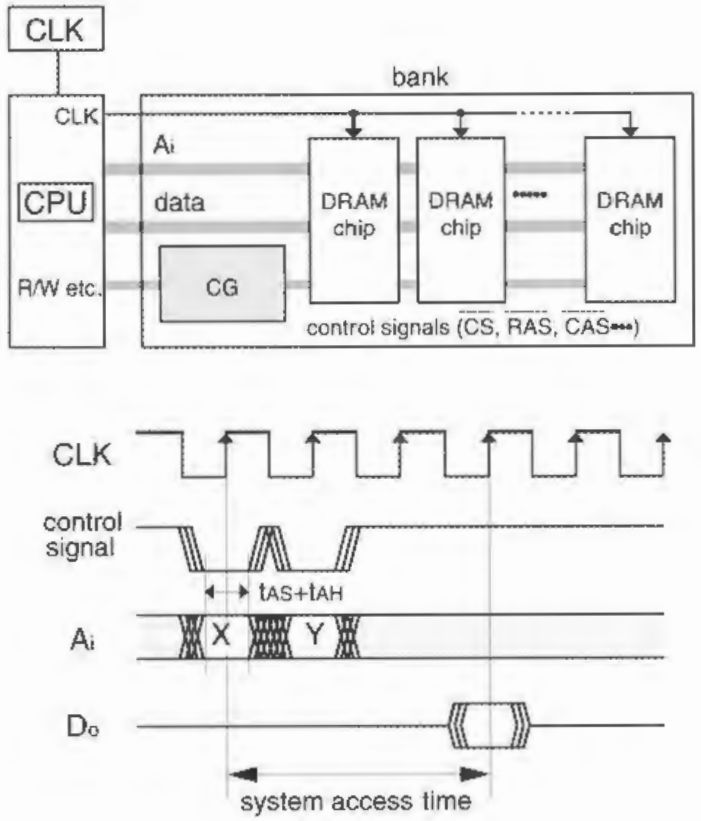


Fig. 6.17. The synchronous operation of the memory system [6.8]. CG, command generator

by asynchronous memory operation, and they can design the memory system just as they design a conventional logic system.

6.3.5 Pipeline/Prefetch Operations

A pipeline operation, which features the parallel operation of a number of circuit blocks combined with latch circuits, has been proposed to increase the throughput of an asynchronous operation DRAM [6.12]. The combination of a pipeline operation and synchronous operation, however, further enhances the throughput. An application of pipeline operation with a latch circuit to the I/O line, as shown in Fig. 6.15, is especially important. It shortens the cycle time of the I/O relevant circuit, which is usually slow in speed because of its inherently heavy capacitance. If the latch circuit (LT) on the common I/O bus line latches the read data from a memory array, and then isolates the latched data from the I/O line, the I/O line can be used for other operations, such as precharge and equalizing operations, while the latched data is processed by the succeeding circuit. If the column-selection circuits (Y select. in the figure) have shift-register and ring-counter functions, a string of m data is sequentially output on the I/O line by inputting only the first addresses to the

column decoder. This operating mode is called the burst mode. In practice, m-bit data are divided into a number of units, each of which consists of a few bits and is independently operable. Thus, any bit within any unit can be accessed, followed by the sequential data of the remaining bits of the unit. The throughput of the burst mode is higher than that of the page mode or the nibble mode discussed in Chap. 3. This is because the page mode, for example, needs not only timing margins between the external clock $\overline{\text{CAS}}$ and addresses (as shown in Fig. 6.14), but also I/O-line precharging. Note that the maximum number of data bits available with one burst mode (i.e. one word-line selection) is m. When a different word line is successively selected, the next burst mode is interrupted by data-line precharging.

Pipeline and/or parallel operations, combined with synchronous operation using the external system clock, offer a faster column cycle. For example, a prefetch scheme doubles the throughput, if two bits from memory array are simultaneously fetched to the data output buffer, and are outputted through a high-speed, parallel-to-serial operation at the buffer.

6.3.6 High-Speed Clocking Schemes

The high throughput of modern DRAMs is supported by high-speed clocking schemes such as command operations, on-chip mode registers, and packet protocols. Command operations allow all of the operating modes to be designated by combinations of two logic levels of command signals, whose pulse widths are almost equal to that of the system clock, as will be explained with an SDRAM. The resulting simplified timing design enables the high speed. On-chip mode registers defines how the DRAM chip operates. For instance, they define column latency and burst length, as explained later. This information is programmed into the registers by users, so that the chip operates based on this information. Packet protocols feature a request packet of addresses, data, and controls transmitted on the system bus, as will be discussed below for a Rambus DRAM.

6.3.7 Terminated I/O Interfaces

Small-swing impedance-matched (terminated) interfaces provide high-speed performance with the reduction of the high switching I/O power, especially in the high-speed region, of the TTL or LVTTL interfaces. Rambus, SSTL (Stub Series Terminated Logic), and HSTL (High-Speed Transceiver Logic) interfaces are good examples. Protocol architectures such as Rambus and SyncLink, which have an extremely high clocking scheme supported by high-density packaging, partly solve the problems caused by the simply widened bus architecture. The interfaces will be explained in more detail below.

6.3.8 High-Density Packaging

The rapid and remarkable progress in high-density chip packaging can be summarized into two categories: achievements involving better physical characteristics (thinner, smaller, and lighter) and those involving a higher pin count. As a result, the LOC (Lead-On-Chip) package (Fig. 2.46) can accommodate ever-enlarged chips in a small package. The ball grid array (BGA) and the chip-size package maximize the pin count while reducing the lead-frame inductance and the I/O loading capacitance, which are essential to limit the ever-increasing area, power, and noise of the chip. Multichip module packaging (MCM) further improves the system performance. The data transfer rate of standard DRAM chips dramatically increases with low power, if high-density packaging that keeps up with high-density chip technology, and low-voltage operation, are applied, as shown in Fig. 6.18. The I/O wiring on the module is short enough to eliminate the termination with reduced capacitance. MCM packaging will make standard DRAMs competitive with embedded DRAMs in the gigabit era, while keeping the low-cost advantage of standard DRAMs.

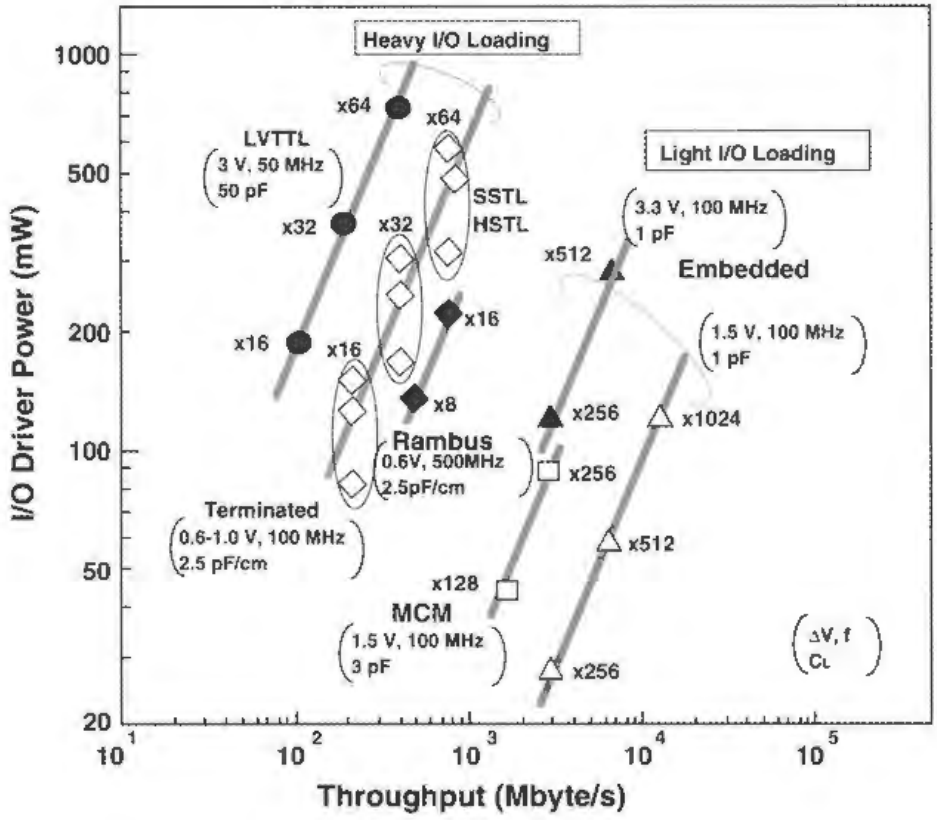


Fig. 6.18. The data transfer rate of a DRAM chip [6.4, 6.17]. x_j, The I/O pin count of a DRAM chip; C_L, the loading capacitance of each I/O line, without termination power

6.4 High-Performance Standard DRAMs

6.4.1 Trends in Chip Development

DRAM chips are now capable of 1.6 Gbyte/s operation [6.15, 6.16] due to the memory-subsystem technologies discussed previously. Synchronous operation with a latch function solves the slow-speed problem of a large memory array. A latch function at the row address buffers controlled by the external clock shortens the row access/cycle times with multibank interleaving, in which the next row selection for another independent bank (i.e. a subarray) can be accepted while processing of the selected bank is in progress. The successive operation of different banks even enables consecutive burst read/write operations with mixed data, while hiding row operations, when high-speed column selections are combined. A huge number of read data latched at many sense amplifiers on data lines can be quickly processed by using the following column selection schemes (Fig. 6.19). A two-stage pipeline applied to the EDO (Extended Data-Out) DRAM [6.17], despite asynchronous operation, extends the data-valid window while reducing the CAS cycle time (t_{HPC}) to a time as short as the array delay (t_D) by eliminating the data-output delay. The data-output latch (DLT) controlled by the timing generator (TG) isolates the output buffer (OB) from the array after $\overline{CAS}$ returns as high (Fig. 6.19a). Hence, the address latch (ALT) can accept the next column selection and the address transition detector (ATD) can begin the column operation after stopping the I/O precharge, while the latched data continue to be output. A wave pipeline applied to the SDRAM (Synchronous DRAM) [6.18, 6.19] further improves the speed, enabling a clock cycle time (t_{CK}) shorter than t_D, which includes the transition time of long lines such as the column address line (AY), the column select line (YS), and I/O line (Fig. 6.19b). The address counter (AC) captures the start address at a read command and generates successive column addresses every cycle according to the burst length (BL) designated in the mode register. To ensure a wide range of operable t_{CK}, the first-in-first-out register (FIFO) controls the column latency (CL). Furthermore, a two-bit prefetch, as in the DDR (Double Data-Rate) SDRAM [6.15, 6.20, 6.21], halves the data cycle time with parallel-to-serial data conversion at a multiplexer (MUX), by aligning data at both edges of the clock (Fig. 6.19c). To eliminate the clock-data delay (i.e. $t_{AC} = 0$), the data-out timing is defined by a clock recovery circuit, such as a DLL (Delay-Locked Loop). Also, the data strobe (DQS) output with the same timing as the data output helps the controller capture data in a small valid window. The power consumption can be lowered by inactivating the clock recovery circuit in the stand-by mode, although a few lock cycles are needed at the transition between the active and stand-by modes [6.19].

The SLDRAM (Synchronous-Link DRAM) [6.22, 6.24] and the D-RDRAM (Direct Rambus DRAM) [6.16, 6.25] operate at extremely high clock frequen-

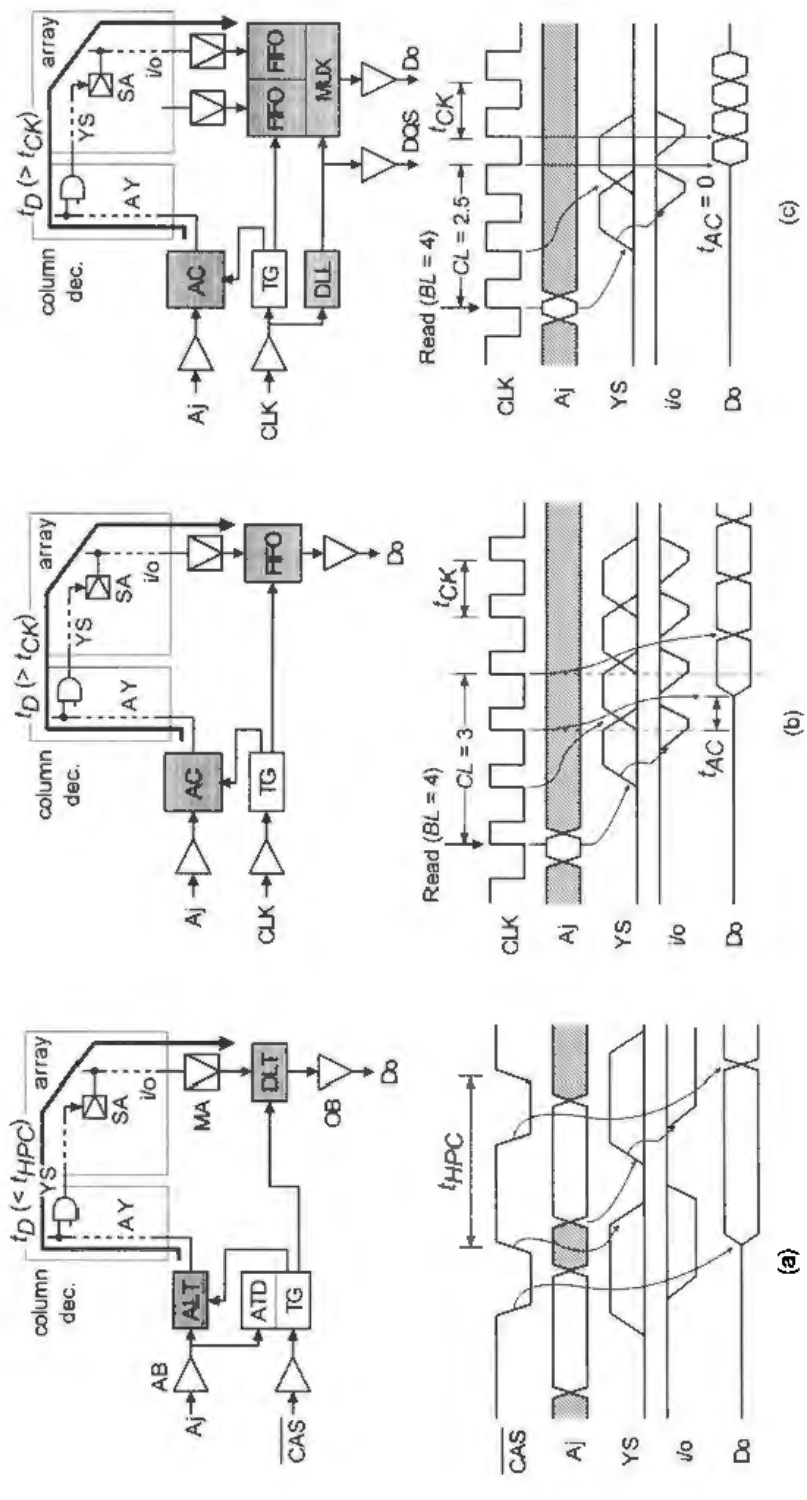


Fig. 6.19. Concepts of high-throughput column architectures [6.14]. (**a**) EDO DRAM, using a two-stage pipeline; (b) SDRAM, using a wave pipeline; (c) DDR DRAM, using a wave pipeline and prefetching

cies. In addition to having the relevant DDR circuits (described above) and terminated interfaces (discussed later), they effectively cancel the bus flight time, which is comparable to the clock cycle time, by using packet protocols and chip-by-chip timing adjustment. The packet protocols ensure a uniform delay for commands and addresses. Variation in the data arrival time from each DRAM chip, which depends on the chip location, is eliminated by auto-calibration of the clock access time of each DRAM chip or the loop-around clock.

Terminated I/O interfaces [6.26] are also needed to enable operation over 100 MHz, which is increasingly difficult to achieve with TTL or LVTTL (Low-Voltage TTL) due to the signal distortion caused by impedance mismatching. SDRAMs use an SSTL (Stub Series Terminated Logic) interface that has a small signal swing of 0.7 V, compared to 2 V with LVTTL, and series terminations (Rs) at stubs and parallel terminations (Rt) at both ends of a bus,

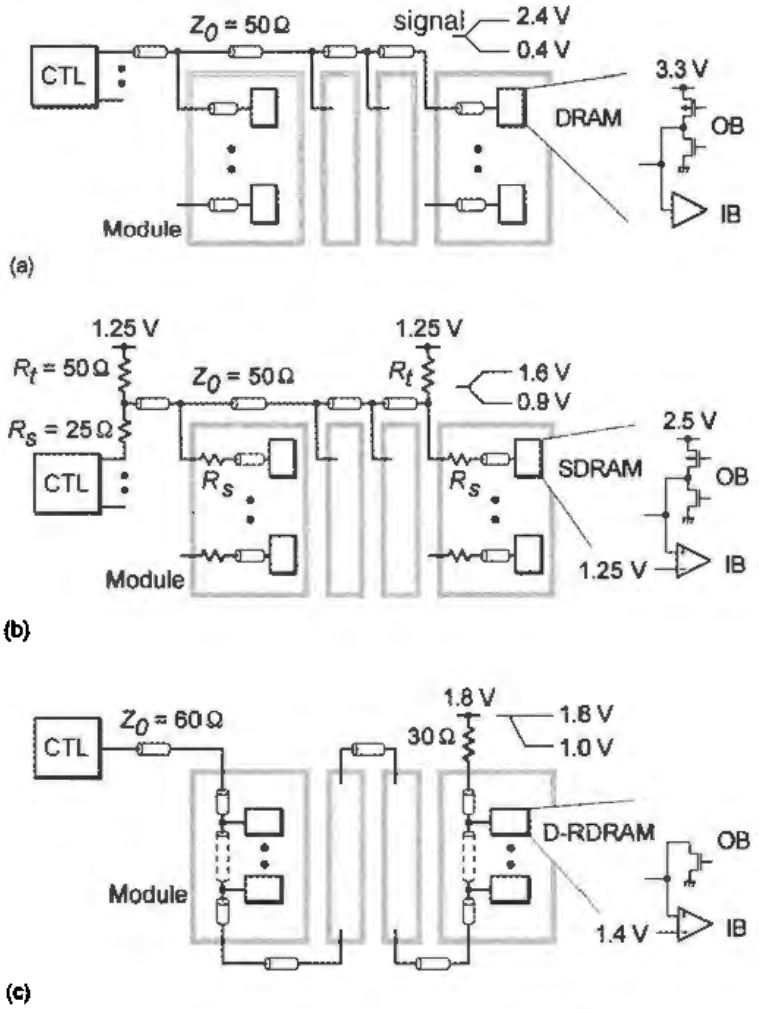


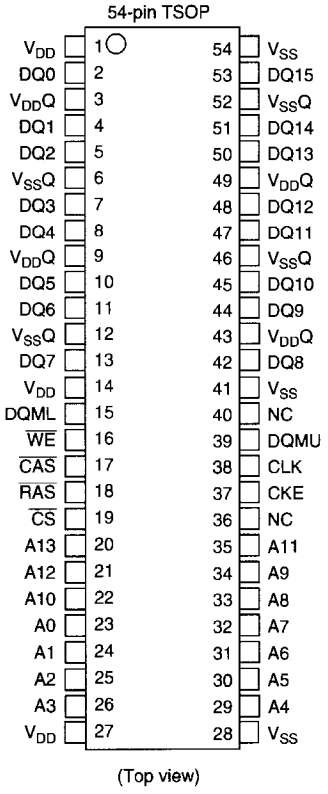
Fig. 6.20. Data buses of typical DRAMs [6.14]. (**a**) LVTTL; (**b**) SSTL for 2.5 V (SSTL_2); (**c**) Rambus channel

as shown in Fig. 6.20. The chip interface uses push-pull output drivers (OB) and differential input buffers (IB) with a 1.25 V reference voltage. While the SLDRAM uses the SSTL-type bus, the D-RDRAM uses a Rambus channel which features a stubless bus and a single termination at the opposite end of the memory controller (CTL). The channel necessitates an open-drain output buffer for the DRAM and high-density packaging.

Table 6.2 compares the various high-throughput DRAMs.

6.4.2 Synchronous DRAM

Synchronous DRAM (SDRAM) has been widely accepted by many system designers. In this section there is a brief explanation about a commercial SDRAM with 1 Mword × 16 bit × four-bank organization [6.27], in which four banks can operate simultaneously and independently. The chip is accommodated in a 400 mil (1 mil = 0.001 inch) 54-pin plastic TSOP (Thin Small-Outline Package). The pin arrangement and pin functions, and the block diagram and characteristics of the chip, are shown in Figs. 6.21 and 6.22, and Table 6.3, respectively.



Pin name	Function
A0 to A13	Address input — Row address A0 to A11 — Column address A0 to A7 — Bank select address A12/A13 (BS)
DQ0 to DQ15	Data-input/output
$\overline{CS}$	Chip select
$\overline{RAS}$	Row address strobe command
$\overline{CAS}$	Column address strobe command
$\overline{WE}$	Write enable
DQMU/DQML	Input/output mask
CLK	Clock input
CKE	Clock enable
V_{DD}	Power for internal circuit
V_{SS}	Ground for internal circuit
$V_{DD}Q$	Power for DQ circuit
$V_{SS}Q$	Ground for DQ circuit
NC	No connection

Fig. 6.21. The pin arrangement and pin functions of a 64 Mb SDRAM [6.27]

Table 6.2. A comparison of the specifications of high-throughput 64 Mb/72 Mb DRAMs [6.14]

Items	EDO DRAM	SDRAM	DDR SDRAM	SLDRAM	D-RDRAM
Input signals	Async. Clocks ($\overline{RAS}$, $\overline{CAS}$) Controls ($\overline{WE}$, $\overline{OE}$) Addresses	Sync. clock (CLK) Commands ($\overline{CS}$, $\overline{RAS}$, $\overline{CAS}$, $\overline{WE}$, BA) Addresses		Sync. clock (CCLK) packet Protocol (command/address)	Sync. clock (CFM, CTM) Packet protocol (command/addresses)
Number of banks	1	2, 4	4, 8	8	16 dependant
Data bus	×4, 8, 16	×4, 8, 16	×4, 8, 16 +data strobe (DQS)	×18 +data strobe (DCLK)	×16, 18
Read/Write latency	1/0	2, 3/0	2, 2.5/1	0.5–16/0.5–16	8–12/6
Interface	TTL/LVTTL	LVTTL/SSTL	SSTL	SSTL type	Rambus channel
Frequency Clock	33–60 MHz(CAS)	100–133 MHz	100–133 MHz	200 MHz	400 MHz
Data	33–60 MHz	100–133 MHz	200–266 MHz	400 MHz	800 MHz
Peak throughput[a]	66–120 Mbyte/s	200–266 Mbyte/s	400–532 Mbyte/s	900 Mbyte/s	1.8 Gbyte/s

[a] At maximum data-bus width.

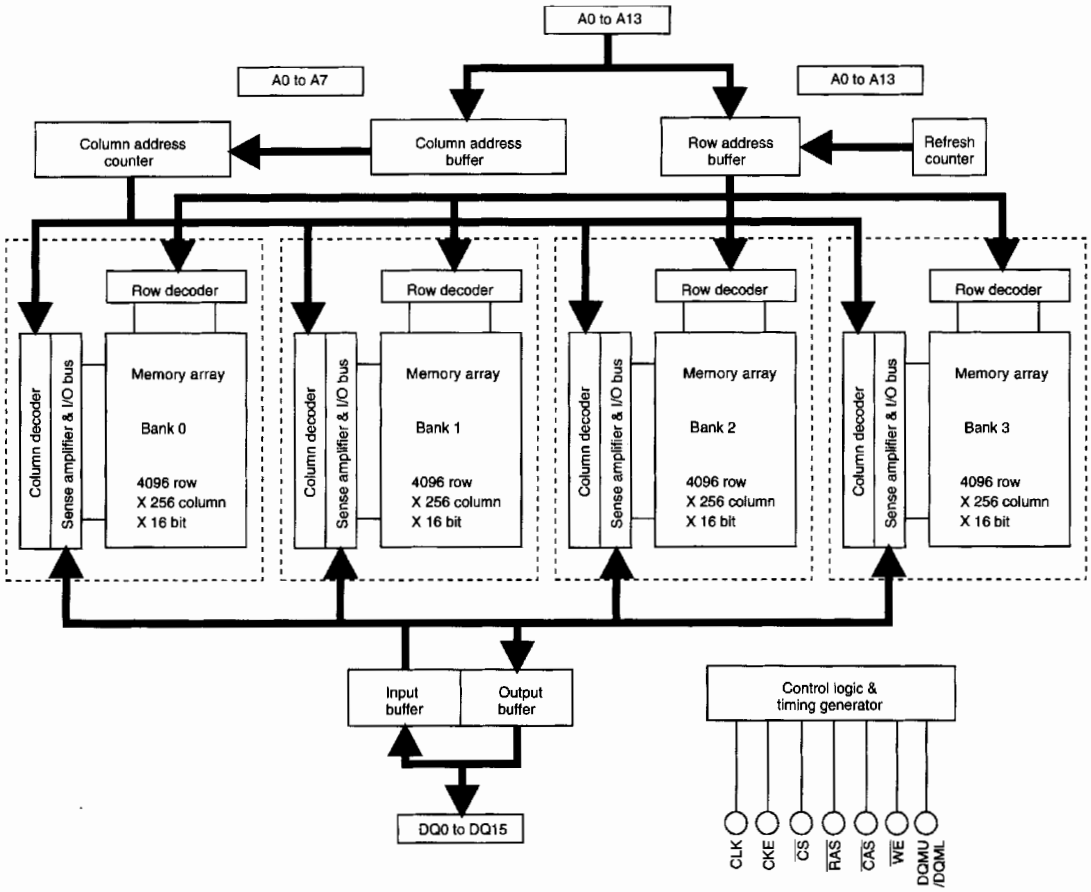


Fig. 6.22. The block diagram of a 64 Mb SDRAM [6.27]

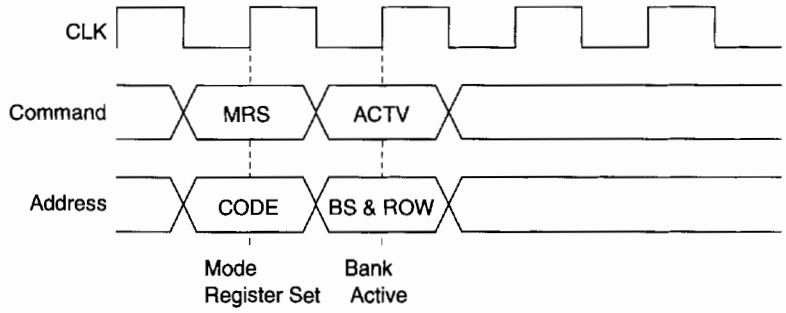


Fig. 6.23. The mode-register set cycle [6.27]

Pin Functions.

- CLK: CLK is the master clock input to this pin. The other input signals are referred at the CLK rising edge.
- $\overline{\text{CS}}$: When $\overline{\text{CS}}$ is Low, the command input cycle becomes valid. When $\overline{\text{CS}}$ is High, all inputs are ignored. However, internal operations (bank active, burst operations, etc.) are held.

Table 6.3. The characteristics of a 64 Mb SDRAM [6.27]

Parameter	Symbol	Specification		Unit	Test condition[a]
		min.	max.		
Ambient temperature	T_a	0	70	°C	
Power supply voltage	$V_{DD}/V_{DD}Q$	3.0	3.6	V	
System clock cycle time	t_{CK}	10	–	ns	
Row access time	(t_{RAC})	–	40	ns	$t_{RCD} = 20$ ns, $CL = 2, C_L = 50$ pF
Row cycle time	t_{RAC}	–	70	ns	$BL = 1$
Operating current	I_{CC1}	–	75	mA	t_{RC} =70 ns
Stand-by current	I_{CC6}		400	μA	self-refresh
Input high voltage	V_{IH}	2.0	$V_{DD} + 0.3$	V	
Input low voltage	V_{IL}	− 0.3	0.8	V	
Output high voltage	V_{OH}	2.4	–	V	$I_{OH} = -4$ mA
Output low voltage	V_{OL}	–	0.4	V	I_{OL} =4 mA
Input capacitance (CLK)	C_{I1}	2.5	4	pF	
Input capacitance (Input)	C_{I2}	2.5	5	pF	
Output capacitance (DQ)	C_O	4	6.5	pF	
Refresh time	t_{REF}	–	64	ms	
Refresh cycles	n	4096	–	n	

[a] CL, $\overline{\text{CAS}}$ latency; BL, burst length.

- $\overline{\text{RAS}}$, $\overline{\text{CAS}}$, and $\overline{\text{WE}}$: Although these pin names are the same as those used in traditional DRAMs, they function in a different way. These pins define operation commands (read, write, etc.) depending on the combination of their voltage levels, as explained later.
- A_0 to A_{11}: The row address (AX_0 – AX_{11}) is determined by the A_0 – A_{11} level at the rising edge of the bank active command cycle CLK. The column address (AY_0 – AY_7) is determined by the A_0 – A_7 level at the rising edge of the read or write command cycle CLK. And this column address becomes the burst-access start-address. A_{10} defines the precharge mode. When A_{10} = High at the precharge command cycle, all banks are precharged. But when A_{10} = Low at the precharge command cycle, only the bank that is selected by A_{12}/A_{13} (BS) is precharged, as explained later.
- A_{12}/A_{13}: A_{12}/A_{13} are bank select signals (BS). The memory array is divided into bank 0, bank 1, bank 2, and bank 3, each of which contains

4096 rows × 256 columns × 16 bits. If A_{12} is Low and A_{13} is Low, bank 0 is selected. If A_{12} is High and A_{13} is Low, bank 1 is selected. If A_{12} is Low and A_{13} is High, bank 2 is selected. If A_{12} is High and A_{13} is High, bank 3 is selected.

- CKE: This pin determines whether or not the next CLK is valid. If CKE is High, the next CLK rising edge is valid. If CKE is Low, The next CLK rising edge is invalid. This pin is used for power-down mode, clock suspend mode, and self-refresh mode.
- DQMU/DQML: DQMU and DQML mask the upper and lower bytes of the DQ data, respectively.
- DQ_0 – DQ_{15}: Data is input to and output from these pins.
- V_{DD} and $V_{DD}Q$: 3.3 V is applied. (V_{DD} is for the internal circuit and $V_{DD}Q$ is for the output buffer.)
- V_{SS} and $V_{SS}Q$: Ground is connected. (V_{SS} is for the internal circuit and $V_{SS}Q$ is for the output buffer.)

Command Operation. The SDRAM recognizes the following commands specified by the $\overline{CS}$, $\overline{RAS}$, $\overline{CAS}$, $\overline{WE}$, and address pins. Typical operations are shown in Table 6.4. After power-on, various operations are under the control of a mode register in the control logic and timing generator shown in Fig. 6.22.

Mode Register Set [MRS]. The SDRAM has a mode register that defines how it operates. After power-on, the contents of the mode register are undefined,

Table 6.4. Various command operations [6.27][a]

Command	Symbol	CKE		$\overline{CS}$	$\overline{RAS}$	$\overline{CAS}$	$\overline{WE}$	A_{12}/ A_{13}	A_{10}	A_0–A_{11}
		$n-1$	n							
Mode register set	MRS	H	×	L	L	L	L	V	V	V
Row address strobe and bank active	ACTV	H	×	L	L	H	H	V	V	V
Column address and read command	READ	H	×	L	H	L	H	V	L	V
Column address and write command	WRIT	H	×	L	H	L	L	V	L	V
Precharge select bank	PRE	H	×	L	L	H	L	V	L	×
Precharge all bank	PALL	H	×	L	L	H	L	×	H	×
Refresh	REF/SELF	H	V	L	L	L	H	×	×	×

[a] H, V_{IH}; L, V_{IL}; ×, V_{IH} or V_{IL}; V, valid address input.

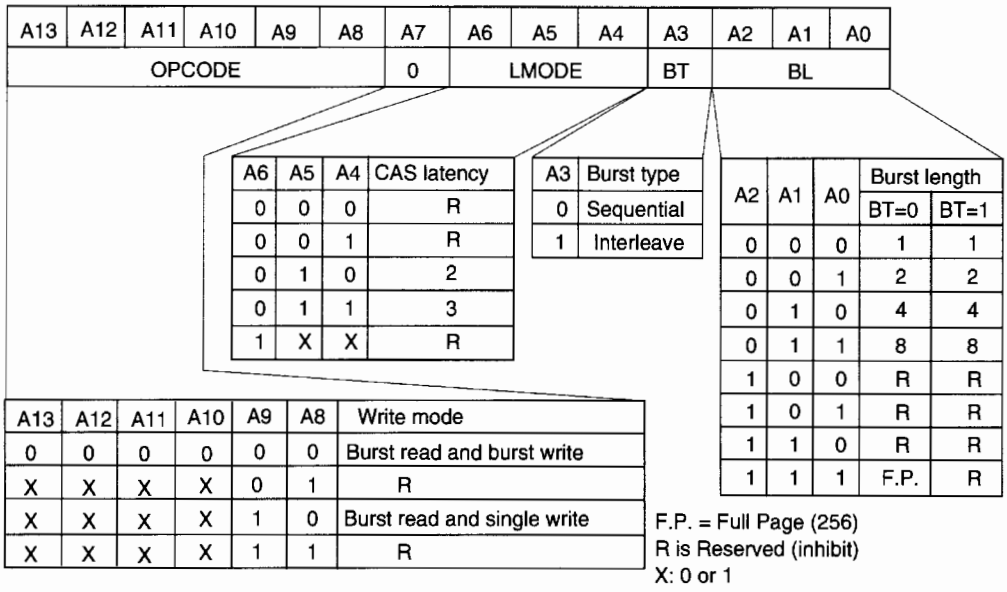


Fig. 6.24. The mode-register configuration [6.27]

so the mode-register set command must be executed to set up the mode register. The configuration is shown in Fig. 6.24. The mode register consists of five sections, each of which is assigned to address pins. It is specified by the address pins (A_0 – A_{13}) at the mode-register set cycle shown in Fig. 6.23.

- A_{13}, A_{12}, A_{11}, A_{10}, A_9, A_8(OPCODE): The SDRAM has two types of write modes. One is the burst write mode, and the other is the single write mode. The following bits specify the write mode. Burst read and burst write: burst write is performed for the specified burst length, starting from the column address specified in the write cycle. Burst read and single write: data is only written to the column address specified during the write cycle, regardless of the burst length.
- A_7: Keep this bit Low at the mode-register set cycle. If this pin is High, the vender test mode is set.
- A_6, A_5, A_4(LMODE): These pins specify the $\overline{\text{CAS}}$ latency.
- A_3(BT): A burst type is specified. When full-page burst is performed, only "sequential" can be selected.
- A_2, A_1, A_0(BL): These pins specify the burst length (BL).

Row Address Strobe and Bank Activate [ACTV]. Before executing a read or write operation, the corresponding bank and the row address must be activated by the bank active (ACTV) command. Bank 0, bank 1, bank 2, or bank 3 is activated according to the status of the A_{12}/A_{13} pin, and row address (AX_0 – AX_{11}) is activated by the A_0 – A_{11} pins at the bank active command cycle. An interval of t_{RCD} (i.e. two clock cycles) is required between the bank

active command input and the following read/write command input. In the case of the same bank, the interval between the two bank-active commands must be no less than t_{RC} (i.e. more than seven clocks for BL = 1), while for the different bank the interval must be no less than t_{RRD} (i.e. more than two clocks).

Column Address Strobe and Read Command [READ]. A read operation starts when a read command is inputted. The output buffer becomes Low-Z in the ($\overline{CAS}$ latency-1) cycle after read command set. The burst length can be set to one, two, four, eight or a full page of 256. The start address for a burst read is specified by the column address (AY_0 – AY_7) and the bank select address (A_{12}/A_{13}) at the read command set cycle. In a read operation, data output starts after the number of clocks specified by the $\overline{CAS}$ latency. The $\overline{CAS}$ latency can be set to two or three. When the burst length is one, two, four, eight, the D_{out} buffer automatically becomes High-Z at the next clock after the successive burst-length data has been output. The $\overline{CAS}$ latency and burst-length data must be specified at the mode register. Figure 6.25 shows schematic read-timing waveforms for a burst length of four [6.27]. When another read command is executed at the same row address of the same bank as the preceding read command execution, the second read can be performed after an interval of no less than one clock. Even when the first command is a burst read that is not yet finished, the data read by the second command will be valid, as shown in Fig. 6.26. When the row address changes on the same bank, however, consecutive read commands cannot be executed; it is necessary to separate the two read commands by a precharge command and a bank-active command. When the bank changes, the second read can be performed after an interval of no less than one clock, provided that the other bank is in the bank-active state. Even when the first command is a burst read that is not yet finished, the data read by the second command will be valid, as shown in Fig. 6.27. Figure 6.28 and Table 6.5 show the detailed waveforms and ac characteristics for the same bank and different row addresses [6.27].

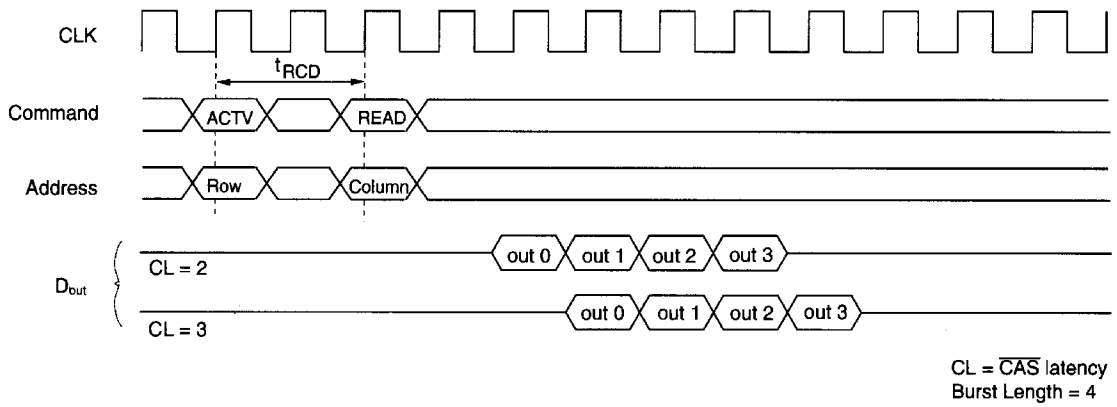


Fig. 6.25. Schematic read-timing waveforms for a burst length of 4 [6.27]

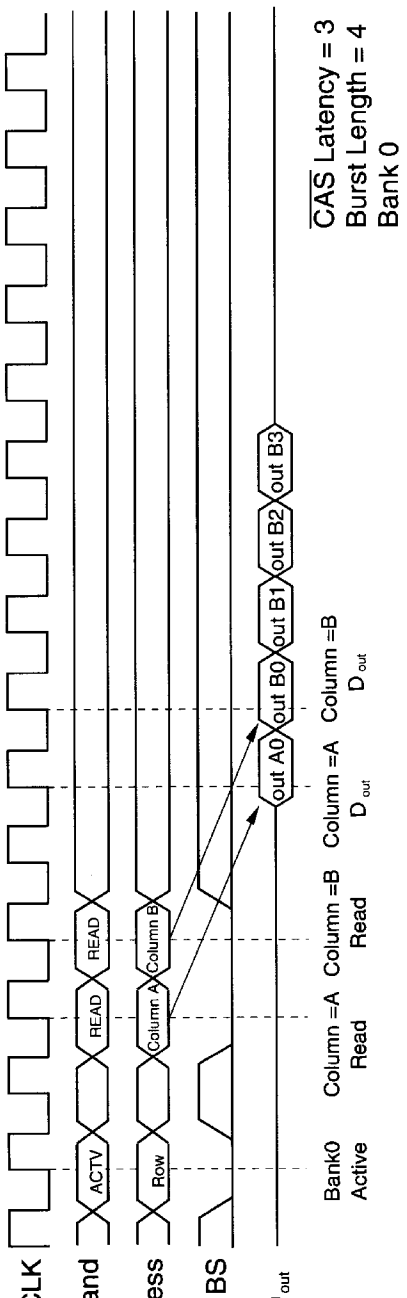


Fig. 6.26. Burst read to burst read at the same row address of the same bank [6.27]

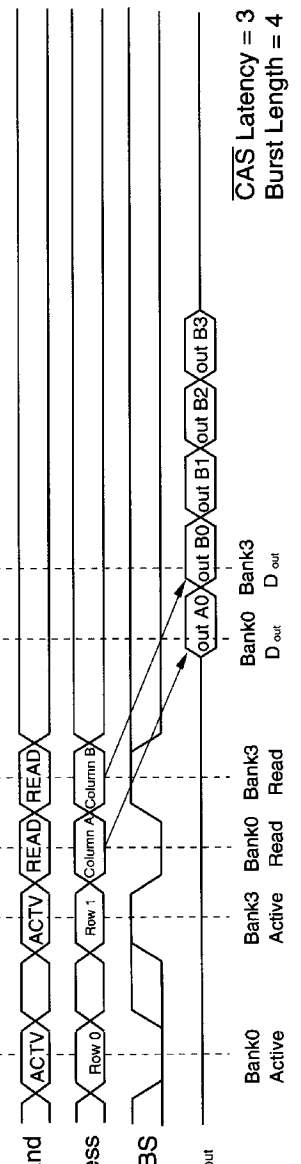


Fig. 6.27. Burst read to burst read at different banks [6.27]

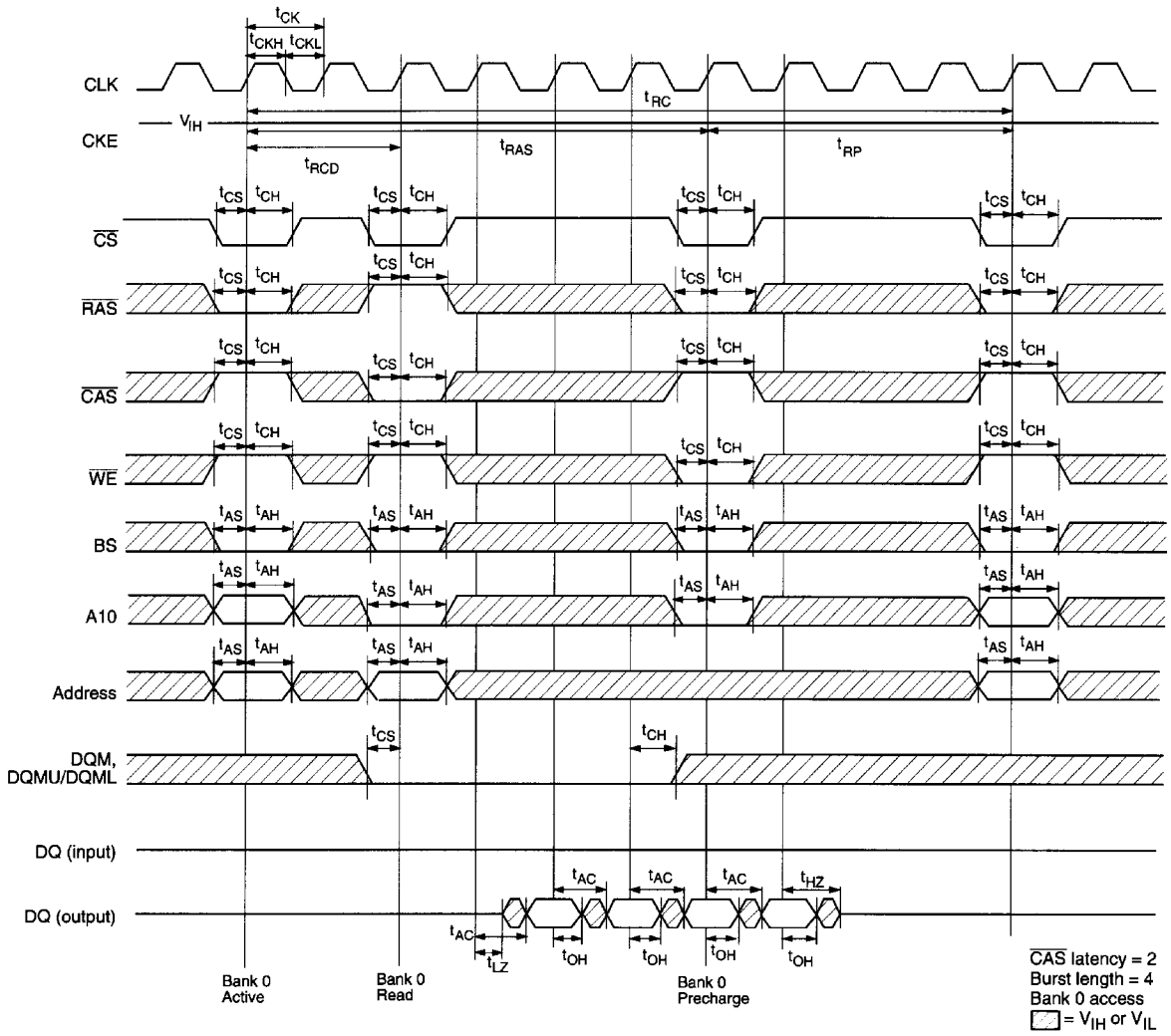


Fig. 6.28. Detailed read-timing waveforms for a burst length of 4 for the same bank (Bank 0) and a different row address [6.27]. The precharge start cycle of Bank 0 is one cycle before the final data is outputted. The next command execution for a different row address starts with the bank active (ACTV) command after t_{RC}

Column Address Strobe and Write Command [WRIT]. This command starts a write operation. The burst write or single write mode is selected by the OPCODE (A_{13}, A_{12}, A_{11}, A_{10}, A_9, A_8) of the mode register. A burst write operation is enabled by setting OPCODE (A_9, A_8) to $(0, 0)$. A burst write starts in the same clock as a write command set. (The latency of data input is zero clocks.) The burst length can be set to one, two, four, eight, and full-page, like burst read operations. The write start address is specified by the column address (AY_0 – AY_7) and the bank select address (A_{12}/A_{13}) at the write-command set cycle. Write transaction timing is very much like read transaction timing. Figure 6.29 shows schematic write-timing waveforms for a burst length of four [6.27]. When another write command is executed at

Table 6.5. The ac characteristics for a $\overline{\text{CAS}}$ latency of 2 [6.27]

Parameter	Symbol	Min. (ns)	Max. (ns)
System clock cycle time	t_{CK}	10	–
CLK high pulse width	t_{CKH}	3	–
CLK low pulse width	t_{CKL}	3	–
Access time from CLK	t_{AC}	–	6
Data-out hold time	t_{OH}	3	–
CLK to Data-out low impedance	t_{LZ}	2	–
CLK to Data out high impedance	t_{HZ}	–	6
Input setup time	$t_{AS}, t_{CS}, t_{DS}, t_{CES}$	2	–
Input hold time	$t_{AH}, t_{CH}, t_{DH}, t_{CEH}$	1	–
Ref/active to Ref/active command period	t_{RC}	70	–
Active to precharge command period	t_{RAS}	50	120 000
Active command to column command (same bank)	t_{RCD}	20	–
Precharge to active command period	t_{RP}	20	–
Write recovery or data-in to precharge lead time	t_{DPL}	10	–
Active (a) to active (b) command period	t_{RRD}	20	–
Transition time (rise to fall)	t_T	1	5

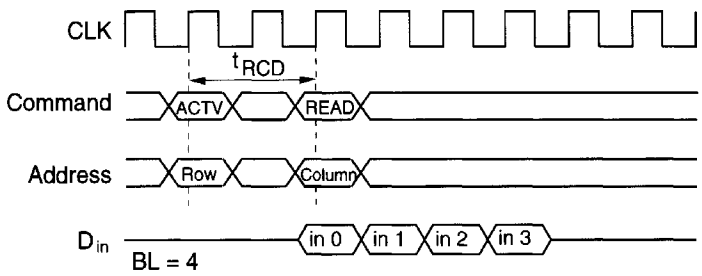


Fig. 6.29. A schematic of write-timing waveforms for a burst length of 4 [6.27]

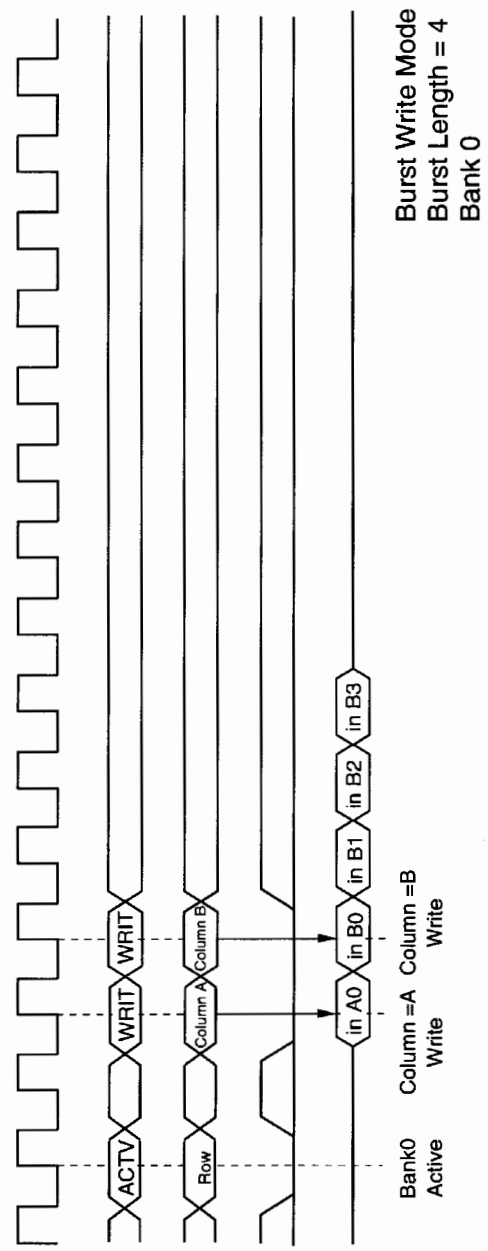


Fig. 6.30. Burst write to burst write at the same row address of the same bank [6.27]

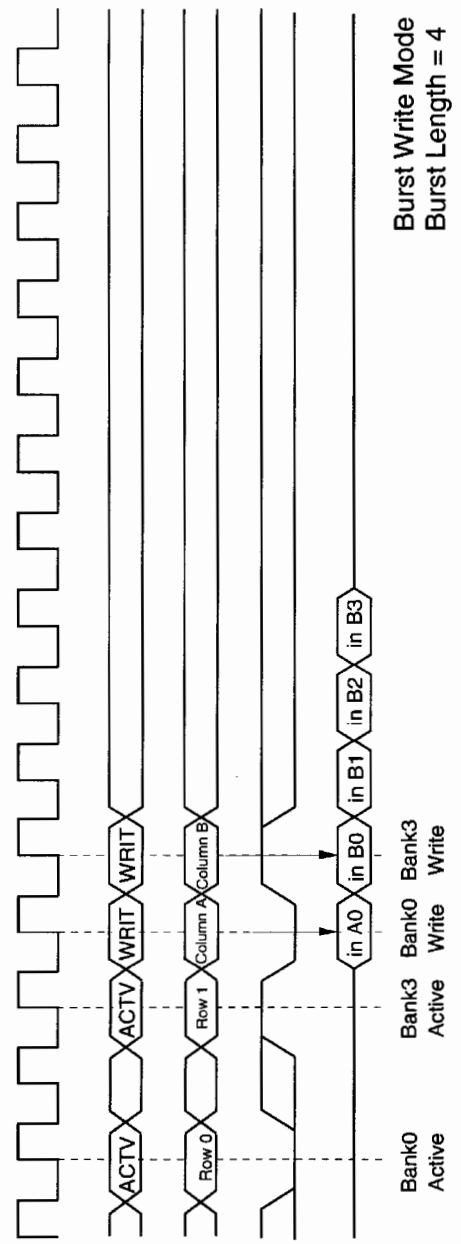


Fig. 6.31. Burst write to burst write at different banks [6.27]

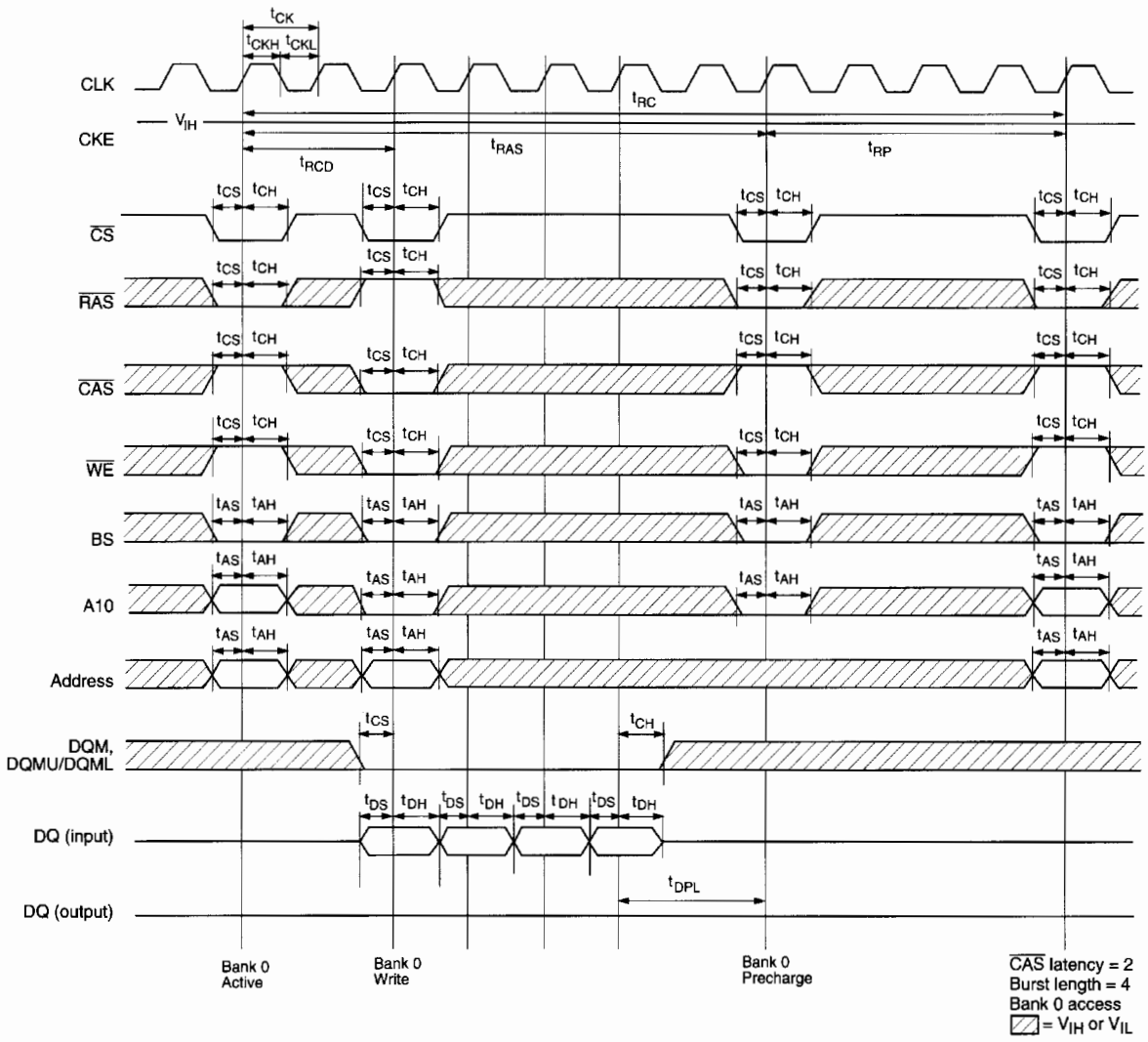


Fig. 6.32. Detailed write-timing waveforms for a burst length of 4 for the same bank and a different row address [6.27]. The ac characteristics are shown in Table 6.5

the same row address of the same bank as the preceding write command, the second write can be performed after an interval of no less than one clock, as shown in Fig. 6.30 [6.27]. In the case of burst writes, the second write command has priority. When the row address changes, consecutive write commands cannot be executed: it is necessary to separate the two write commands by a precharge command and a bank-active command. When the bank changes, the second write can be performed after an interval of no less than one clock, provided that the other bank is in the bank-active state. In the case of burst writes, the second write command has also priority, as shown in Fig. 6.31 [6.27]. Figure 6.32 shows the detailed waveforms for the same bank and different row addresses [6.27].

Precharge Selected Bank [PRE]. This command starts the precharge operation for the bank selected by A_{12}/A_{13}. If A_{12} and A_{13} are Low, bank 0 is selected. If A_{12} is High and A_{13} is Low, bank 1 is selected. If A_{12} is Low and A_{13} is High, bank 2 is selected. If A_{12} and A_{13} are High, bank 3 is selected.

Precharge All Banks [PALL]. This command starts a precharge operation for all banks.

Refresh [REF/SELF]. This command starts the refresh operation. There are two types of refresh operation: the one is auto-refresh, and the other is self-refresh.

6.4.3 Rambus DRAM

In Rambus DRAM (RDRAM) necessary data are serially transferred taking several clock cycles on a narrow bus, which is characterized by an extremely high-speed interface, with packet protocols. Thus, despite a narrow bus width, a large number of data are obtained within a short period of several clock cycles. In the following section, Direct RDRAM (D-RDRAM), which is the third generation of RDRAM as a result of modifications of the first- and second-generation RDRAMs (i.e. Base RDRAM and Concurrent RDRAM), will be explained by simply calling it RDRAM.

Figure 6.33 shows the bus interface called the Rambus channel. It features 18-bit data lines with 9 bits for each of data A (DQA) and B (DQB), and 8-bit address/command lines with 3 bits (ROW) for chip selection, bank selection, row addresses, and so on, and the remaining 5 bits (COL) for column addresses and commands. Two pairs of differential clock lines, $T_{\rm clk}$ and $R_{\rm clk}$, are for write and read operations, respectively. High speed is realized by a small-swing (0.8 V) impedance-matched interface, a bus structure strictly specified for the input/output characteristics of RDRAM and its package and wiring rules on the board, a doubled data rate of 800 MHz obtained by aligning data at both edges of clock of 400 MHz, and a loop-around clock. In the loop-around clock, write data, for example, from the controller and the

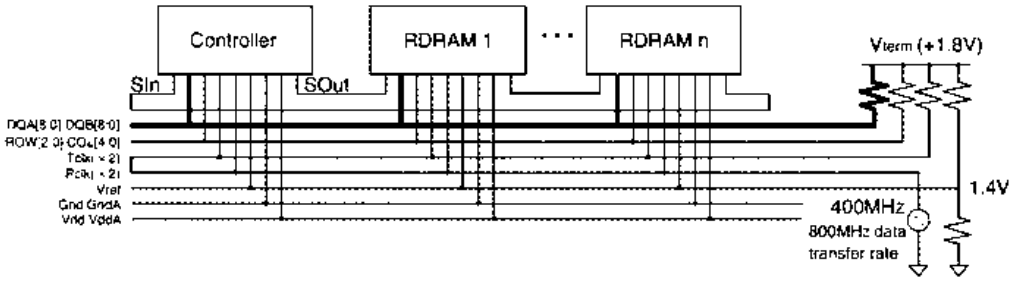


Fig. 6.33. A memory controller and RDRAMs connected to terminated transmission lines [6.28–6.30]

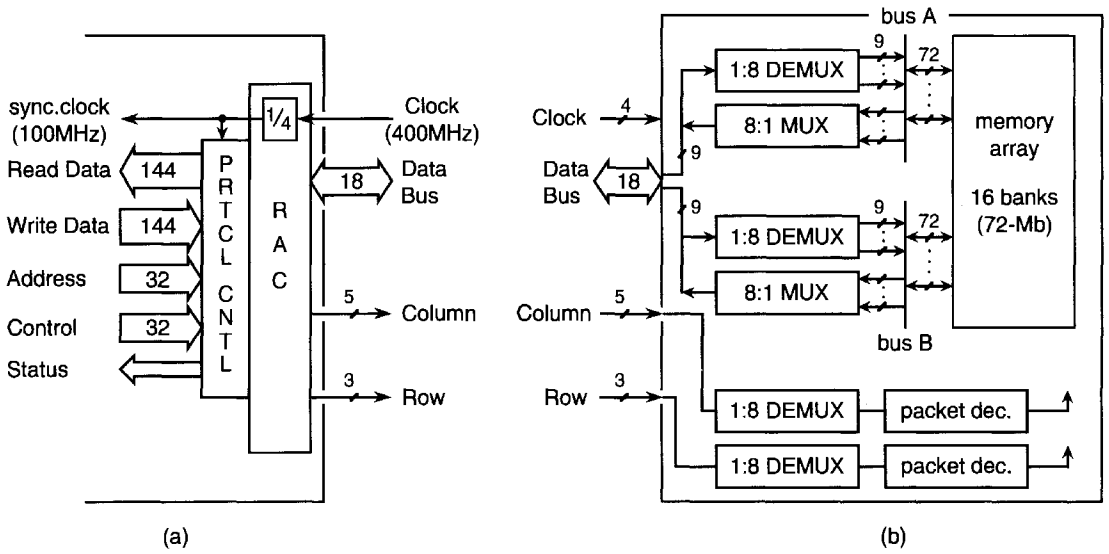


Fig. 6.34. The configurations of a memory controller (**a**) and an RDRAM (**b**) [6.28–6.30]. RAC, Rambus ASIC Cell; PRTCL CNTL, protocol controller

write clock (T_{clk}) run in the same direction simultaneously, so as to cancel skews between the clock pulse and data pulses.

Figure 6.34 shows schematic configurations of the interface of a controller LSI and a 72 Mb RDRAM. The controller comprises RAC (Rambus Asic Cell), which converts an 18-bit external memory data bus operable at 400 MHz clock to a 144-bit internal bus operable at 100 MHz clock, and a protocol control circuit to deal with packets. A 72 Mb RDRAM consists of 16 banks, 4.5 Mb($= 512 \times 64 \times 144$) for each, two multiplexers (MUXs) to convert 144-bit read data to 18-bit read data, two demultiplexers (DEMUXs) to convert 18-bit write data to 144-bit write data, a demultiplexer and packet decoder for three row addresses, a demultiplexer and packet decoder for five column addresses, and a control register.

Figure 6.35 shows the packet structure [6.28]. There are three kinds of packets: the data packet for data A and B (DQA and DQB), the row address (ROW) packet, and the column address (COL) packet. All necessary information is serially transferred at both edges of the clock, taking four clock cycles (i.e.10 ns at 400 MHz). The data packet is composed of 18 bytes($= 144$ bits) in total for data A and data B, with 18 bits at every edge. The row address packet stores three kinds of information for selection: chip (DR [4:0]), bank (BR [3:0]), and row address (R [8:0]). The column address packet is used for designating chip (DC [4:0]), bank (BC [3:0]), and column address (C [5:0]). COP [3:0] is the operation code that designates operations such as read, write, and refresh.

Figure 6.36 shows a read transaction of 36-byte data. As soon as successive selections for chip, bank, and row address are performed in the period ACT_{a0}, a word line is activated so that the resulting cell signals along the word line

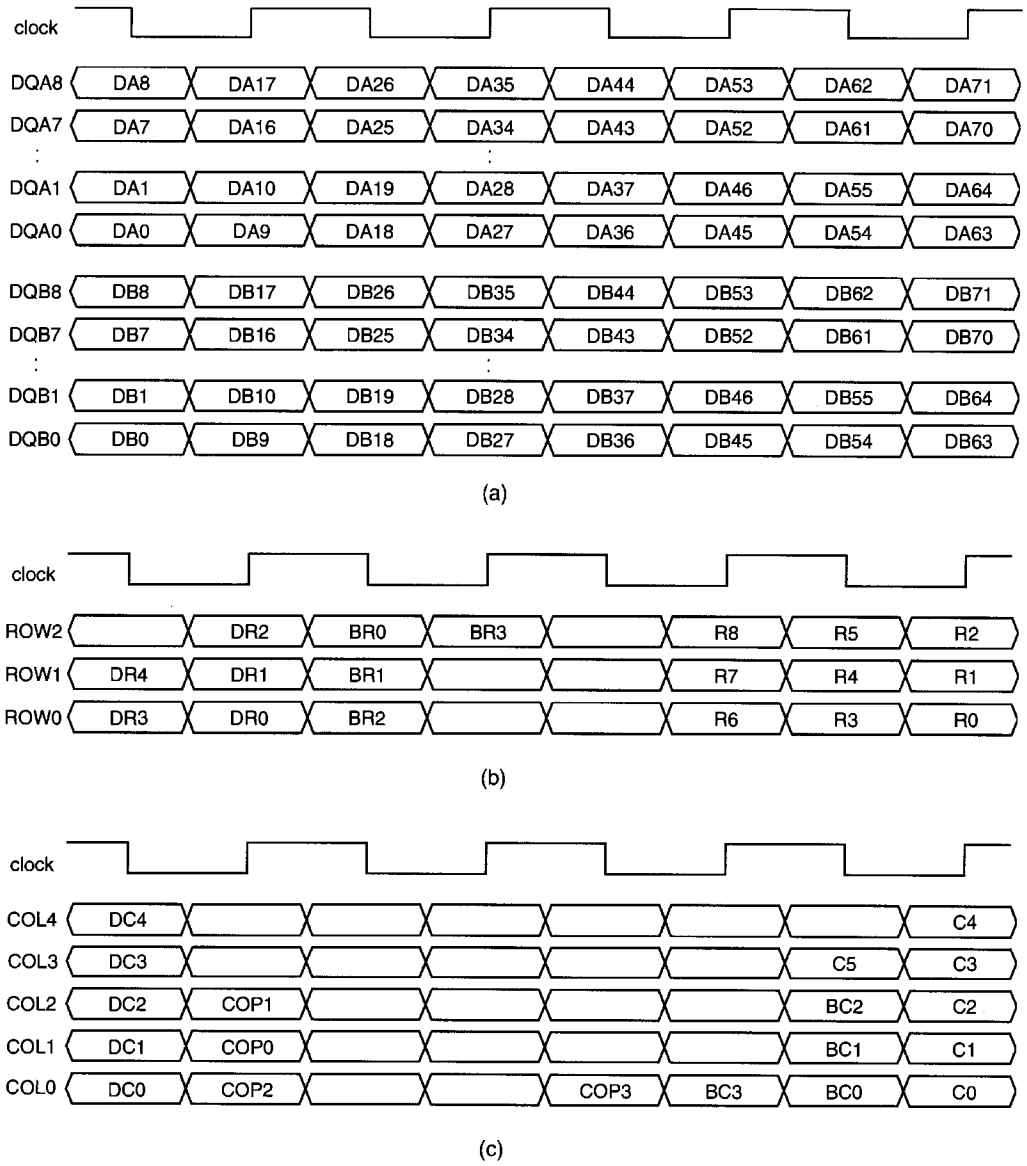


Fig. 6.35. A packet structure consisting of a data packet (**a**), a row-address packet (**b**), and a column-address packet (**c**) [6.28]

are amplified and then latched at sense amplifiers. After receiving the column addresses and operation code in the period RD_{a1}, the DRAM core sends the selected 18-byte data of the latched data to the data busses A and B in Fig. 6.34b, and then to the memory controller at Q(a1). Another 18 bytes of data are available at Q(a2) by the succeeding column selection at RD_{a2}. After the data have been transferred to the memory controller, the DRAM core is precharged under the control of an autoprecharge command designated by COP in the column address packet, so that the next row cycle starts.

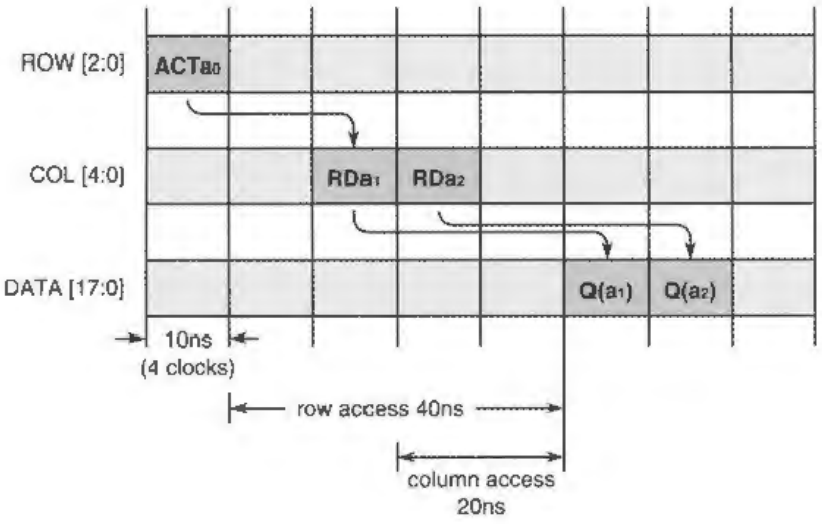


Fig. 6.36. The read transaction [6.30]

6.5 Embedded Memories

Embedded memories [6.14] enable low power consumption, high performance, and flexible system design. Figure 6.37 shows the power comparison of a 32-bit MPU system [6.9]. For a system using discrete MPU and standard TTL-interface DRAMs, in which the external data bus width is 32 bits, DRAMs dominate the total system power. The power, however, was reduced to one-

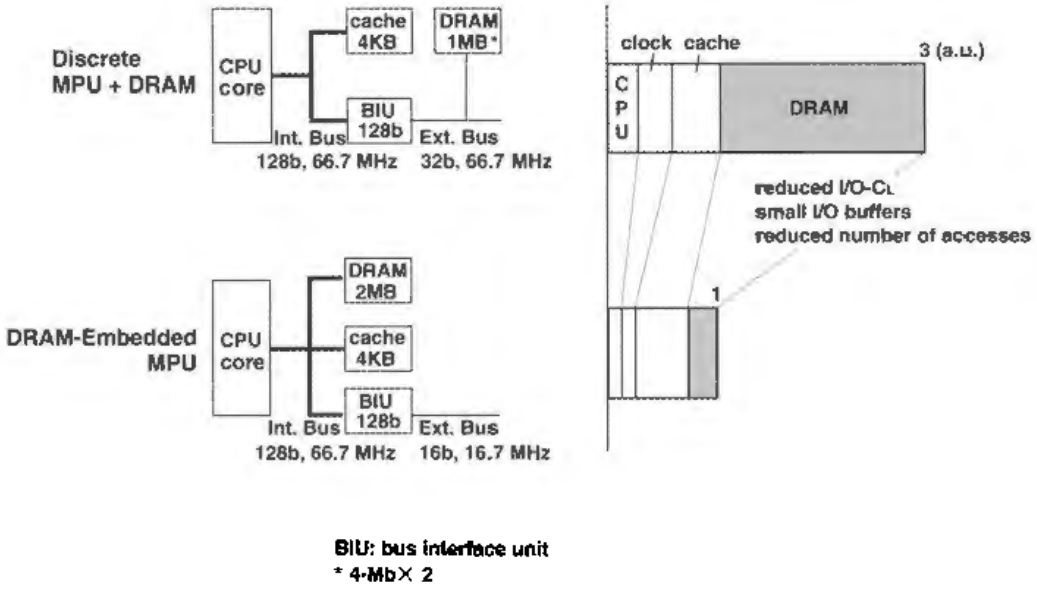


Fig. 6.37. The power reduction of a 32-bit MPU system with an embedded DRAM [6.9]

third by the DRAM-embedded system, which enables a wide internal data bus width of 128 bits. The reduced I/O capacitances, small internal buffers, and reduced number of accesses to DRAM that are caused by the wide data-bus width are responsible for the reduction. A 64 Mb DRAM macro core for ASICs using a 0.25 μm, p-sub, triple-well, five Al-layers CMOS process has also been reported [6.31]. Its random access and cycle times of 6.8 ns and 9.1 ns are record-setting performances, which are about five times faster than those ever reported for standard DRAMs. A throughput of 12.8 Gbytes/s is also outstanding. Moreover, a 12 ns access-time 64 Mb secondary cache on a 64-bit MPU [6.32], and a 3.7 ns address access-time 1 GHz 8 Mb secondary cache [6.33] have been reported. The origins of such high performance are low internal-node capacitances and the use of a multidivided array composed of small subarrays of 8 Kb to 16 Kb, multilevel Al-wiring running over the memory array, high-speed circuits such as non-address multiplexing and direct sensing (see Fig. 3.63), and a wide data bus of 1024. An embedded Flash memory also enables a short time-to-market for system products, quick specification changes after the start of mass production, and easy maintenance by end users. A 4 Mb embedded Flash memory is now ready for the market.

Embedded DRAMs can achieve exceptionally high throughput, which is particularly important for graphics and image applications. The massively parallel processing capability that the embedded DRAM offers meets the requirement for Tbyte/s throughput (Fig. 6.6) in the gigabit era. The key is the use of a wider internal I/O bus array shown in Fig. 6.12b, combined with low DL power circuits. Figure 6.38 compares an embedded DRAM [6.17] with a standard DRAM in more detail. For the standard DRAM, each unit of data from a number of subarrays is read out to an external I/O pin. The speed is degraded by the column selection, a relatively large internal i/o line capacitance of about 3 pF for a 0.4 μm 16 Mb process, the main amplifier, the large data output buffer, and the heavy I/O capacitance that comes from the pad, the package pin, and the wiring capacitance on the memory board. For the embedded DRAM, however, each i/o line could be connected directly to each logic circuit block. Thus, each i/o line can be regarded as the I/O line in the standard DRAM, and thus the data output buffer and large capacitances in the standard DRAM can be eliminated. In this case one i/o line is shared by eight DLs to match the layout pitch of the logic circuit block, allowing a huge number of data lines. Each i/o line (i/o_1) capacitance that determines the internal operating frequency is negligibly small, perhaps less than 1 pF. Thus, all the massive data, read out on all the data lines by selecting one word line, are processed simultaneously at an extremely high speed. Note that the number of data lines increases with memory capacity, whereas the number of i/o's in the standard DRAM cannot necessarily increase because of the difficulties explained earlier. Furthermore, the data-line and the i/o capacitances decrease with high-density device technology, while

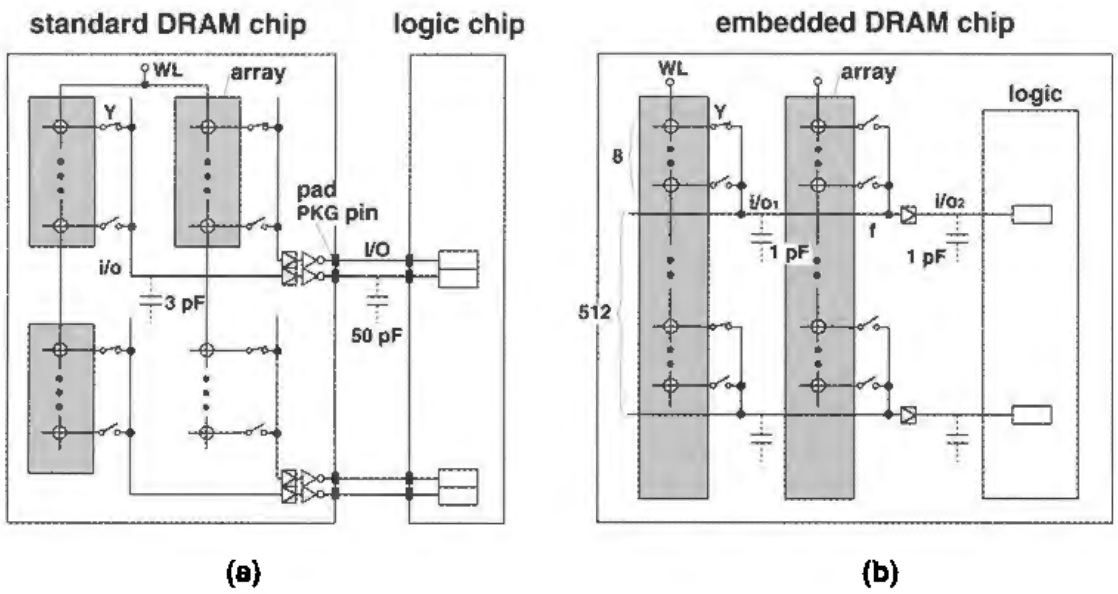


Fig. 6.38. A comparison of (a) a standard DRAM and (b) an embedded DRAM [6.17]. Operating frequency at i/o, f, which is $I/(C \cdot \Delta V)$, is estimated by the use of scaling factors normalized by those of a 0.4 μm 16 Mb chip: 2.8 at 256 Mb and 13.1 at 16 Gb. I, C, and ΔV are the i/o_1 driving current, the i/o_1 capacitance, and the i/o_1 voltage swing, respectively

the external I/O capacitance never reduces without packaging innovation. Consequently, the performance gap between these two kinds of DRAMs will increase with increasing memory capacity and with higher-density device technology. Figure 6.39 summarizes the estimated throughputs of DRAM chips. For the standard DRAM, the column mode frequency has increased by successive use of the page mode, the fast page mode, EDO, and sychronous DRAM by a factor of roughly $\sqrt{2}$ per generation. The number of I/O pins rapidly increased until the 4 Mb generation, then the rate of increase slowed to about $\sqrt{2}$ per generation due to the increasing difficulty, as discussed before. As a result, the standard DRAM chip will realize doubled-throughput per generation, reaching 10 GB/s in the 16 Gb generation with 0.6 GHz and 128 I/Os. For the embedded DRAM, a 1.67 time speed improvement, based on the device scaling theory and a doubled number of I/Os per generation, are assumed. Thus, the embedded DRAM will realize a 3.3 time throughput improvement per generation, reaching 7 TB/s in the 16 Gb generation with 1.7 GHz and 32 k i/o's. Note the ever-increasing gap in throughput between the two kinds of DRAMs, as explained previously. Such high performance of the embedded DRAM suggests the emergence of new-system DRAM chips.

Despite the great potential of embedded memories, there are still many challenging issues [6.10] regarding fabrication cost, design automation that can cope with a wide variety of memory macros, testing, and a chip-shrinking

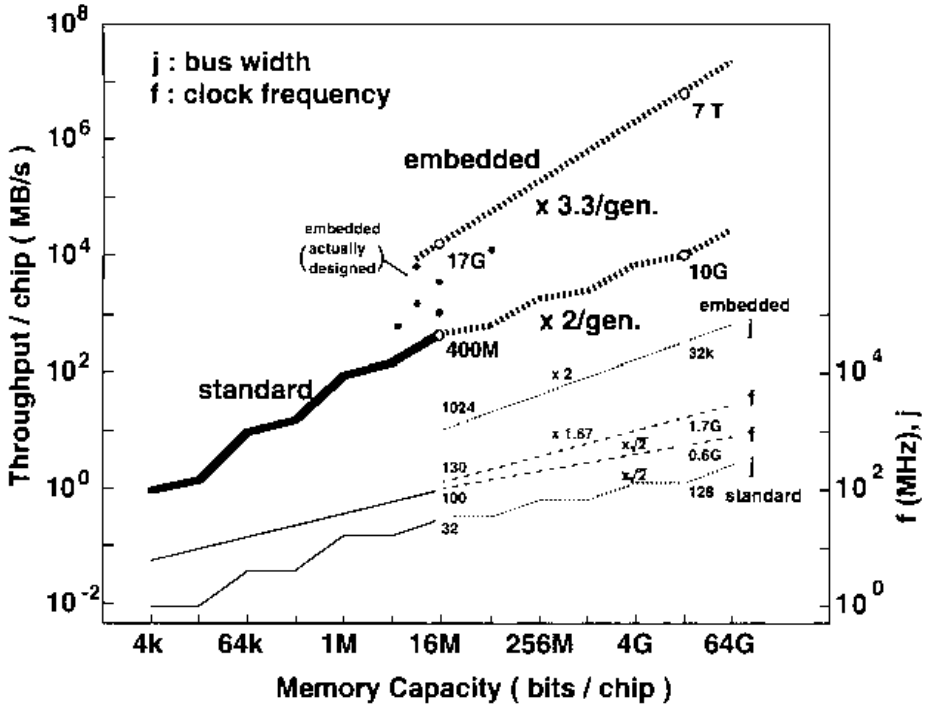


Fig. 6.39. The estimated throughput of DRAMs [6.17]

strategy to keep up with the size reduction of standard memory chips. The process cost is a particular concern, especially for embedded DRAMs. Even if the added cost, if any, of an embedded DRAM can be justified by performance advantages at the system level, compared to a standard off-chip DRAM, the fabrication cost must be reduced. Simply merging the high-speed logic process and the standard high-density DRAM process would result in an unacceptably complicated process, since the two processes are not compatible. In general, the logic process needs low-resistivity multilevel metal layers (up to seven to eight levels of aluminum or copper) for high speed. On the other hand, the DRAM process needs a different process tailored to an extremely dense memory cell. It uses fewer (two or three) low-resistivity metal layers. However, refractory metal (poly-Si or polycide) layers which are suitable for self-aligned process and fine patterning despite being resistive materials, a self-aligned contact, and storage-node processing are indispensable. Thus, simplification of the combined process is essential for embedded DRAMs, so that both high speed and high density can be achieved at low cost. If this is impossible, however, there are two alternative approaches for embedded DRAMs. One is a high-speed-oriented memory design, where the logic-compatible process is used as much as possible despite any increase in chip area caused by an enlarged memory cell. In this case, the processing cost for a typical embedded DRAM is expected to be about 25% higher than for the base process of a high-end CMOS logic

chip [6.10]. The other alternative is high-density-oriented memory design using the memory-compatible process as much as possible. The resultant slower speed might be offset by the much higher throughput of the embedded DRAM.

7. Low-Power Memory Circuits

7.1 Introduction

The low-power RAM circuit [7.1–7.4] is a major area of interest in low-power LSI research. Successive advances in the low-power RAM circuit have been able to suppress chip-power consumption, which increases with increasing memory capacity, chip area, and speed. As a result, these advances – coupled with high-density memory-cell technology – have allowed chip power consumption to be maintained or lowered (Fig. 7.1), although the memory capacity of DRAM chips has increased rapidly by over six orders (1 Kb to 4 Gb) over the past 30 years. Consequently, the low-power RAM circuit has realized low-cost, high-reliability chips because it allows plastic packaging, a low operating current, and a low junction temperature. In addition, it has managed the ever-increasing memory-subsystem power caused by the increasingly high throughput requirements of various processing systems such as personal computers. Moreover, it forms not only the basis of other LSI memory chips, such as Flash memory and ROM, but also the basis of on-chip memory subsystems, such as embedded DRAMs (merged DRAM and logic) [7.3] and SRAM caches, both of which have become increasingly important in modern memory systems.

RAM chip power is a prime concern of the subsystem designer, since it dominates memory-subsystem power. To reduce active chip power, the designers of low-power RAM circuits have focused on three key issues so far [7.1]: reductions in the charging capacitance, operating voltage, and dc current. Of these issues, a reduction in the operating voltage has become relatively important not only to reduce power, but also to ensure device reliability in scaled-down devices, and also to extend the use of LSIs to battery-based portable systems. Historically, DRAM designers initiated and then led the field in low-voltage LSI research, because their first priority was on higher-density chips, which were obtained through scaled-down FETs consequently resulting in lower breakdown voltages. Low-voltage (2–3 V) circuits have been used in actual 16 Mb and 64 Mb DRAM products, although their external supply voltages are 5 V or 3.3 V, being internally lowered by on-chip voltage down-converters to standardize the power supply [7.1–7.4]. Recent exploratory research on low-voltage operations of 3 V or less suggests the great potential of CMOS circuits, although reducing voltage inevitably

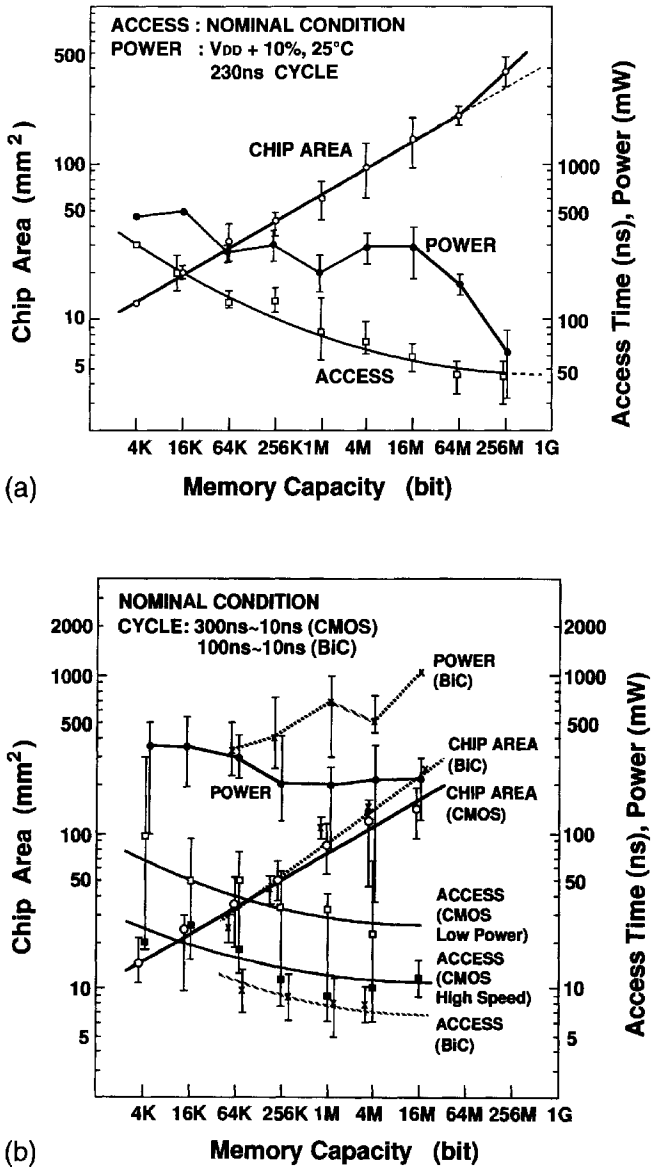


Fig. 7.1. Trends in DRAM chip performance [7.1]. The bit width for I/O pin is mostly 1 bit or 4 bits for DRAMs and high-speed SRAMs, and mostly 8 bits for low-power SRAMs. (**a**) DRAM; (**b**) SRAM

imposes memory-cell development, focusing attention on the high signal-to-noise (S/N) ratio design [7.1–7.4] discussed in Chap. 4. In fakt, a 64 Mb DRAM chip with an operating voltage of 1.5 V has been reported [7.5], affording an active current of 23 mA at a 230 ns cycle time. Reducing the data-retention power in DRAMs has been increasingly important for battery backup applications (where SRAMs are normally used), since DRAMs are inherently less expensive to produce than SRAMs. A data-retention current as low as 3 μA at 2.6 V has been reported for 4 Mb chip [7.6], as shown in Fig. 7.2. Such a low current finally allows the design of DRAM battery

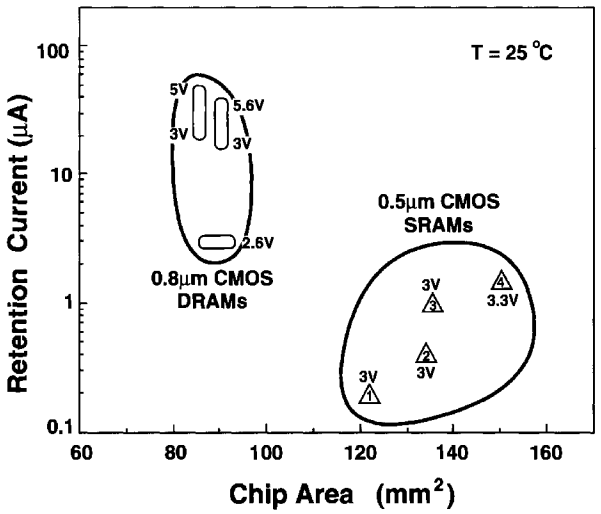


Fig. 7.2. The data-retention current versus the chip area for 4 Mb RAM chips [7.1]. 4 Mb SRAM: 1, 2, 3, TFT load cells; 4, poly load cell

operation, even in active operation mode, which is a key in battery-based handheld equipment, such as mobile communication systems.

For SRAMs, there has been also a strong requirement for low power. In the early days, SRAM chip development was focused on low-power applications, especially with very low stand-by 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 [7.7], primarily led by cache applications in high-speed microprocessors. ECL BiCMOS SRAM chips with access times of less than 10 ns are good examples. In this case, the power and chip area are of less concern, as shown in Fig. 7.1. For low-power applications, however, the primary concern is for CMOS technology to realize high speed with minimum power. Tremendous efforts have been made to minimize power, with more emphasis on static current reduction and low-voltage operation on long signal transmission lines. Consequently, there has been a decrease in CMOS SRAM power dissipation, in spite of a rapid increase in chip area and improved operating speed. However, this reduction trend does not look very significant compared with the trend in DRAMs. This is because high-speed characteristics have been more emphasized in the recent development of 1 Mb or larger SRAM chips. A typical example of a low-power, high-density SRAM chip is a 16 Mb CMOS SRAM that achieves an access time of 15 ns and a power dissipation of 165 mW at 30 MHz with 3 V power supply [7.8]. As for the data-retention current, recent 4 Mb CMOS SRAM chips achieve sub-μA data-retention currents, as shown in Fig. 7.2, which are still about one order of magnitude less than those of 4 Mb DRAM chips, because they do not require a refresh operation.

In addition to the above low-power circuit for RAM chips, system approaches [7.2] have been important. The wide-bit I/O memory chip configu-

ration for simultaneously processing a large number of data bits has become more popular as chip memory capacity increases, since it effectively reduces subsystem power, with a resulting lower chip count for a fixed subsystem memory capacity. Moreover, small package technology combined with high-density chip technology has reduced ac power, with less capacitance for address lines, control lines, and data input/output bus lines on memory boards. Lowering the voltage swing on chip-to-chip data-bus lines reduces the ac power, which is always increasing in line with the strong requirement for higher data-throughput for memory subsystems.

DRAMs and SRAMs both have reduced power dissipation through the use of common low-power circuit designs. However, there are few papers that clarify which circuits are common or different based on a systematic comparison. This chapter consists of two parts: low-power RAM subsystem design and low-power standard CMOS RAM chip design. In the first part, power sources and power reductions in RAM subsystems are discussed, with emphasis given to the importance of RAM-chip power reduction. In the second part, power sources and power reductions of RAM chips are discussed in terms of three key issues: charging capacitance, operating voltage, and dc static current, differentiating between DRAMs and SRAMs. First, in Section 7.3, power sources for active and data retention modes in a chip are described and essential differences in power sources between DRAMs and SRAMs are clarified. Next, the power reduction circuits for each power source are separately reviewed in the following sections, first for DRAMs and then for SRAMs. Note that the subthreshold dc current of a MOSFET, which will be a source of both active and data-retention power for future DRAMs and SRAMs supplied with an ultra-low voltage of less than 2 V, is discussed in Chap. 8.

7.2 Sources and Reduction of Power Dissipation in a RAM Subsystem

The major sources of active power in a modern CMOS RAM subsystem are the active chips, whose power is a product of the activated chip power and the number of simultaneously activated chips. Other minor sources are the capacitances of data-bus lines, address lines, and control lines on a memory board. In particular, the data-bus lines are of special concern. They cause a heavy capacitance compared with other lines because of their inherently large values in number and capacitance. Active chip power and ac capacitive power have both been reduced by low-power circuits, miniaturized device technology, wide-bit I/O chip configuration, small-package technology, and low-voltage interfaces. On the other hand, the sources of subsystem retention power are all the memory chips in the subsystem, and thus the retention power has also been reduced by low-power circuits and by improving device

technology. In this section, sources and reductions of power dissipation in a typical DRAM subsystem are described in terms of wide-bit I/O chip-configuration, small-package technology and low-voltage interfaces.

7.2.1 Wide-Bit I/O Chip Configuration

The wide-bit I/O configuration [7.1--7.3] is becoming more favorable in terms of memory subsystem designs as chip memory-capacity increases. It reduces subsystem power, and offers ease of use by reducing the chip count needed by the system and adding flexible add-on memory capability, if low-power chip technology is applied at each successive generation. For example, a 4 MByte (B), 32 bit data-bns subsystem needs 32 memory chips for a 1 Mb chip of 1 Mword × 1 bit configuration, as shown in Fig. 7.3a. With the wide-bit I/O configuration, however, the number of chips needed is eight for a 4 Mb chip (1 Mword × 4 bit) and two for a 16 Mb chip (1 Mword × 16 bit). Generally, in addition to increasing the memory capacity of a chip, widening the data-bit width increases the chip power, chip area, and I/O pin count, as shown in Fig. 7.4 [7.3], because of the resultant increased number of column and sense circuits. However, the chip power has been suppressed with low-power circuits for a moderate number of I/Os of less than 16. Thus, the wide-bit I/O configuration combined with larger memory capacity dramatically reduces subsystem power with a smaller chip count. The smaller chip count also allows design flexibility, so subsystem memory capacity can be added on with an increment of as small as 4 MB with a minimum memory capacity of 4 MB.

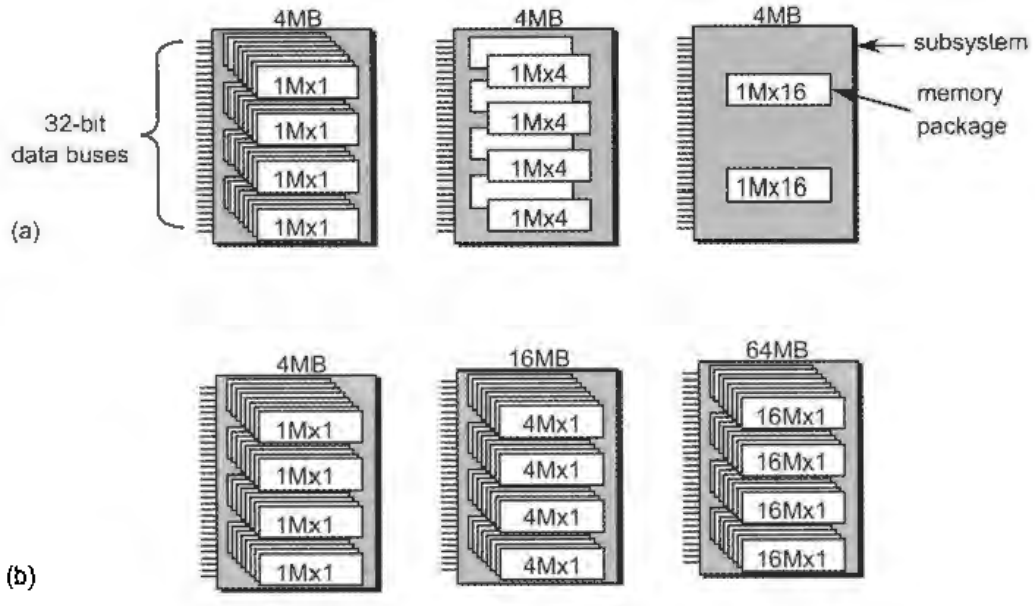


Fig. 7.3. The impact of a multibit data configuration chip on a 32-bit data-bus system. 1M×1, 1 Mword×1 bit configuration [7.2]. (**a**) Subsystem using multibit data configuration chip; (**b**) subsystem using fixed data bit configuration chip

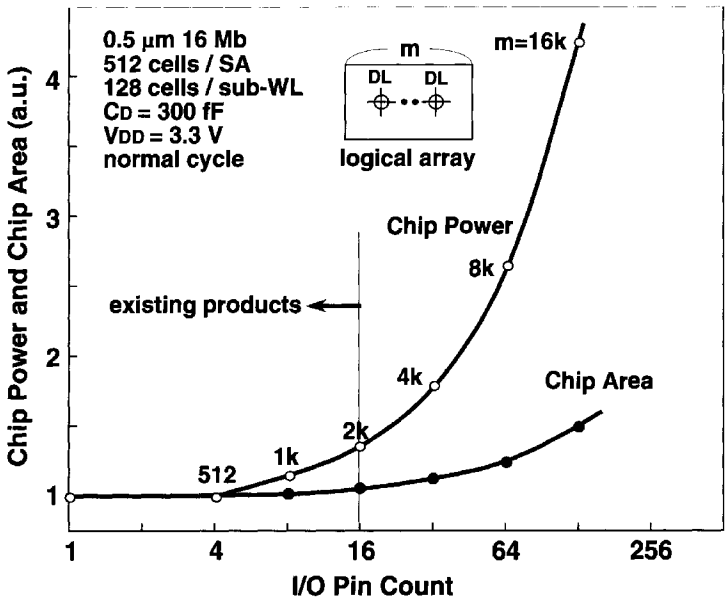


Fig. 7.4. I/O pin count versus chip power and chip area [7.3]

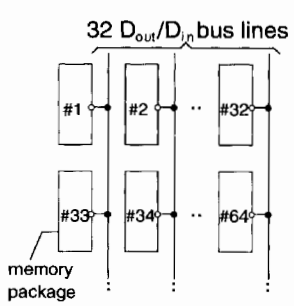
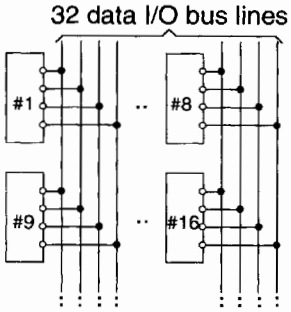
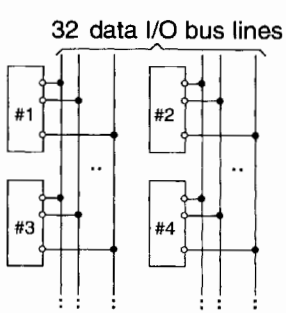
These features are essential for small systems such as personal computers and handheld equipment, in which power, physical size, and price are critical. On the other hand, a fixed data-bit configuration suffers from disadvantages such as larger power as well as design inflexibility, as shown in Fig. 7.3b. Table 7.1 shows the more detailed advantages of the wide-bit I/O configuration. The data are from a catalog specification [7.9] which guarantees the power at the worst combination within variations of V_{DD}, temperature, and the fabrication process. Obviously, a drastic reduction in subsystem power is achieved with the smaller chip count resulting from wide-bit I/O configuration combined with large memory capacity, and low-power chip-technology advancement, as shown in Table 7.2. For example, the chip component of normal-mode active power for a 110 ns cycle is reduced from 17.4 W for a 1 Mb chip subsystem down to 1.1 W for a 16 Mb chip subsystem. Note that inactive (non-selected) chips consume a negligibly small power compared with active (selected) chips. Thus, chip power is almost determined by the number of chips simultaneously activated, independent of the subsystem memory capacity.

A recent paper [7.10] suggests the possibility of at least 256 b configuration compared to the 32 b configuration for current experimental 1 Gb chips, but the number of I/Os is eventually restricted by rapid increases in chip power, chip area, and package size.

7.2.2 Small Package

In addition to the lower chip count described above, small-package technology reduces ac capacitive power by reducing the capacitance on the memory board. A common I/O pin assignment, in which the data input (D_{in}) pin

Table 7.1. 5 V DRAM subsystem with 32 data-bit bus line [7.2]

Memory chip	**1 Mb** (1 Mword × 1 bit)	**4 Mb** (1 Mword × 4 bit)	**16 Mb** (1 Mword × 16 bit)
Memory subsystem			
Number of chips simultaneously activated	32	8	2
Normal mode active power (t_{RC} = 110 ns)	17.45 W	4.85 W	1.15 W
Chip/ac[a]	17.4/0.05 W	4.8/0.05 W	1.1/0.05 W
Page mode active power (t_{PC} = 40 ns)	15.63 W	4.93 W	1.23 W
Chip/ac[a]	15.5/0.13 W	4.8/0.13 W	1.1/0.13 W
Data-retention power for 8 MB subsystem	108.8 mW	17.6 mW	11.2 mW
Memory density on memory board (ratio)	1	4	8

[a] $32 \times 1/2C(\Delta V)^2 f$ for TTL interface: $C = 20$ pF for 8 MB subsystem, $\Delta V = V_{DD} - V_T = 4$ V for 5 V V_{DD}, $f = 9$ MHz or 25 MHz.

Table 7.2. 5 V DRAM chip and package advancement [7.5]

Memory capacity		**1 Mb** (1 Mword × 1 bit)	**4 Mb** (1 Mword × 4 bit)	**16 Mb** (1 Mword × 16 bit)
Power supply, V_{DD}		5 V	5 V	5 V (internal 3.3 V)
Active power	Normal mode t_{RC} =110 ns	545 mW max.	605 mW max.	550 mW max.
	Page mode, t_{pc} =40 ns	484 mW max.	605 mW max.	550 mW max.
Data-retention power		1.7 mW max.	1.1 mW max.	2.8 mW max.
Refresh cycle, n		512	1024	4096
Feature size		1.3 μm	0.8 μm	0.5 μm
Chip size (ratio)		1	1.5	2.5
Package size		8.5 mm×16.9 mm (1.0)	8.5 mm×16.9 mm (1.0)	11.2 mm×27.1 mm (2.1)

and the data output (D_{out}) pin are common, enables small-package design by halving the number of data pins, which increases with the wide-bit I/O configuration. Furthermore, remarkable progress in package structure can accommodate ever-more enlarged chips at every successive generation in a small package. Typical examples are the LOC (Lead-On-Chip) package and the CSP (Chip-Size Package), as discussed in Chap. 2. Figure 7.5 shows the resultant small packages [7.2] achieved despite the wide-bit I/O configuration and a large memory capacity. The memory density on a memory board is almost doubled when 4 Mb chips are replaced by 16 Mb chips, as shown in Tables 7.1 and 7.2, by reducing the parasitic capacitances of data-bus lines.

7.2.3 The Low-Voltage Data-Bus Interface

Recently, further ac power reduction has been strongly required because the ever-increasing data-throughput of memory subsystems, which results from rapidly improved microprocessor performance, increases ac power. The reduction is especially vital for the DRAM column mode which offers a data-throughput that is inherently higher than that of the normal read–write mode. The details are given in Chap. 6.

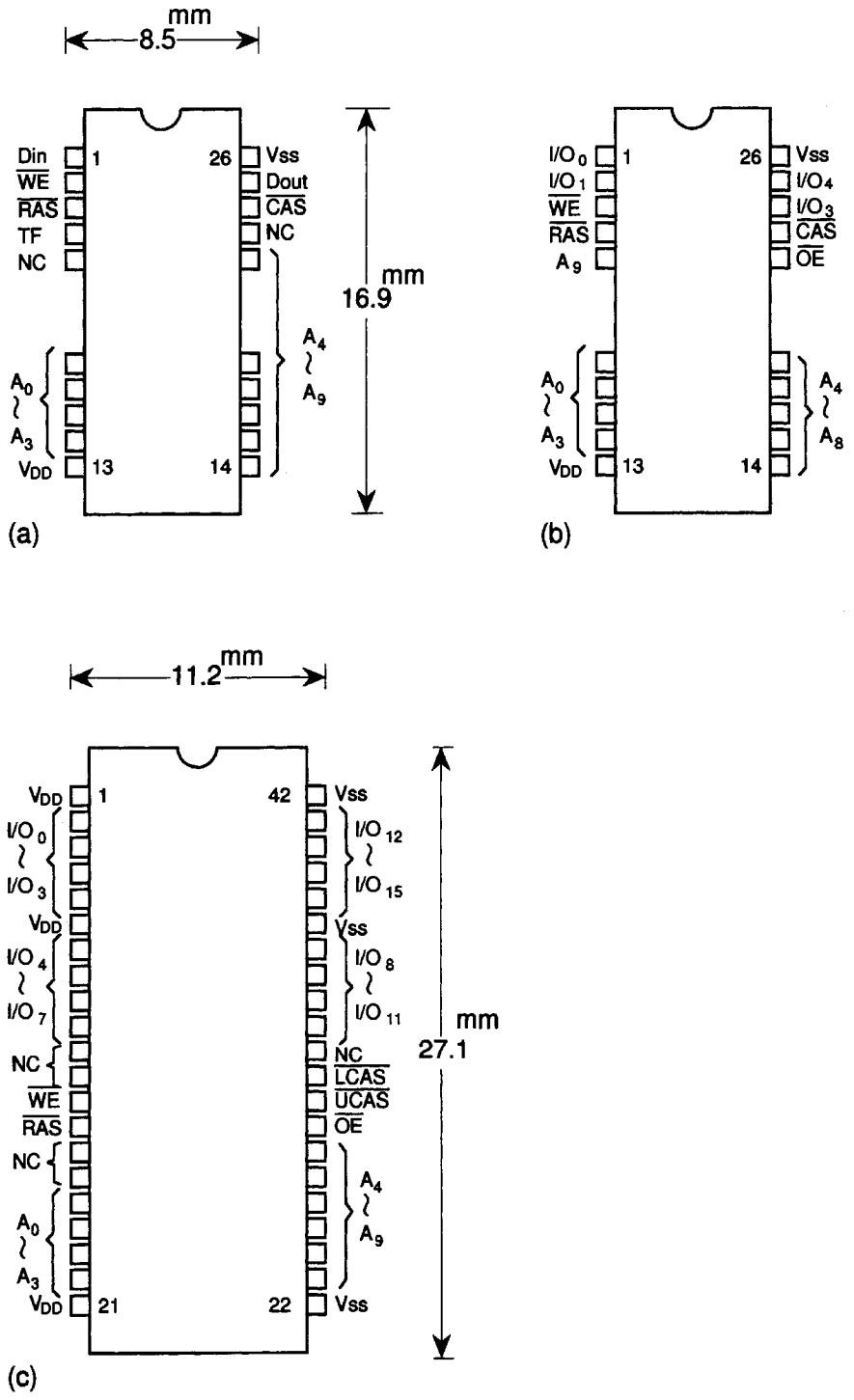


Fig. 7.5. The top view of an SOJ package and its pin assignment [7.2]. A_0–A_9, addresses, D_{in}, data input; D_{out}, data output; I/O_0–I/O_{15}, data input/output; $\overline{RAS}$, $\overline{CAS}$, $\overline{UCAS}$, $\overline{LCAS}$, clocks; $\overline{WE}$, write enable; $\overline{OE}$, output enable; NC, no connection; V_{DD}, power supply; V_{SS}, ground; TF, test function; SOJ, Small Outline J-Leaded Package. (**a**) 1M×1; (**b**) 1M×4; (**c**) 1M×16

In a traditional address-multiplexed DRAM chip, as shown in Fig. 7.6, a selected word line, WL, is activated after strobing the corresponding set of row addresses with an external clock called the Row Address Strobe ($\overline{\text{RAS}}$). As a result, all of the stored information in the memory cells along the selected WL is read on the corresponding data lines, and then amplified by the corresponding amplifiers. Each amplified signal is held on the data line as read information. Any read information on the selected data line can be read out to the data output (D_{out}) pin by strobing the corresponding set of column addresses with another clock called the Column Address Strobe ($\overline{\text{CAS}}$). A different word line is activated by the succeeding $\overline{\text{RAS}}$ strobing of the corresponding row addresses. This is the so-called normal mode, which always entails word-line activations. Here, during a $\overline{\text{RAS}}$ activation ($\overline{\text{RAS}}$: Low), any amplified signal held on the data line can be successively read out to the D_{out} pin by successive data-line selection (column mode) that is performed only by $\overline{\text{CAS}}$ and address activations. The column mode ($\overline{\text{CAS}}$) cycle (t_{PC}) is inherently much faster than the normal mode ($\overline{\text{RAS}}$) cycle (t_{RC}). The difference in speed originates from the fact that a $\overline{\text{CAS}}$ cycle is

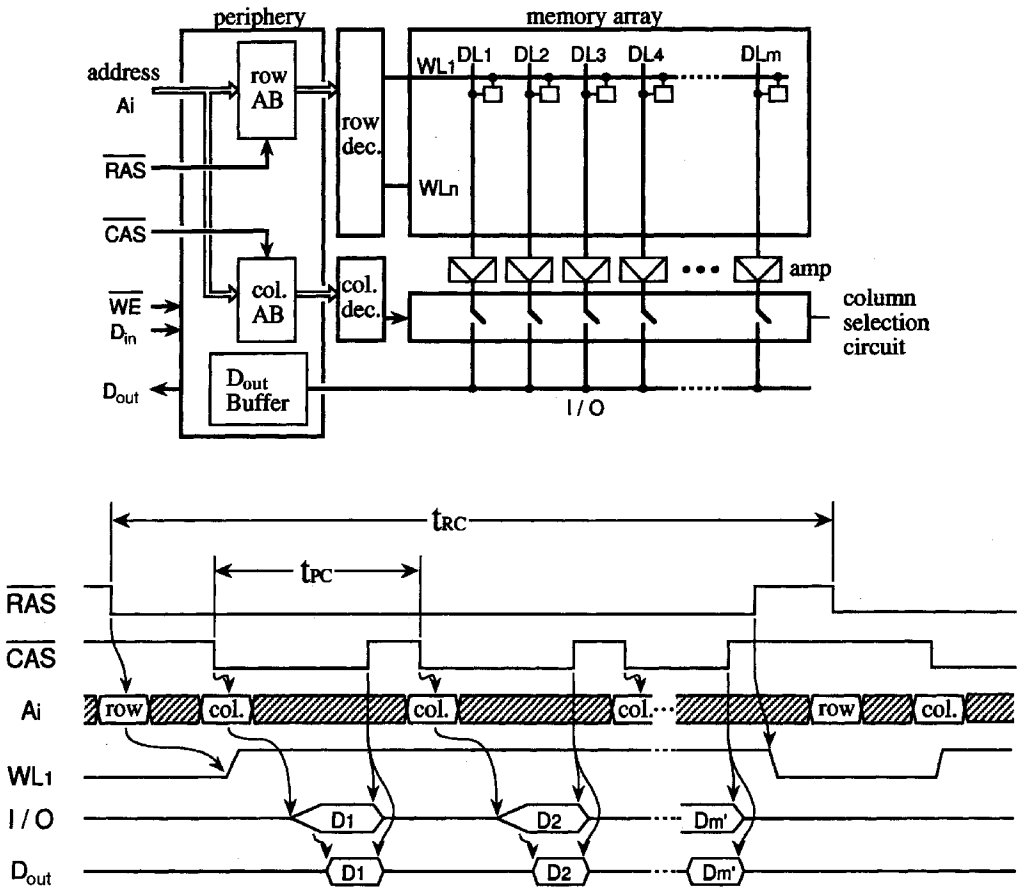


Fig. 7.6. The concept of DRAM chip operation [7.2]. AB, address buffer; t_{RCmin} is for $m' = 1$

accomplished only by column circuit operation, while a $\overline{\text{RAS}}$ cycle entails both row and column circuit operation. The inherently slow operation of row circuits, which include relatively heavy loads such as the word and data lines that are responsible for the large physical size of the memory array, makes the difference larger. Thus, the minimum t_{PC} is about 40 ns, while the minimum t_{RC} is about 110 ns in a modern DRAM chip. The page mode, nibble mode, and static column mode are typical examples of the column mode, as discussed in Chap. 3 and 6. Note that for a 4 bit I/O configuration, for example, a memory array is divided into four subarrays which are simultaneously accessible. Thus, a 4-bit parallel read operation from four subarrays can be performed.

We will now evaluate the ac capacitive power of data-bus lines at the minimum column cycle. For the 5 V TTL (Transistor Transistor Logic) shown in Fig. 7.7, each bus line power is expressed by

$$P = \frac{1}{2}C(\Delta V)^2 f \,, \tag{7.1}$$

where C, ΔV, and f are the loading capacitance, voltage swing, and frequency, respectively. It is found that the ac power of an 8 Mb, 32-bit data-bus-line subsystem becomes relatively large as the width of the data I/O increases, despite a still small portion of the total power, as shown in Table 7.1. In fact, the power is as small as 0.13 W even for a 40 ns page mode cycle, comprising 10% of the total power of a 16 Mb chip subsystem. In this subsystem, two

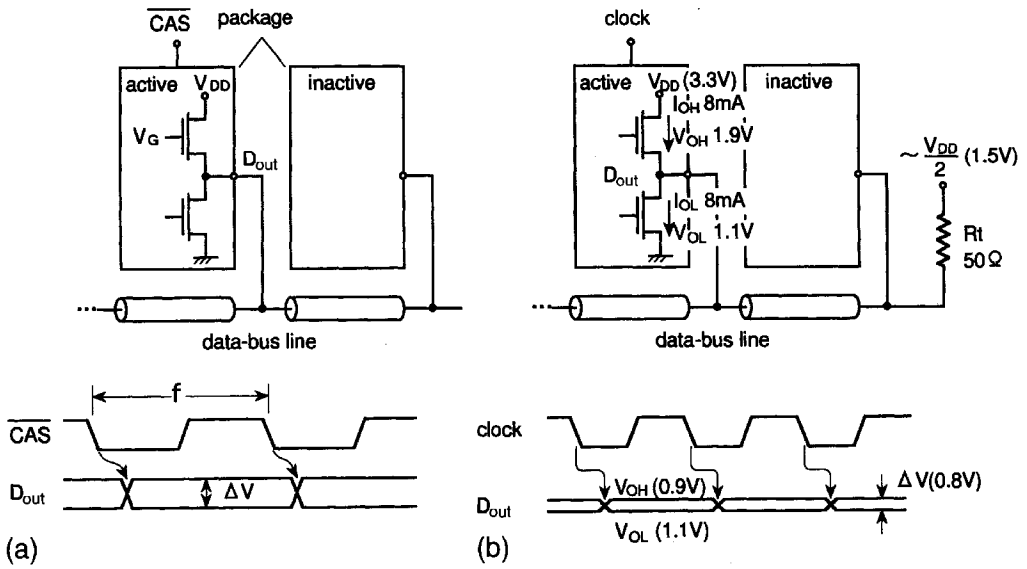


Fig. 7.7. Typical interfaces [7.11]. (**a**) TTL ($V_{\text{DD}} = 5$ V), LVTTL ($V_{\text{DD}} = 3.3$ V); (**b**) CTT (Center Tapped Terminated)

$$\begin{aligned} \Delta V &= V_{\text{DD}} - V_{\text{T}} = 4\,\text{V} && \text{for TTL } (V_{\text{G}} = V_{\text{DD}}) \\ &= V_{\text{DD}} = 3.3\,\text{V} && \text{for LVTTL } (V_{\text{G}} \geq V_{\text{DD}} + V_{\text{T}}) \\ &= 0.8\,\text{V} && \text{for CTT} \end{aligned}$$

16 Mb chips are connected to each data-bus line, with a resulting capacitance of 20 pF, and ΔV is 4 V ($= V_{DD} - V_T$, where V_T is the transistor threshold voltage), since the gate voltage of the D_{out} transistor is usually V_{DD}. For a large system in which many chips are connected on each data-bus line, however, the ac power component is more prominent because of the large capacitance. For the example of a 16 Mb chip subsystem, it occupies about 37% of the total power for $C = 100$ pF. Low-V_{DD} operation, enabling low-voltage interfaces, reduces not only the chip power, but also the ac power, as shown in Table 7.3. The LVTTL (Low Voltage TTL) reduces subsystem power by 34% with reduced voltages of 3.3 V V_{DD} and 3.3 V ΔV. However, even LVTTL can no longer manage power increase due to more enhanced cycle time. The chip power also increases, since the D_{out} transistor, which is enlarged according to the high-speed driving of the bus capacitance, increases the internal node capacitance of the chip.

Recently, new interfaces featuring small-amplitude impedance-matched bus lines have been proposed, as described in Chap. 6. A small amplitude minimizes the ac capacitive power, while impedance matching allows extremely high speed transmission. These new interfaces enable high-speed column

Table 7.3. The impact of low-voltage 16 Mb DRAM operation on subsystem power [7.2][a]

Power		5 V **TTL chip** (1 Mword × 16 bit)	3.3 V **LVTTL chip** (1 Mword × 16 bit)
Active chip power	Normal mode, $t_{RC} = 110$ ns	550 mW max.	360 mW max.
	Page mode, $t_{pc} = 40$ ns	550 mW max.	360 mW max.
Data-retention chip power		2.8 mW max.	1.4 mW max
Normal mode power of 8 MB subsystem t_{RC}=110 ns		1.15 W	0.75 W
Chip/ac		1.1/0.05 W	0.72/0.03 W
Page mode power of 8 MB subsystem t_{pc}=40 ns		1.23 W [1]	0.81 W[0.66]
Chip/ac		1.1/0.13 W	0.72/0.09 W
Data-retention power of 8 MB subsystem		11.2 mW	5.6 mW

[a]Feature size=0.5 µm, refresh cycle=4096, 32-bit data buses, C =20 pF.

cycles of 2–20 ns with low power. This is when high-speed DRAM chip operation is achieved using pipeline column-circuit operation synchronous to an external clock. The power consumption at each CTT interface, as shown in Fig. 7.7, is difficult to express because of the formation of a complicated transmission line [7.11, 7.12]. However, it is roughly estimated, by using a lump circuit approximation, as

$$P = \{I_{\mathrm{OL}} V_{\mathrm{OL}} + R_{\mathrm{T}} I_{\mathrm{OL}}^2\}(1 - \mathrm{duty}) + \{I_{\mathrm{OH}}(V_{\mathrm{DD}} - V_{\mathrm{OH}}) + R_{\mathrm{T}} I_{\mathrm{OH}}^2\}\mathrm{duty} + \frac{1}{2} C(\Delta V)^2 f \,, \tag{7.2}$$

where I_{OL} and I_{OH} are currents for the low signal voltage (V_{OL}) and the high signal voltage (V_{OH}), respectively, and R_{T} is a termination resistance. The signal swing ΔV is 0.8 V for 8 mA I_{OL}, 8 mA I_{OH}, and 50 Ω R_{T}. The total power consumed at 32 bus lines is about 500 mW even for 100 pF and a 10 ns cycle, as shown in Fig. 7.8. Thus, the interface provides low power with high speed compared with TTL and LVTTL, even in the high-speed region of a less than 40 ns cycle. Note that even for new interfaces the chip-power component dominates the total power. This is because it increases due to additional on-chip interface circuits, while the interface power component is still low enough in the high-speed region. With regard to the data-retention power of the memory subsystem, the main determinant is also the chip power component. A wide-bit configuration combined with low-power chip technology reduces the data-retention power, as shown in Table 7.1.

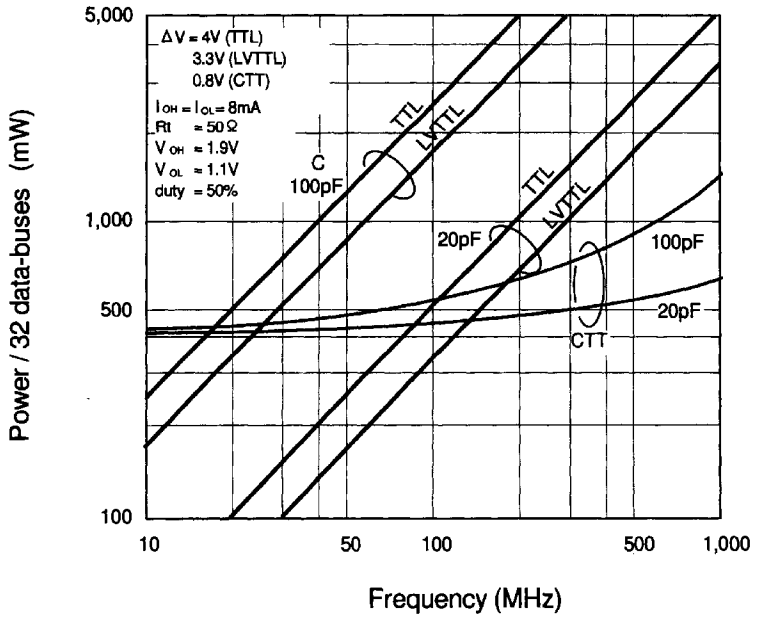


Fig. 7.8. Interface power for various interfaces [7.2]

7.3 Sources of Power Dissipation in the RAM Chip

7.3.1 Active Power Sources

Figure 7.9 shows a simplified memory chip architecture [7.1] for investigating its power dissipation. The chip comprises three major blocks of power sources: the memory cell array, the decoders (row and column), and the periphery. Note that all of the m cells on one word line are simultaneously activated in this logical array model. A unified active power equation for modern CMOS DRAMs and SRAMs is approximately given by

$$P = V_{\mathrm{DD}} I_{\mathrm{DD}} \ , \tag{7.3}$$

$$I_{\mathrm{DD}} = m i_{\mathrm{act}} + m(n-1) i_{\mathrm{hld}} + (n+m) C_{\mathrm{DE}} V_{\mathrm{INT}} f + C_{\mathrm{PT}} V_{\mathrm{INT}} f + I_{\mathrm{DCP}} \ , \tag{7.4}$$

for normal read cycle, where V_{DD} is an external supply voltage, I_{DD} is the current of V_{DD}, i_{act} is the effective current of the active or selected cells, i_{hld} is the effective data-retention current of inactive or non-selected cells, C_{DE} is the output node capacitance of each decoder, V_{INT} is an internal supply voltage, C_{PT} is the total capacitance of the CMOS logic and driving circuits in the periphery, I_{DCP} is the total static (dc) or quasi-static current of the periphery, and f is the operating frequency ($= 1/t_{\mathrm{RC}}$, where t_{RC} is the cycle time). The major sources of I_{DCP} are the column circuitry and the differential amplifiers on the I/O lines. Other contributors to I_{DCP} are the refresh-related circuits and the on-chip voltage converters essential to DRAM operation;

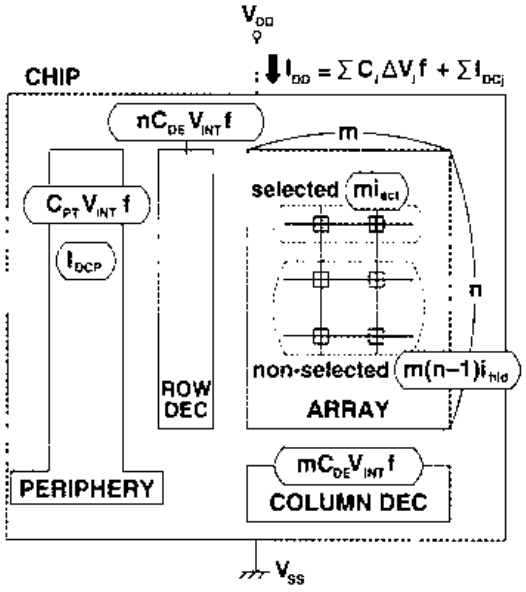


Fig. 7.9. Sources of power dissipation in a RAM chip [7.1]

these include a substrate back-bias generator, a voltage down-converter, a voltage up-converter, a voltage reference circuit, and a half-V_{DD} generator. The dc currents of these circuits are virtually independent of the operating frequency. Hence, for high-speed operation, I_{DCP} becomes relatively small compared with other ac current components. At high frequencies, the data-retention current, $m(n-1)i_{hld}$, is negligible – as shown by the small cell leakage current and small periphery current necessary for the refresh operation in DRAMs. In SRAMs, i_{hld} is also quite small, as described later. The decoder charging current, $(n+m)C_{DE}V_{INT}f$, is also negligibly small in modern RAMs that incorporate a CMOS NAND decoder [7.13], because only one out of the n or m (nodes) is charged at every selection: $n+m=2$. The active current for both DRAMs and SRAMs is shown in Fig. 7.10. Note that I_{DD} increases with increasing memory capacity; that is, with increased m and n.

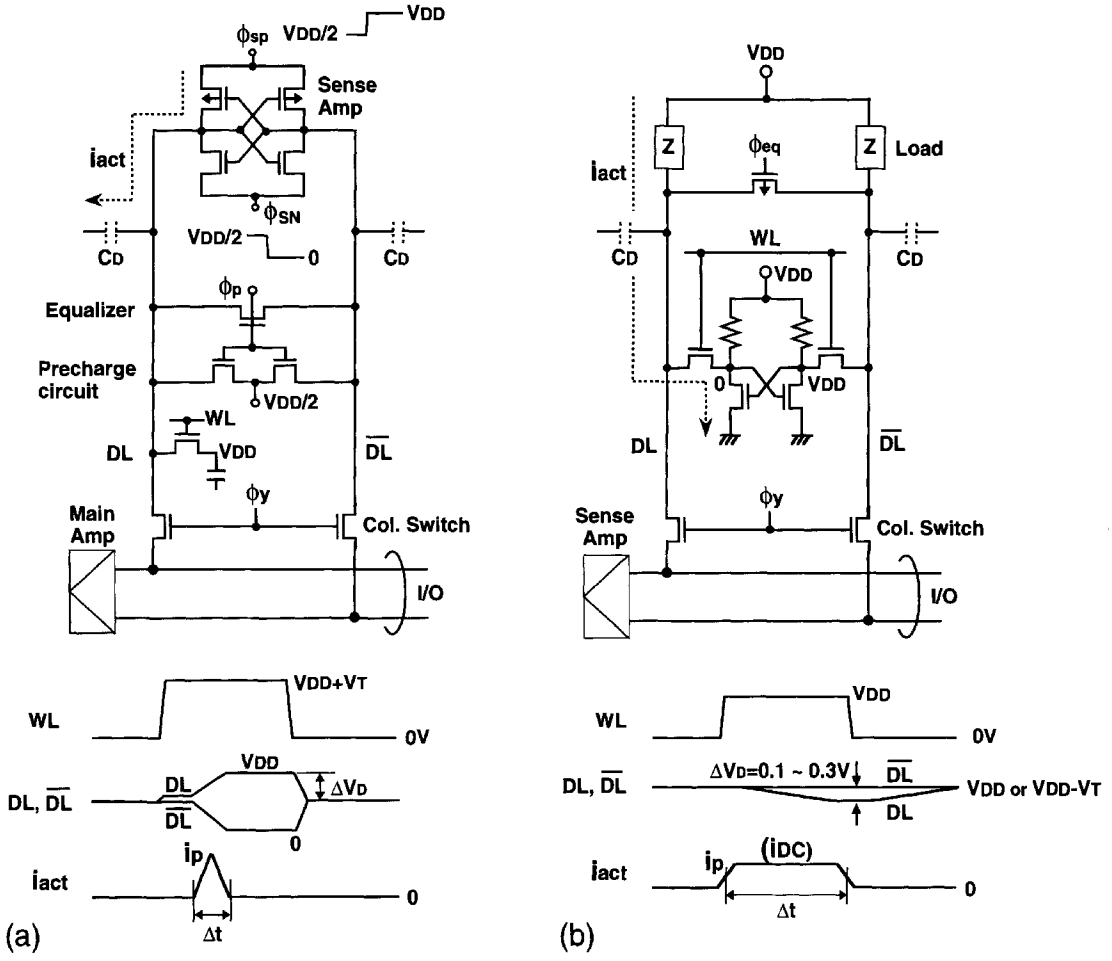


Fig. 7.10. A comparison between the DRAM (**a**) and SRAM (**b**) read operations [7.1]. A half-V_{DD} precharging scheme is assumed in the DRAMs

DRAM. The destructive readout characteristics of a DRAM cell necessitate successive operations of amplification and restoration for a selected cell on every data line. This is performed by a latch-type CMOS sense amplifier on each data line. Consequently, the data line is charged and discharged with a large voltage swing of ΔV_{D} (usually 1.5–2.5 V) and with a charging current of $C_{\mathrm{D}}\Delta V_{\mathrm{D}}$ where C_{D} is the data-line capacitance. Hence, (7.4) is simplified to

$$I_{\mathrm{DD}} \simeq [mC_{\mathrm{D}}\Delta V_{\mathrm{D}} + C_{\mathrm{PT}}V_{\mathrm{INT}}]f + I_{\mathrm{DCP}} \ . \tag{7.5}$$

Equations (7.3) and (7.5) show that the following issues are the keys to reducing the active power for a fixed cycle time: reducing the charging capacitance (mC_{D} and C_{PT}), lowering the external and internal voltages (V_{DD}, V_{INT}, and ΔV_{D}), and reducing the static current (I_{DCP}). In particular, emphasis must be placed on the reduction of the total data-line dissipation charge ($mC_{\mathrm{D}}\Delta V_{\mathrm{D}}$), since it dominates the total active power, as described in Chap. 3. However, the dissipation charge must be reduced while maintaining an acceptable signal-to-noise (S/N) ratio, since they are closely related to each other. The S/N ratio is an extremely important issue for stable operation, as discussed in Chap. 4. This stems from the principle of DRAM cell operation, that the cell signal is not only small, but also read out on the floating data line, which is susceptible to noise. The signal, ν_{s}, is approximately expressed as

$$\nu_{\mathrm{s}} \simeq (C_{\mathrm{S}}/C_{\mathrm{D}})V_{\mathrm{DD}}/2 = (C_{\mathrm{S}}/C_{\mathrm{D}})\Delta V_{\mathrm{D}} = Q_{\mathrm{S}}/C_{\mathrm{D}} \tag{7.6}$$

for the half-V_{DD} precharging scheme ($\Delta V_{\mathrm{D}} = V_{\mathrm{DD}}/2$) described in Chap. 1, where C_{S} and Q_{S} are the cell capacitance and the cell signal charge, respectively. It is obvious from (7.5) and (7.6) that reducing C_{D} is effective for both reducing I_{DD} and increasing ν_{s}, while reducing ΔV_{D} degrades ν_{s}, despite the reduction in I_{DD}. This implies the importance of increasing C_{S} and/or decreasing noise instead.

SRAM. The non-destructive readout characteristics of SRAM never require restoration of cell data, allowing the elimination of a sense amplifier on each data line. To obtain a fast read, the cell signal on the data line is made as small as possible, and the resulting small signal is transmitted to the common I/O line through the column switch, so as to be amplified by a sense amplifier. Since the cell signal is developed as a result of the ratio operation of a data-line load and a cell, a ratio current i_{DC} flows along the data line during word-line activation, Δt. Here, the data-line charging current is negligibly small due to a very small ΔV_{D} (usually 0.1–0.3 V), although it is prominent for the write operation, as described later. Thus, (7.4) for the read operation is expressed as

$$I_{\mathrm{DD}} \simeq [mi_{\mathrm{DC}}\Delta t + C_{\mathrm{PT}}V_{\mathrm{INT}}]f + I_{\mathrm{DCP}} \ . \tag{7.7}$$

Table 7.4. A comparsion of determinants of active cell current between DRAMs and SRAMs [7.1] (estimated based on data in *ISSCC Digests*; DRAMs are 1 Mb to 16 Mb and SRAMs are 256 Kb to 4 Mb)

	DRAM	**SRAM**
Number of cells on a data-line pair	256–512	256–1024
C_D (pF)	0.2–0.3	1.0–2.0
v_S (V)	0.1–0.2	0.1–0.3
ΔV_D (V)	1.5–2.5	0.1–0.3
$i_p/i_{DC}(r)$ (μA)	20–50	100–200
Δt (ns)	10–20	5–50
m	2k–8k (128–512[a])	64–128[a]
$mi_p/mi_{DC}(r)$ (mA)	100–160 (6–10[a])	6–25[a]

[a]Partial activation of multidivided word line.

To reduce the active power, the three issues of static current, voltage, and capacitance, as in DRAMs, are vital. However, the static current charge, $mi_{DC}\Delta t$, should be reduced more intensively, because it dominates the total active current, which differs between SRAMs and DRAMs. Obviously, the S/N ratio issue is not so serious as in DRAMs because of the ratio operation.

Eventually, DRAMs and SRAMs have evolved so that they use similar circuit techniques, although emphasis on each of the three issues is different between the two types of RAM, as described in the following sections. To clearly show the state-of-the-art RAM designs, the ranges of cell design parameters for the active current in DRAMs and SRAMs shown in Fig. 7.10 are compared in Table 7.4. Note the peak current, mi_P, which is a good measure of active power. A partial activation scheme of a multidivided word line, as described later, reduces the DRAM current down to the SRAM level, although this is still at an experimental stage for DRAMs.

7.3.2 Data-Retention Power Sources

DRAM. In the data-retention mode, a memory chip cannot be accessed from outside and the data are retained by the refresh operation. The refresh operation is performed by reading the data of the m cells on a word line and restoring them for each of the n word lines in order. Note that n in the logical array in Fig. 7.9 corresponds to the number of refresh cycles in the catalog specification. A current given by (7.5) flows every time m cells are

refreshed at the same time. The frequency f at which the refresh current flows is n/t_{REF}, where t_{REF} is the refresh time of the cells in the retention mode, and could be increased with a reducing junction temperature. Thus, from (7.5), the data-retention current is given by

$$I_{\mathrm{DD}} \simeq [mC_{\mathrm{D}}\Delta V_{\mathrm{D}} + C_{\mathrm{PT}}V_{\mathrm{INT}}](n/t_{\mathrm{REF}}) + I_{\mathrm{DCP}} \,. \tag{7.8}$$

The t_{REF} could be much longer than the t_{REFmax} that is guaranteed in the catalog specification, as explained in Chap. 3. This is because t_{REFmax} is for the active mode operating at the maximum frequency of around 10 MHz, where the cell leakage current is maximized with the highest junction temperature. On the other hand, t_{REF} is for an extremely slow refresh frequency ($n/t_{\mathrm{REF}} \ll 62\,\mathrm{KHz}$), where the current is minimized with the lowest junction temperature. In any event, I_{DCP} becomes relatively large for other ac current components because of the small n/t_{REF}. This implies the necessity of reducing both the ac and the dc components.

SRAM. In low-power CMOS SRAMs, the static cell leakage current, mni_{hld}, is the major source of the retention current because of the negligibly small value of I_{DCP}. The leakage current or retention current has been maintained at a small value by memory cell innovations, as explained in Chap. 1.

7.4 Low-Power DRAM Circuits

This section summarizes low-power DRAM circuits, since the full details are given in Chap. 3.

7.4.1 Active Power Reduction

The power for a fixed cycle time has been gradually decreased in spite of the increase in memory capacity, as shown in Figs. 7.11 and 7.12 [7.1]. This is due to the low-power circuits developed at each generation. For a given memory capacity chip, successive circuit advancements have produced a power reduction equivalent to two to three orders of magnitude over the past decade. Figure 7.12 shows the power dissipation of a 64 Mb DRAM, hypothetically designed with the NMOS circuit of the 64 Kb generation in 1980, compared with the CMOS circuit presented in 1990 [7.5]. Almost the same process and device technologies are assumed. The drastic reduction in power by about two orders of magnitude is due to many sophisticated circuits: partial activation of a multidivided data line and a shared I/O, which reduce mC_{D} in (7.5); a CMOS NAND decoder, which reduces $(n+m)C_{\mathrm{DE}}$; an external supply voltage (V_{DD}) reduction from 5 V to 3.3 V, half-V_{DD} data-line precharge and on-chip voltage down-conversion, which reduce V_{DD}, ΔV_{D}, and V_{INT} in (7.3) and (7.5); and CMOS drivers and pulse operation of the column

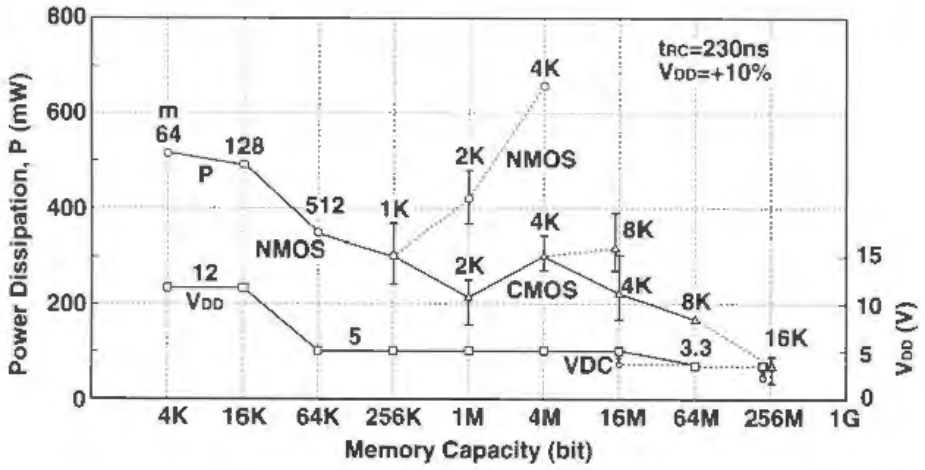


Fig. 7.11. Trends in the power dissipation of commercial DRAM chips [7.1]

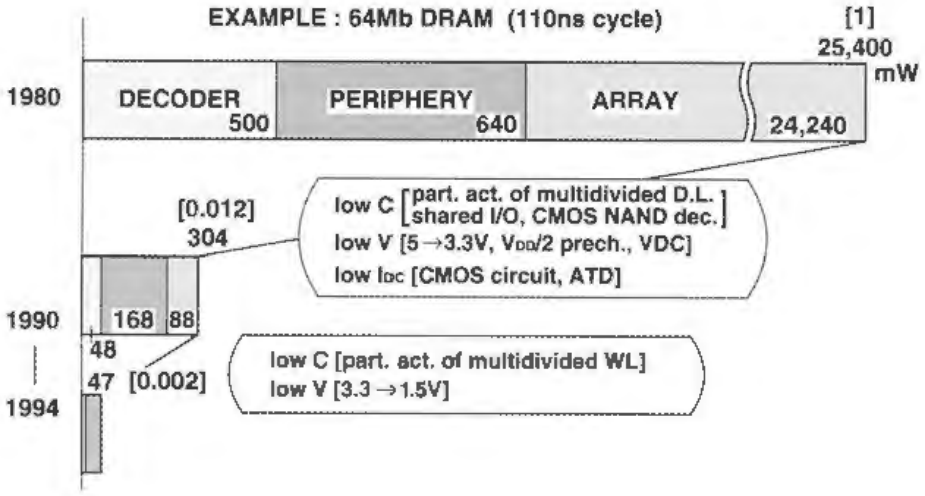


Fig. 7.12. Low-power circuit advancement of DRAM chips over the past decade [7.1]

circuitry and the amplifiers, which reduce the peripheral static current, I_{DCP}. An exploratory 64 Mb DRAM chip, also hypothetically designed using the state-of-the-art technology, could further reduce the active power to about one-tenth that of the 1990 chip. Subthreshold current reduction, explained in Chap. 8, to enable a 1.5 V V_{DD} and partial activation of a multidivided word line, is responsible for this reduction. In this section, these key circuit technologies are reviewed. Here, the details of CMOS circuits are omitted because they are well known, although they are crucial in reducing the power requirement of megabit DRAMs [7.13].

Charging Reduction of Capacitance. The charging capacitance of a data line is reduced by partial activation of multidivided data lines, which is now

a widely accepted technique in commercial 16 Mb and 64 Mb chips. Partial activation of multidivided word lines further reduces the capacitance, although this is still in the experimental stage. Thus, a combination of multidivisions of the data line and the word line minimizes the capacitance.

Partial Activation of Multidivided Data Line. With increasing memory capacity, the ever-increasing number of memory cells connected to one data line causes an increased C_D and thus an increased power requirement, as well as a poor S/N ratio, as expressed by (7.5) and (7.6). One practical solution is to divide a data line into several sections and to activate only one section. In the early days, the number of divisions was increased by the multiple column decoder scheme: this simply increased the number of Y (column) decoders, each of which is placed at each data-line division, by using an additional chip area. However, the division was almost completed in the 16 Mb generation with the combination of a shared SA, a shared I/O, and a shared Y decoder, as shown in Fig. 3.35 [7.14]. Partial activation is performed by activating only one sense amplifier along the data line. Thus, the total data-line charging capacitance, $C_{DT}(= mC_D)$, has been minimized with the help of an increase in the number n (Fig. 3.37). The total data-line dissipating charge, $Q_{DT}(= mC_D\Delta V_D)$, has also been suppressed as much as possible with the help of the ever-reducing operating voltage (see Fig. 3.38).

Partial Activation of Multidivided Word Line. For this multidivision, similar to that of the SRAM, each word line is divided into several subword lines (SWLs) (Fig. 3.53) [7.15]. Any SWL is selected by the coincidence of the selected main word line (MWL) and the selected row line (RX). Only the signals from the cells on the selected SWL are simultaneously amplified by the corresponding amplifiers, allowing partial activation of a word line. Thus, the multidivision method provides less data-line charging capacitance than no division. C_{DT} could be reduced down to about 100 pF at 256 Mb (Fig. 3.38), assuming a C_D of 200 fF, a ΔV_D of 1 V, and that the number of activated data lines is reduced by 1/32. Consequently, the architecture almost halves the chip power of the partial activation of the data line. However, this architecture seemingly does not meet the requirements of the traditional address multiplexing scheme, since it increases the number of row address signals or introduces a speed penalty that is involved in additional selection.

Refresh Time Increase. Even though a reduction of C_D is achieved by the use of the partial activations described above, the reduction of m is never achieved without the help of an increase in the maximum refresh time of the cell, t_{REFmax}. This stems from the need to preserve the refresh-busy rate [7.1, 7.4], γ, expressed by

$$\gamma = t_{RCmin}/(t_{REFmax}/n) = (M/m)(t_{RCmin}/t_{REFmax}) \ , \tag{7.9}$$

where t_{RCmin} and M are the minimum cycle time and memory capacity, respectively. This expresses the percentage of time that is not accessible from

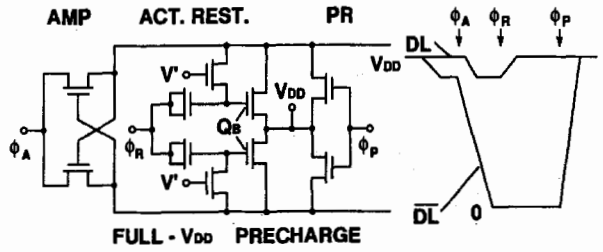
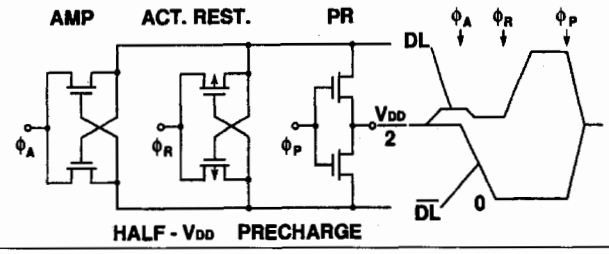
outside of the chip. A smaller γ is preferable, since it involves less conflict between the refresh operation and normal operation. Hence, for a fixed M, it is necessary to maintain mt_{REFmax} to keep γ constant, assuming a fixed t_{RCmin}. This implies a reduced m accompanied by an increased t_{REFmax}. Moreover, to quadruple M, mt_{REFmax} must be quadrupled. This has been achieved by almost doubling both m and t_{REFmax}, which results from compromising the power with the cell leakage current [7.4]. As a result, $n(= M/m)$ has gradually increased with each successive generation, with an increased t_{REFmax} for commercial DRAMs (Fig. 3.16). Alternative choices for n and m, as shown in Fig. 7.11 and 3.38, have been eventually rejected by this compromise. However, it seems difficult to maintain the pace of doubling t_{REFmax} at each generation, because it is determined by the cell leakage current. One solution is the use of a new refreshing scheme [7.15] favorable to partial activation of a multidivided word line. Note that the traditional scheme uses the same n; that is, the same m for both normal and refresh operations. However,the new scheme uses a reduced m for normal operation, which is a determinant of maximum power, while maintaining the same m as in the traditional scheme for the refresh operation, to preserve γ. The resulting power reduction allows an increased t_{REFmax} with a reduced junction temperature.

Operating Reduction in Voltage. The reductions in V_{DD}, from 12 V to 5 V in the 64 Kb generation and then to 3.3 V in the 64 Mb generation, and half-V_{DD} data-line precharge, which has been widely used since the advent of commercial 1 Mb DRAMs, are well known contributors to low power. On-chip voltage down-converters, which are employed in commercial 16 Mb DRAMs to maintain a standard 5 V supply, also contribute to low power. This low-power operation necessitates an improvement in the S/N ratio of the memory cell, as described in Chap. 4.

Half-V_{DD} Data-Line Precharge. An excellent circuit for reducing the array operating current is half-V_{DD} data-line precharge [7.1]. Table 7.5 shows full-V_{DD} precharge with an NMOS sense amplifier (SA), which was popular until the 256 Kb generation, and half-V_{DD} data-line precharge, which is favorable to the CMOS SA. In full-V_{DD} precharge, the voltage difference on the data lines is amplified by applying pulse Φ_{A} and then the resultantly degraded high level is restored to the full-V_{DD} level by applying pulse Φ_{R} to the active restore circuit. In contrast, the restore operation in half-V_{DD} precharge is simply achieved by applying pulse Φ_{R} to the cross-coupled P-MOSFETs. In principle, half-V_{DD} precharge halves the data-line power of full-V_{DD} precharge, with a halved data-line voltage swing. A large spike current caused during the restoring or precharging periods is also halved with less noise generation, allowing a quiet array. More details are given in Chap. 3 (see Fig. 3.58 and 3.59).

The On-Chip Voltage Down-Converter (VDC). For a fixed external supply V_{DD}, a voltage down-conversion scheme combined with scaled-down devices reduces the chip power (Table 2.4). An internal voltage, V_{INT}, reduced by

Table 7.5. A half-V_{DD} precharging scheme [7.1]

	Circuit and timing	Power	Area	Speed
N		1	1	1
C		0.46	0.7	1

a scaling factor ($k > 1$), permits the use of scaled devices in core circuits while maintaining the same electric field. This eventually provides a low power (reduced by $1/k$) with a higher speed and a smaller chip area. This is justified by the negligibly low VDC current (about 3% of the total 16 Mb chip current) and the negligibly small VDC area (less than 1% of the 16 Mb chip area), as shown in Chap. 5. The breakdown voltage of devices in the VDC can be adjusted to a sufficiently high level through slight modifications involving device parameters. The following is a summary, although the detailed design is described in Chap. 5. The keys to designing a VDC are the provision of a stable and accurate output voltage under a rapidly changing load current and provision of an on-chip burn-in capability. Figure 7.13 shows a schematic of a typical VDC [7.16] and the step response for the load current, I_L. The almost fixed output voltage, V_{DL}, is about 3.3 V for a 16 Mb DRAM with $V_{DD} = 5$ V. For accuracy and load current driving capability, it consists of a current-mirror differential amplifier (Q_1–Q_4, Q_S) and an output transistor (Q_6). As shown in Table 7.4, the array current for a DRAM is fairly large compared with that of an SRAM. The peak height is more than 100 mA with a peak width of around 20 ns. Thus, the gate width of Q_6 has to be more than 1000 μm. In order to minimize the output voltage drop ΔV_{DL}, the gate voltage of Q_6, V_G, has to respond quickly when the output goes low. An amplifier current I_S of 2–3 mA enables such a fast response time. A bias current source I_B is needed to clamp the output voltage when the load current becomes almost zero. To ensure enough loop stability with a minimized area and operating current, phase compensation is indispensable.

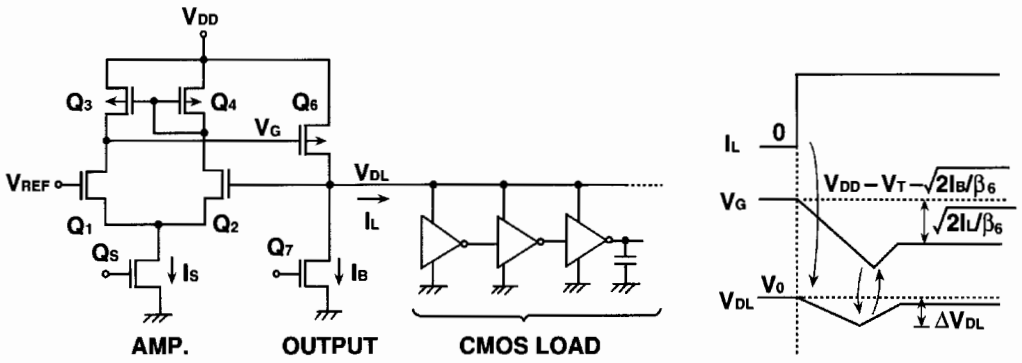


Fig. 7.13. A voltage down-converter [7.1, 7.16]. (**a**) Schematic circuit; (**b**) step response for loading

For stable operation, the reference voltage, V_{REF}, must be accurate over wide variations of V_{DD}, process, and temperature, because the voltage level determines the amount of cell signal charge as well as the speed performance. A band-gap V_{REF} generator and a CMOS V_{REF} generator [7.17] utilizing the threshold voltage difference have been proposed to meet these requirements. A burn-in operation, with the application of a high stress voltage to devices, is indispensable in VLSI production both for reliability testing and for chip screening. For this purpose, the V_{REF} generator is designed to output an increased voltage when V_{DD} is higher than the value for normal operation. Otherwise, the fixed voltage fails to apply a higher stress voltage.

Reduction of dc Current. Figure 7.14 shows the column signal path circuitry [7.1], which is a main source of static current. It consists of a pair of data lines, a column switch, a pair of I/O lines, a load circuit for the I/O lines, and a differential amplifier (the "main amp") for detecting the small signal voltage on the I/O lines. Static (dc) current flows from the I/O line load to the data lines while the column switch is on. Also, the main amplifier consumes dc current for amplifying the signal, because it usually employs differential amplifiers that are similar to those of SRAMs. The current reductions are essential, especially for the wide-bit I/O chip, because of the increased number of column circuitries. The dc currents are shut down with the pulse-operation technique, which will be described in detail in the SRAM section. For example, in the static column mode where a DRAM operates as an SRAM for the column address signals, the column switch and the main amplifier are activated only when address-signal transition occurs. The ATD generates such control pulses. Other main amplifiers, such as current sense amplifiers, that comprise a negative feedback circuit or a current-mirror circuit have also been proposed for DRAM, as explained in Chap. 3. Various low-power amplifiers for SRAM, which are explained later, are also useful for DRAM.

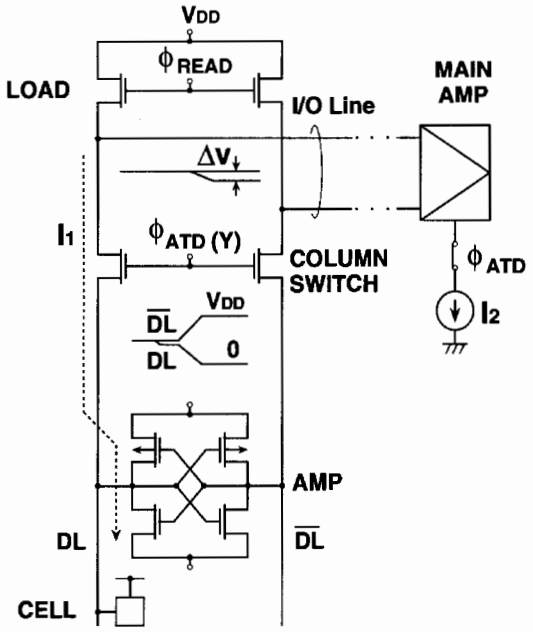


Fig. 7.14. Pulse operation of column signal path circuitry [7.1]. Typical parameters for 16 Mb DRAMs are as follows: $I_1/I_2 = 0.5/1.0$ mA, $\Delta V = 0.5$ V, $C_{I/O} = 2$ pF

7.4.2 Data-Retention Power Reduction

The reduction of both the dc and the ac current components in data-retention mode is a prime concern. Minimizing the power of on-chip voltage converters, such as the VDC, the voltage (V_{DH}) up-converter, the substrate back-bias (V_{BB}) generator, the V_{REF} generator, and the half-V_{DD} generator, reduces the dc current component. Extending the refresh time and reducing the refresh charge reduces the ac current component, by reducing I_{DCP}, $1/t_{REF}$, and $mC_D\Delta V_D$, respectively, in (7.8). Since the low-power generator designs are discussed in detail in Chap. 5, extension of the refresh time is briefly explained here.

- *Refresh Time Extension.* To extend the refresh time (t_{REF}) according to a reduced junction temperature in data-retention mode, a self-refresh control with an on-chip temperature detection circuit and the use of a cell-leakage monitor circuit on the chip [7.1] have been proposed (Fig. 3.71 and 3.72).
- *Refresh Charge Reduction.* One practical way to reduce the refresh charge is to reduce n (increase m) from that for active operation. This effectively reduces the operating frequency for the periphery while keeping the array power constant. Another possible way is to reduce the voltage swing of the data lines in data-retention mode. The resultantly reduced signal charge and an increase in the soft error rate are additional issues with this scheme. A charge recycle refresh scheme [7.1] has been proposed to reduce the data-line dissipating charge. In this scheme, the charges used in one array, which

are conventionally poured out in every cycle, are transferred to another array and used, thus enabling the data-line charging current to be halved.

7.5 Low-Power SRAM Circuits

7.5.1 Active Power Reduction

The partial activation of a multidivided word line and pulse operation of word-line circuitry are typical examples of low-power circuits, which drastically reduce the dc current that dominates the total active current with a decreasing m and Δt in (7.7). Pulse operation of the column/sense circuitry is another example of low-power techniques. Column/sense circuitry inevitably includes differential circuits, which unfortunately consume much dc current to achieve high speed. Therefore, pulse operation is essential in reducing I_{DCP} in (7.7). ATD (Address Transition Detection) plays an important role in the pulse operations for word-line and column/sense circuitry. For column/sense circuitry, lowering the operating voltage while maintaining a high-speed amplification capability for small signals is also critical.

Reduction of the dc Current.

The Partial Activation of Multidivided Word Lines. The multiple row decoder scheme and the double-word-line scheme [7.1], which divides a word line into subword lines, greatly reduces the static current of SRAMs. A more sophisticated word-line division, called DWL (Divided Word Line), adopts a two-stage hierarchical row decoder structure [7.18], as shown in Fig. 7.15. The DWL scheme requires two levels of metal layers; one for a main word line and the other for a data line. The number of subword lines connected to one main word line in the data-line direction is generally four (at most eight), compromising one area allocated to a main row decoder with another area

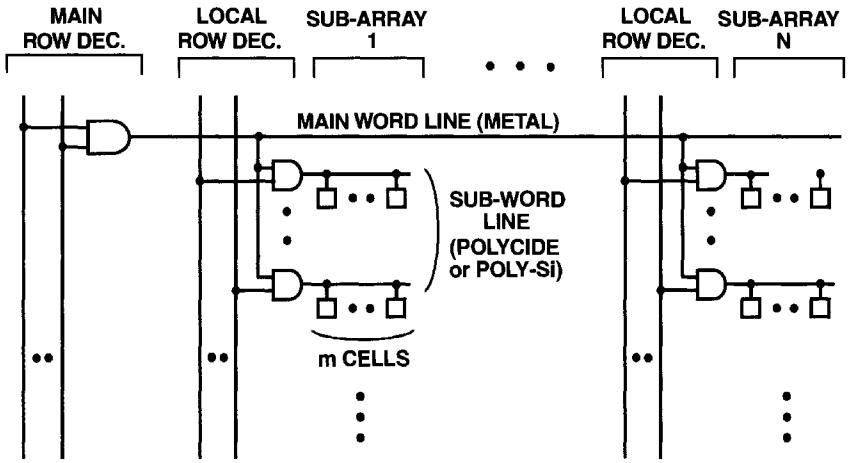


Fig. 7.15. A divided word-line structure (DWL) [7.1, 7.18]

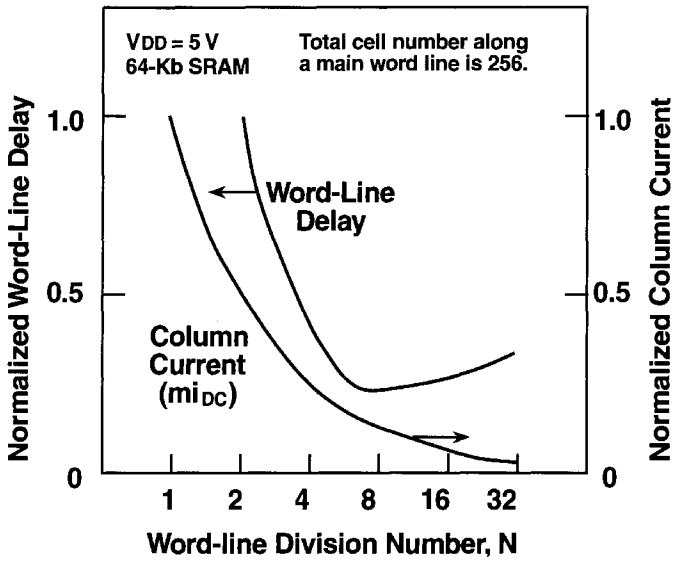


Fig. 7.16. The decrease in the word-line delay and the column current due to a divided word-line structure (DWL) [7.1, 7.18]

allocated to a local row decoder. DWL features two-step decoding to select one subword line, which greatly reduces the capacitance of the address lines to a row decoder and the word-line RC delay. Figure 7.16 [7.18] shows that the column current (mi_{DC}) and the word-line delay decrease as the number of word-line divisions increases. The DWL scheme has been used in most high-density SRAMs of 1 Mb and greater. In a recent 16 Mb SRAM [7.8], a word line was divided into 32 subword lines: $N = 32$ and $m = 256$. The cell current was reduced to one thirty-second of its original level with the DWL scheme. However, this scheme eventually results in a long main word line that has a load capacitance that increases due to an increase in the number of local row decoders. This is because the word-line division number (N) must increase while m (the number of cells selected simultaneously) is kept small. To overcome this problem, two approaches [7.1] have been proposed: a combination of a multiple row decoder and a DWL, and a three-stage hierarchical row decoder scheme. An experimental 4 Mb SRAM with a three-stage hierarchical decoder reveals a reduction in the total capacitance in the decoding pass of 30% and a reduction in delay of 20% compared with the DWL scheme.

The Pulse Operation of Word-Line Circuitry. The duration of the active duty cycle can be shortened by pulsing the word line for the minimum time required for reading and writing in a cell array, as shown in Fig. 7.17 [7.19]. This reduces the power by the duty ratio of the pulse duration to the cycle time. The word-activating pulse, ϕ_X, is obtained by lengthening the original ATD pulse, described shortly, sufficiently to build up the data-line signal and latch the amplified signal by ϕ_L. This scheme is usually employed with a pulsed sense amplifier and a latch circuit, as shown below in Fig. 7.22.

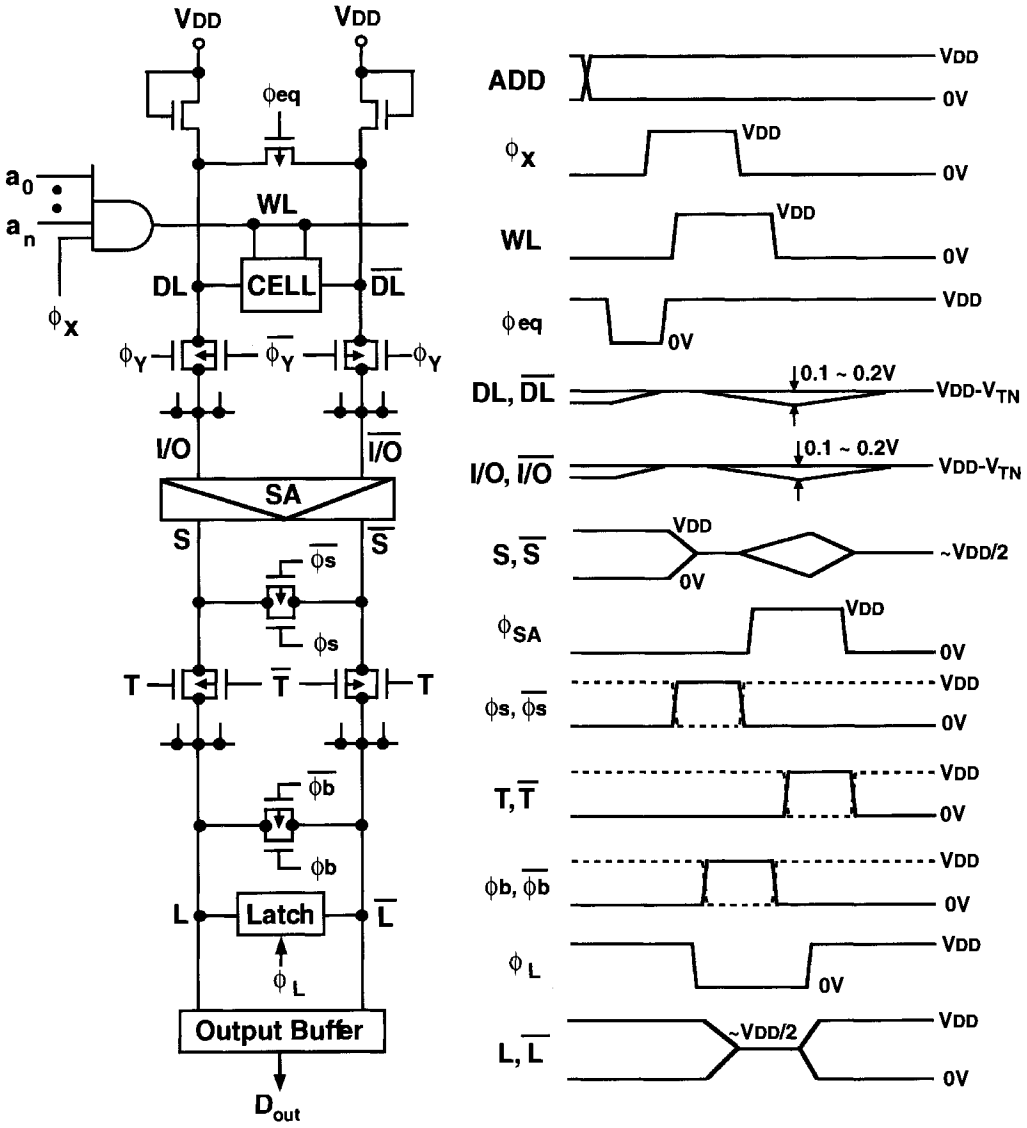


Fig. 7.17. A simplified view of the pulsed operation of a word line, a sense amplifer (SA), and a latch circuit [7.1, 7.19]

One of the data-line signal pairs (DL, $\overline{\text{DL}}$) is selected by ϕ_y and $\overline{\phi_y}$, and transmitted to I/O and $\overline{\text{I/O}}$, respectively. The I/O signals are amplified by a sense amplifier SA. The amplified S and $\overline{\text{S}}$ signals from a subarray are selected by signals T and $\overline{\text{T}}$ and then are transmitted to the latch, where the signals are latched to keep the data output valid after the word line and sense amplifier are inactivated. An on-chip pulse-generating scheme using ATD first appeared in SRAMs, and then in DRAMs. An ATD circuit comprises delay circuits and an exclusive OR circuit, as shown in Figs. 7.18a,b [7.1]. An ATD pulse $\phi(a_i)$ is generated by detecting "L" to "H" or "H" to "L" transitions of any input address signal a_i, as shown in Fig. 7.18c. All of the ATD pulses generated from all of the address input transitions are summed

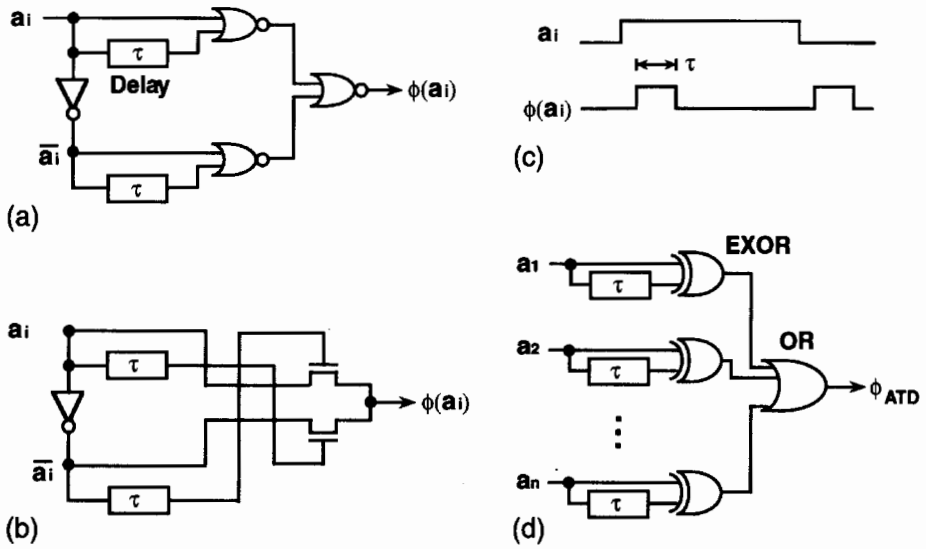


Fig. 7.18. Address transition detection circuits [7.1]. (**a, b**) ATD pulse-generating circuits; (**c**) an ATD pulse waveform; (**d**) a summation circuit of all ATD pulses generated from all address transitions

up to one pulse, ϕ_{ATD}, as shown in Fig. 7.18d. This summation pulse is usually stretched out with a delay circuit and used to reduce power or speed up signal propagation.

An additional dc current in the write cycle, as shown in Fig. 7.19 [7.1], is another concern. An accurate array current expression for the read cycle, exclusive of the peripheral ac and dc currents in (7.7), is given by

$$I_{\mathrm{DDA}}(\mathrm{r}) = [mi_{\mathrm{DC}}(\mathrm{r})\Delta t + mC_{\mathrm{D}}\Delta V_{\mathrm{r}}]f \ , \tag{7.10}$$

and a corresponding current for the write cycle is given by

$$I_{\mathrm{DDA}}(\mathrm{w}) = [(m-p)i_{\mathrm{DC}}(\mathrm{r})\Delta t + pi_{\mathrm{DC}}(\mathrm{w})\Delta t + pC_{\mathrm{D}}\Delta V_{\mathrm{w}}]f \ , \tag{7.11}$$

where $i_{\mathrm{DC}}(\mathrm{r})$ and $i_{\mathrm{DC}}(\mathrm{w})$ are data-line static currents in the read and write cycles; ΔV_{r} and ΔV_{w} are data-line voltage swings in the read and write cycles, respectively; and p is the number of data which are simultaneously written into cells. In a typical 5 V 4 Mb SRAM, $I_{\mathrm{DDA}}(\mathrm{r})$ and $I_{\mathrm{DDA}}(\mathrm{w})$ are 6.4 mA and 10.4 mA, respectively, assuming $i_{\mathrm{DC}}(\mathrm{r}) = 100\,\mu\mathrm{A}$, $i_{\mathrm{D}}C(\mathrm{w}) = 1.0\,\mu\mathrm{A}$, $\Delta V_{\mathrm{r}} = 0.2\,\mathrm{V}$, $\Delta V_{\mathrm{w}} \simeq V_{\mathrm{DD}}$, $C_{\mathrm{D}} = 1\,\mathrm{pF}$, $m = 128$, $p = 8$, $\Delta t = 30\,\mathrm{ns}$, and $f = 10\,\mathrm{MHz}$. Inherently large $pi_{\mathrm{DC}}(\mathrm{w})$ and $pC_{\mathrm{D}}\Delta V_{\mathrm{w}}$ make $I_{\mathrm{DDA}}(\mathrm{w})$ larger than $I_{\mathrm{DDA}}(\mathrm{r})$, especially for wide-bit SRAMs, which make p large. To reduce $I_{\mathrm{DDA}}(\mathrm{w})$, both $i_{\mathrm{DC}}(\mathrm{w})$ and C_{D} must be decreased. To reduce $i_{\mathrm{DC}}(\mathrm{w})$, the variable impedance load [7.1] makes all the data-line load impedances high in the write cycle by cutting off Q_3 and Q_4 with the write enable signal, $\overline{\mathrm{WE}}$, as shown in Fig. 7.19. In some SRAMs the loads are entirely cut off during the write cycle to stop any dc current. C_{D} is reduced by partially

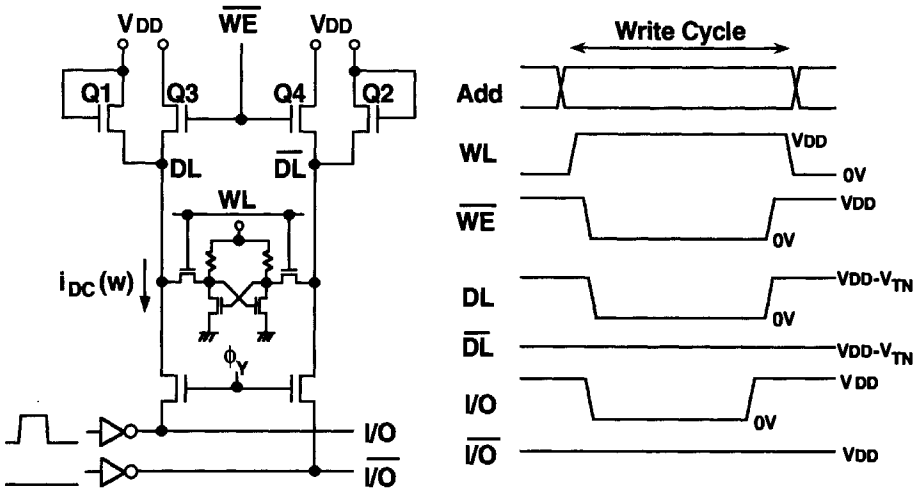


Fig. 7.19. A simplified diagram of data-line load control with a write enable signal $\overline{\text{WE}}$ (variable impedance data-line load) [7.1]

activating the data lines that are divided into two or more portions, just as in DRAMs.

Pulse Operation of Column/Sense Circuitry. A dc current of 1–5 mA flows in a sense amplifier on the I/O line. The power dissipation becomes the larger portion of the total chip power as the number of I/O lines increases to obtain higher data throughput for high-speed processors. There have been many low-power sense amplifiers for SRAMs [7.7], which can be categorized into current-mirror sense amplifiers, latch-type sense amplifiers, and current sense amplifiers. Each amplifier has advantages and disadvantages in terms of meeting the target specifications of SRAMs. They all operate by using ATD pulses.

Current-mirror sense amplifiers have been widely used in SRAMs, but they consume a lot of power and have a relatively low voltage gain. Figure 7.20 shows the switching scheme of well-known current-mirror sense amplifiers [7.19]. Two amplifiers are serially connected to obtain a full supply-voltage swing output, since one stage of the amplifier does not provide enough gain for a full swing. A positive pulse, ϕ_{SA}, activates the sense amplifiers for just long enough to amplify the small input signal; then the amplified output is latched. Otherwise, a large current continues to flow. Hence, the switching scheme combined with a pulsed word-line scheme, reduces the power, especially at relatively low frequencies, as shown in Fig. 7.21 [7.1]. Further current reduction is gained by sense amplifier current control, which switches the current to the minimum level required for maintaining data from a high level that is only needed during amplification.

Latch-type sense amplifiers such as PMOS cross-coupled amplifiers [7.7, 7.20] as shown in Fig. 7.22, greatly reduce the dc current after amplification and latching, because of the small current consumption after latching. More-

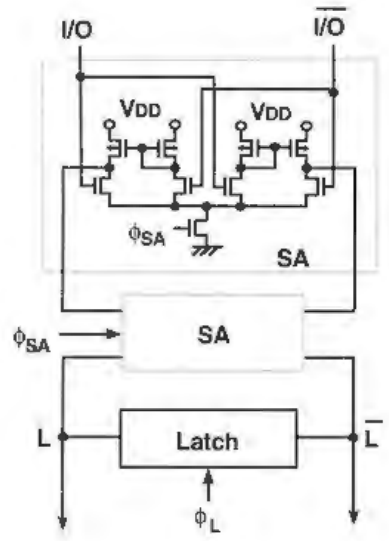


Fig. 7.20. A current-mirror sense amplifier [7.1, 7.19]

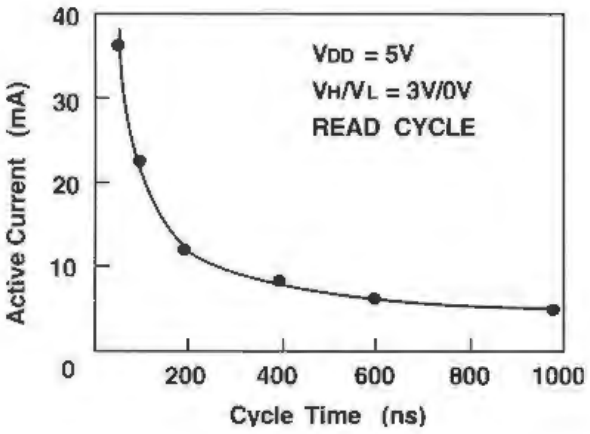


Fig. 7.21. The measured active current in a 4 Mb CMOS SRAM with pulse word-line and pulse sense amplifier control [7.1]. V_H and V_L are high and low input voltage levels, respectively

over, they provide a nearly full supply-voltage swing, with positive feedback of outputs to PMOSFETs. Therefore, they will become a strong candidate for wide-bit SRAMs in the future. Amplifier (a) is a conventional latch-type amplifier which is commonly used in DRAMs. A small signal voltage close to V_{DD} from data lines is amplified to V_{DD} or 0 V, allowing a full-V_{DD} swing at the inputs. Thus, the recovery time of the inputs (i.e. the data line and the I/O line) is fairly long, and the charging/discharging power is quite large. Because of the non-destructive features of the SRAM cell, a full-V_{DD} swing is unnecessary. Amplifiers (b)–(d) solve the problem with input connections to the gates of the MOSFETs. The current in those amplifiers is less than one-fifth of that in a current-mirror amplifier, despite affording very fast sensing speed. However, latch-type amplifiers require much more accurate

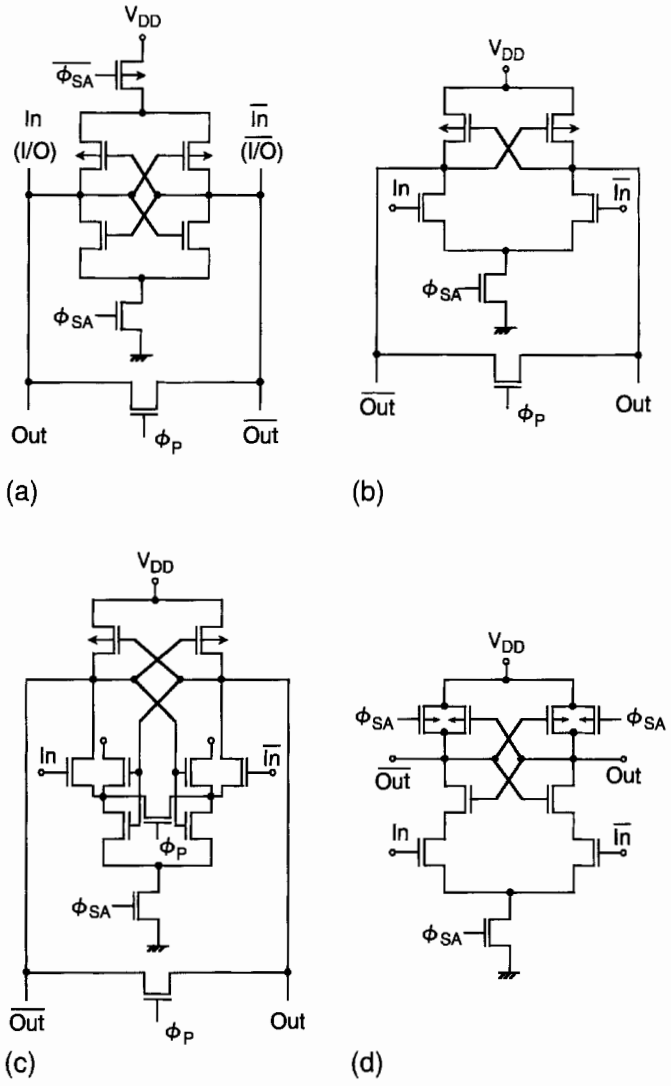


Fig. 7.22a–d. Latch-type sense amplifiers [7.7, 7.20]

timing controls for stable operation. They must be activated after an input signal has fully been developed. In addition, equalizers must equilibrate the paired outputs of the amplifier before amplification.

Current sense amplifiers [7.7, 7.21–7.24] permit a small voltage swing on the data line, reducing the time for changing the data-line voltage as well as the data-line power dissipation. Figure 7.23 shows a current amplifier connected to data lines [7.21]. It consists of four equal-sized PMOSFETs (Q_1–Q_4) in a cross-coupled configuration. The amplifier is selected by turning on Q_3 and Q_4. Currents (I_0's) then flow through Q_1 and Q_3, and Q_2 and Q_4. The drains of Q_3 and Q_4 are connected to I/O lines which are close to ground level. This means that these FETs operate in saturation. The data-line loads (Z) are low impedances to ensure that the data lines are always close to V_{DD} during read-access. The gate–source voltage (V_1) of Q_1

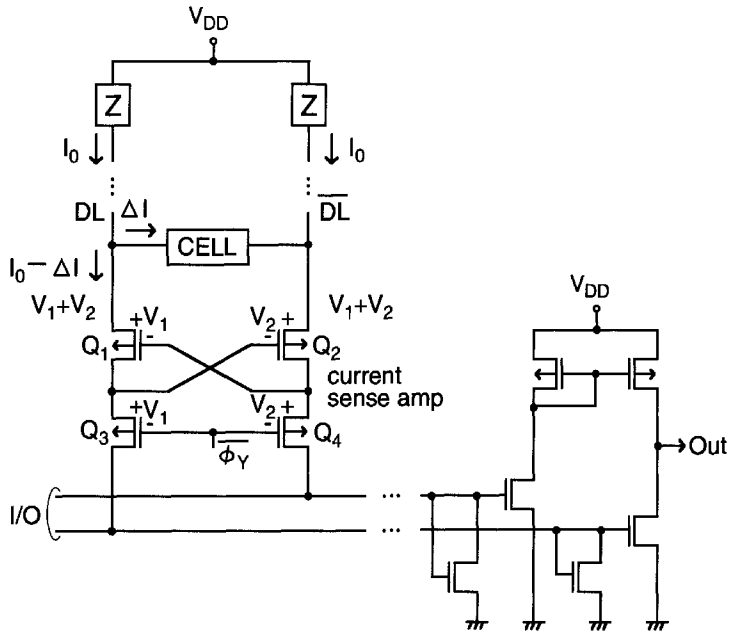


Fig. 7.23. A current sense amplifier connected to data lines, followed by an output section [7.21]

is equal to that of Q_3, since their sizes and currents are equal and both are in saturation. The same applies to Q_2 and Q_4, with V_2 for the gate–source voltage. Since Q_3 and Q_4 are selected, the left data line has voltage $V_1 + V_2$, and the right data line also has voltage $V_1 + V_2$. Therefore, the potentials of the two data lines are equal, independent of the current distribution, thus constructing a virtual short-circuit across the data lines. Since the data-line voltages are equal, the data-line load currents are also equal. Thus, when the cell is accessed and draws current ΔI, the right-hand leg of the amplifier must pass more current than the left leg. In fact, the difference between these currents is ΔI, the cell current. The drain currents of Q_3 and Q_4 are passed to current-transporting I/O lines. The resultant differential I/O current is converted to a large signal voltage by diodes and a current mirror on the I/O lines. Thus, the amplifier achieves a small sensing delay, which is insensitive to the data-line capacitance.

Figure 7.24 shows a current sense amplifier combined with PMOS data-line loads [7.22]. It facilitates low-voltage, low-power, and high-speed operation. A normally on equalizer reduces the data-line voltage swing to less than 30 mV. The current-mirror configuration of the amplifier makes a current difference of $\Delta I/2$ between Q_1 and Q_5 (Q_4 and Q_8), which eventually discharges and charges the outputs, S and $\overline{\text{S}}$. The bias-voltage generator provides an intermediate voltage of 1−1.5 V at 3.3 V V_{DD} to increase the gain by operating Q_1–Q_4 close to the saturation region. An extremely small data-line voltage swing of less than 30 mV, and elimination of the need for pulsed data-line equalization, allow fast sensing. Note that the required voltage swing in

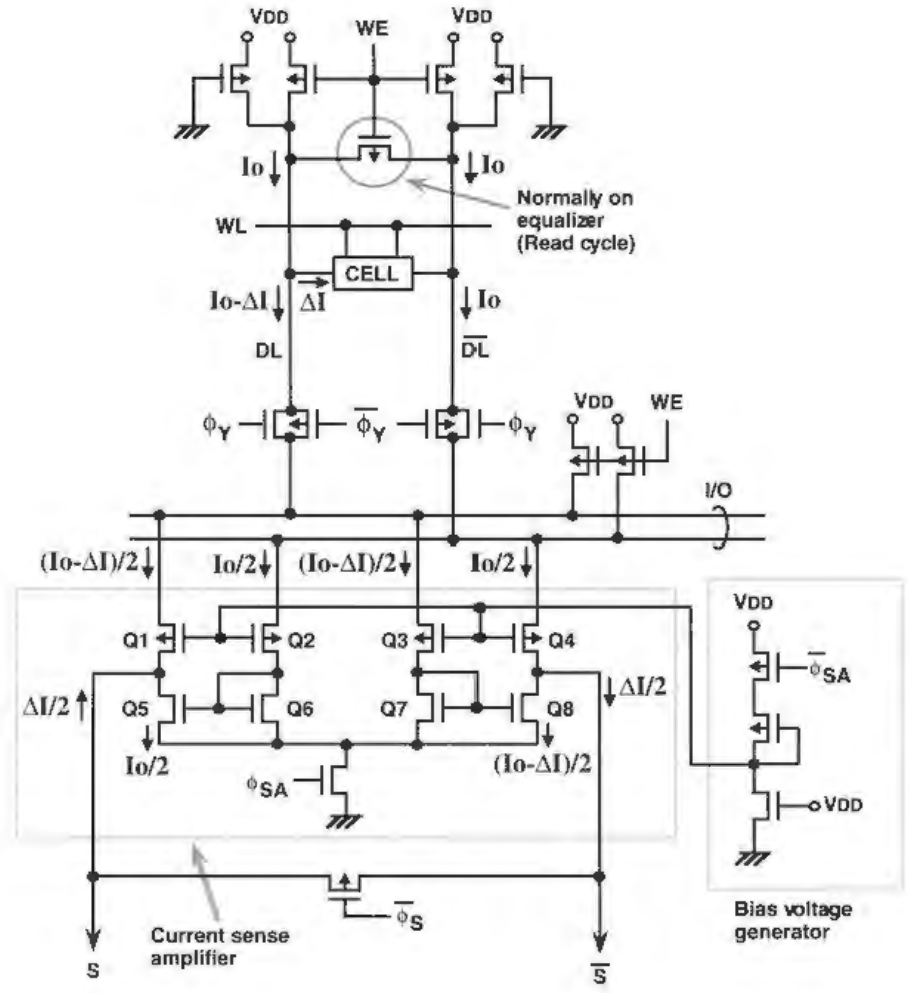


Fig. 7.24. A current sense amplifier [7.22]

a conventional voltage amplifier is 100–300 mV, as shown in Table 7.4. For a fixed delay of 1.2 ns, the amplifier reduces the current consumption by about 2 mA compared with a conventional current-mirror voltage amplifier.

Reduction of the Operating Voltage. A 5 V power supply and NMOS data-line loads, as shown in Fig. 7.17, have been widely used for SRAMs. An NMOS V_T drop provides intermediate input voltages, which are essential to obtain large gains and fast sensing speeds, to the sense amplifiers. For lower V_{DD} operation, however, if a resultant data-line signal voltage close to V_{DD} is amplified, PMOS data-line loads without the V_T drop are more suitable. This is made possible by level-shifting the data-line voltages. A resulting intermediate voltage allows it to use conventional voltage amplifiers for the succeeding stage. The use of NMOS source followers and scaled low-V_T NMOSs are good examples of level-shifting [7.1].

On-chip Voltage Down-Conversion [7.1]. On-chip power-supply conversion was first used for a 256 Kb SRAM, to internally supply 3.3 V to the 0.7 μm devices with a 5 V V_{DD}. The SRAM has a power-down mode to reduce the stand-by chip current to 5 μA. A 4 Mb SRAM, which turns off one of two VDCs (voltage down-converters) to provide a 50 μA stand-by current, automatically shuts down the two VDCs when the external supply voltage is reduced to 3.3 V, to obtain a 1 μA data retention current. An experimental VDC which achieves a sub-μA stand-by current has been reported. The VDC circuits in SRAMs are basically the same as the DRAM VDCs shown in Chap. 5.

Reduction of the Charging Capacitance. Charging capacitance reduction techniques, initially targeted at obtaining a high speed in SRAMs, also contribute to power reduction. These techniques include data-line division and I/O line division. Figure 7.25 shows the partial activation of a multidivided data line [7.25, 7.26], which is similar to that of the DRAM. Each heavily capacitive data line is multidivided. The resulting subdata lines are connected to a lightly capacitive metal global-data line through subarray selection switches. The combination of one global-data line and one selected subdata line reduces the data-line capacitance, affording a low power and high speed. The insertion of a predecoding stage between an address buffer and a final decoder also optimizes both speed and power, and has been used in most SRAMs.

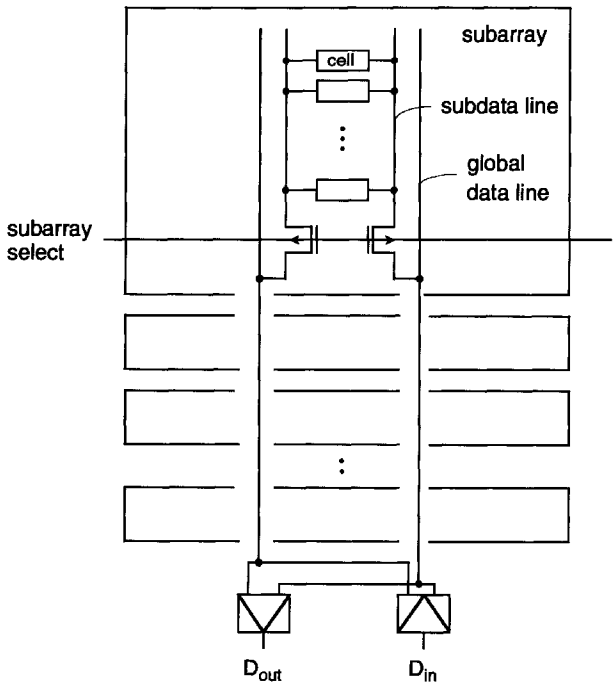


Fig. 7.25. A multidivided data line [7.25, 7.26]

7.5.2 Data-Retention Power Reduction

On-chip voltage generators have not been widely used in commercial SRAMs, although there have been many attempts, that differ from the DRAM approach. This is mainly because the power-supply standardization issue has not been so critical as that of the DRAM, according to the SRAM's market transition, from general-purpose to specific uses. Thus, the data-retention current has been sufficiently reduced solely by improvements in the memory cell. Switching the power-supply voltage to 2 V or to 3 V, from 5 V in normal operation, further reduces the data-retention current. However, this voltage-switching approach is eventually restricted by the ultra-low-voltage ability of SRAMs, which will be described in Chap. 8.

8. Ultra-Low-Voltage Memory Circuits

8.1 Introduction

The reduction of the operating voltage is essential not only to reduce power dissipation, but also to ensure reliablity for miniaturized devices. Further reduction of the supply voltage allows users to design ultra-low-power systems or battery-based mobile equipment. One target is 0.9 V, the minimum voltage of a NiCd cell. Even in this case, higher performance will eventually be required, due to the ever-increasing demand for digital signal processing capability. To simultaneously achieve low voltage and high-speed operation, both device miniaturization and threshold voltage (V_T) scaling are indispensable. Fortunately, the state-of-the-art device technology in miniaturization has at last progressed to the extent of realizing quite high speed even at low voltages of less than 2 V. Even SOI CMOS technology – more suitable for ultra-low-voltage operations – is being intensively developed. V_T scaling, however, is highlighted as an emerging issue, because the reduction of V_T increases the MOSFET subthreshold current [8.1–8.6], even in CMOS chips. The resulting dc current eventually dominates even the active chip current, losing the low-power advantage of CMOS circuits that we take for granted today. The issue is essential not only for the design of large-capacity RAM chips with a feature size of 0.1 μm or less in the future, but also for the design of ultralow-voltage LSIs such as medium memory-capacity RAM chips and MPU chips with an existing fabrication process tailored to scaled V_T. Circuit development for suppressing the subthreshold current was initiated by RAMs [8.15–8.18] with gate–source back-biasing and multi-V_T schemes, based on a few exploratory developments of 1.5 V DRAMs [8.13, 8.59]. This was quickly followed by logic circuits [8.19] aimed at extensive uses in logic-oriented chips. After that, in addition to improvements of the schemes above, a wide variety of variable V_T schemes [8.20] have been proposed. In addition to the subthreshold current issue, there are other key design issues [8.5] that affect ultra-low-voltage RAM designs. These are stable memory cell operation to cope with reduction of the memory-cell signal charge, suppression of or compensation for design parameter variations that enhance speed variations at ultra-low voltages, and power-supply standardization, which will be more difficult for general-purpose RAM designs.

This chapter describes ultra-low-voltage (0.5–2.0 V) RAM circuits, with emphasis on the reduction of the subthreshold current. First, the state-of-the art curcuit designs are summarized in terms of four key design issues listed above. Second, ultra-low-voltage DRAM circuits are discussed, focusing on subthreshold current reduction for the memory cell and the peripheral circuit. To cope with the emerging memory-embedded system LSIs, subthreshold current reduction for logic circuits is also discussed. Third, ultra-low-voltage SRAM circuits are explained, emphasizing cell driving schemes. Finally, the potential of SOI technology is discussed from the viewpoint of low-voltage circuit design.

8.2 Design Issues for Ultra-Low-Voltage RAM Circuits

8.2.1 Reduction of the Subthreshold Current

MOSFET threshold-voltage (V_T) scaling is a key issue in the simultaneous achievement of low-voltage and high-speed operation. The high-speed operation of CMOS circuits necessitates a scaled-down V_T, because speed is roughly inversely proportional to $V_{DD} - V_T$. However, when V_T becomes small enough to no longer cut off the MOSFET, a MOSFET subthreshold dc current is developed, which increases exponentially with decreasing V_T, as discussed below.

To evaluate the subthreshold current caused by V_T scaling, the definition of V_T must be clarified. There are two kinds of V_T [8.3, 8.7]: the extrapolated V_T that is familiar to circuit designers, and constant-current V_T, as shown in Fig. 8.1. The extrapolated V_T is defined by extrapolating the saturation current on the $\sqrt{I_{DS}} - V_{GS}$ plane, and by neglecting the tailing current actually developed at approximately V_T. Our major concern is with the subthreshold current, which is developed at $V_{GS} = 0$ V. If V_T is high enough, the subthreshold current is zero. With decreasing V_T, however, the subthreshold current starts to be developed at a V_T higher than expected. This current is not expressed in the definition. Thus, the constant-current V_T is indispensable in evaluating the current. V_T is defined as a V_{GS} for a given current density on the $\log I_{DS} - V_{GS}$ plane. This constant-current V_T is empirically estimated to be smaller than the extrapolated V_T by about 0.2 V for a current density of 2 nA/μm.

The subthreshold leakage current [8.8], I_L, is given by

$$I_L = W \cdot \frac{I_0}{W_0} \cdot 10^{-V_T/S} , \tag{8.1}$$

$$S \simeq \frac{kT_j}{q} \cdot \left(1 + \frac{C_{DP}}{C_{OX}}\right) \ln 10 , \tag{8.2}$$

where W is the gate width of the FET, I_0/W_0 is the current density to define V_T, S is the subthreshold swing, C_{OX} is the gate capacitance, C_{DP}

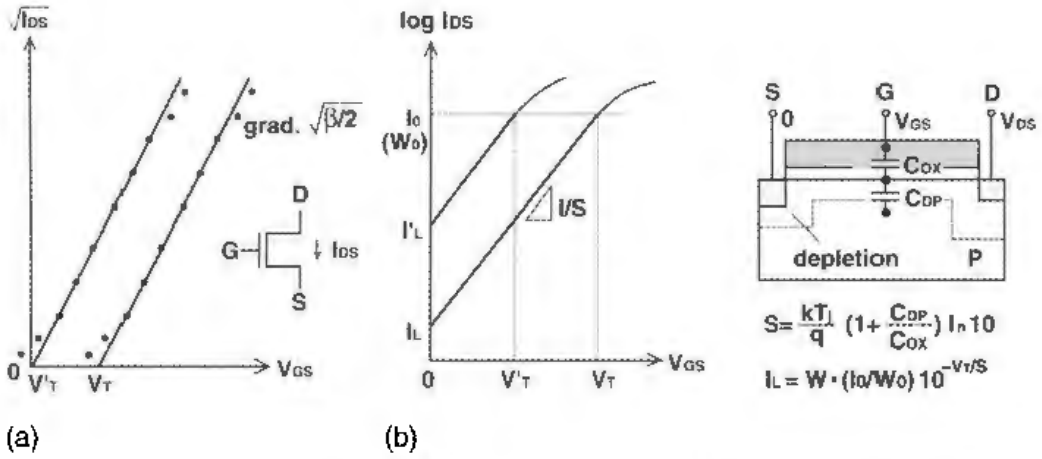


Fig. 8.1. The definition of V_T [8.3]. (**a**) extrapolation; (**b**) constant current

is the depletion-layer capacitance, and T_j is the junction temperature. To reduce the subthreshold current for a given V_T, the W of each MOSFET relevant to the subthreshold current must be reduced. Reduction of S is also effective. This is realized by lowering T_j, which also reduces the current with raised V_T. Thus, liquid-nitrogen temperature (−196 °C) operation [8.9-8.11] has been proposed, although it is not suitable for general-purpose CMOS LSIs. The SOI device [8.12] also reduces S with reduced C_{DP}/C_{OX} from about 80 mV/decade (dec.) of bulk MOSFETs to about 60 mV/dec. at room temperature. Even the SOI device never reduces S to less than 60 mV/dec., because S is always larger than $(kT_j/q)\ln 10$ ($\simeq$ 60 mV/dec.), independent of the values of C_{DP} and C_{OX}, as seen in (8.2). The subthreshold swing S of practical bulk CMOS devices is about 100 mV/dec. at the highest T_j of 100 °C for a usual design. This implies that the subthreshold current increases by one decade with a V_T decrease of only 100 mV. Note that the subthreshold current of PMOSFETs in a usual CMOS logic circuit (Fig. 8.2) is larger than that of NMOSFETs because of a larger W and a larger S, as discussed later.

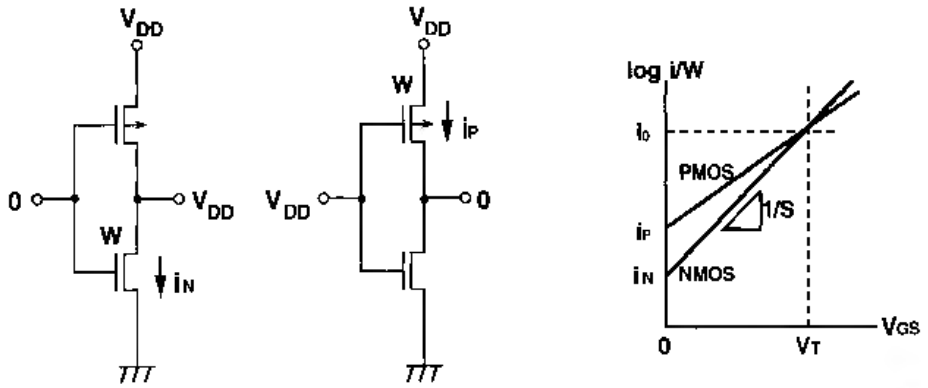


Fig. 8.2. The increase in the subthreshold current in a CMOS inverter [8.3]

The subthreshold current disables the detection of defective chips by monitoring the quiescent power-supply current (the so-called I_{DDQ} test). It also causes unusual memory-cell operation because it discharges the DRAM-cell stored node, and it increases the SRAM-cell retention current. Moreover, the current would dramatically increase not only the stand-by current, but also the active current of the chip, because many inactive iterative circuit-blocks, such as row and column decoders and drivers, start to generate large subthreshold currents. Note that the necessary V_T for RAM cells is larger than that for the peripheral circuits because of their different requirements. The details are discussed in the following.

The Memory-Cell Current. In the DRAM cell the subthreshold leakage current of a cell FET flows from the cell storage node to the data line while the data line is at a low level, as shown in Fig. 8.3 [8.5] and Fig. 4.14. This degrades the data-retention time of the DRAM cells. Since a prolonged data-retention time (t_{REFmax}) is needed, the 1-T DRAM cell needs an ever-higher V_T with increasing memory capacity, as shown in Fig. 4.17. Consequently, of all LSIs, the DRAM cell needs the highest V_T. In SRAMs, two sources (I_L, I'_L) for the leakage current are established in a cell. As a result of current accumulation in numerous cells, a SRAM array suffers from a huge data-retention current along with a decreasing V_T [8.5], as shown in Fig. 8.4. It should be noted that the current of a 0.25 μm 1 Mb SRAM array would exceed 10 A at a 50 °C junction temperature, assuming a ± 0.1 V variation in V_T for a 0.1 V nominal V_T that is needed for 1 V operation. If an acceptable data-retention current for the array is a few μA, the nominal V_T at room temperature would be higher than 0.75 V for the same variation in V_T. Once V_T is chosen to be 0.75 V, the cell is difficult to operate at a V_{DD} below 0.75 V, because the cross-coupled FETs are almost cut off. The SRAM cell array, the block that has effectively the largest channel width, thus calls for the largest V_T for the cell FETs in a chip. This large V_T requirement makes ultra-low-voltage design more difficult.

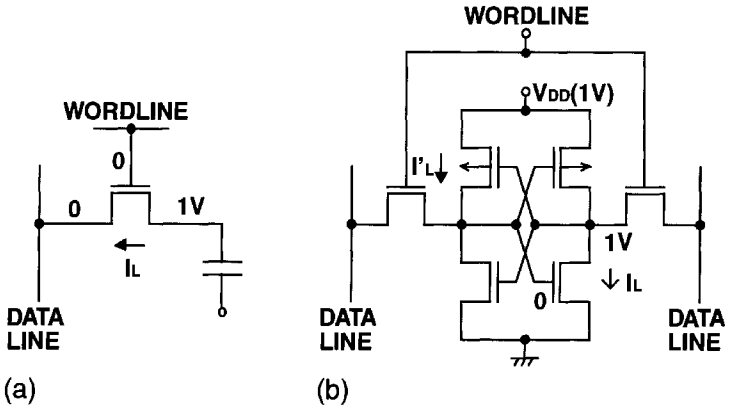


Fig. 8.3. The degradation of memory cell characteristics due to a reducation in V_T [8.5]. (**a**) DRAM cell; (**b**) SRAM cell

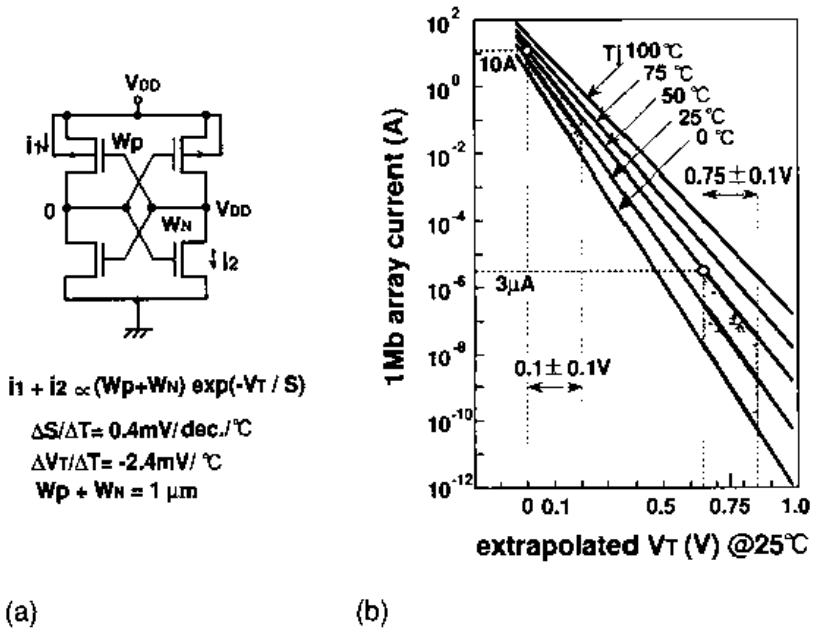


Fig. 8.4. The calculated subthreshold current of a SRAM cell-array [8.5]. (**a**) The current sources in a full CMOS cell; (**b**) the subthreshold current of a 0.25 μm 1 Mb SRAM array versus the extrapolated V_T

The Peripheral-Circuit Current. As for the peripheral circuit, the subthreshold current is a source of dc current for the basic CMOS circuit shown in Fig. 8.2 [8.3]. For a V_{DD} higher than approximately 2 V, where the extrapolated V_T can be higher than around 0.4 V, either the NMOSFET or the PMOSFET in the inverter is completely cut off. For V_{DD} less than 2 V, however, a subthreshold current becomes prominent, and is enhanced with a high temperature and variation in V_T, as expressed by (8.1). The current eventually dominates even the active current of the memory chip. Such is the case for a memory-embedded system LSI chip and a logic chip (the MPU/ASIC chip).

Figure 8.5 [8.3] shows a chip that comprises random logic circuits and iterative circuit blocks, each of which is regarded as a circuit with an effectively large channel width. At present, attention is mainly being paid to the subthreshold current issue during the stand-by period, since the V_T is still too high. With a further reduction in V_T, however, even the numerous circuits, especially the iterative circuit blocks that are inactive during the active period, will start to generate a subthreshold current. The subthreshold dc chip current (I_{DC}) from the overwhelmingly large number (i.e. effectively large W) of inactive circuits increases exponentially with a reduction in V_T, and would increase not only the data-retention current but also the active current. As a result, I_{DC} would exceed even the ac chip current (I_{AC}), which is the total charging current for capacitive loading of the small number of active circuits, and would eventually dominate the active current of the chip

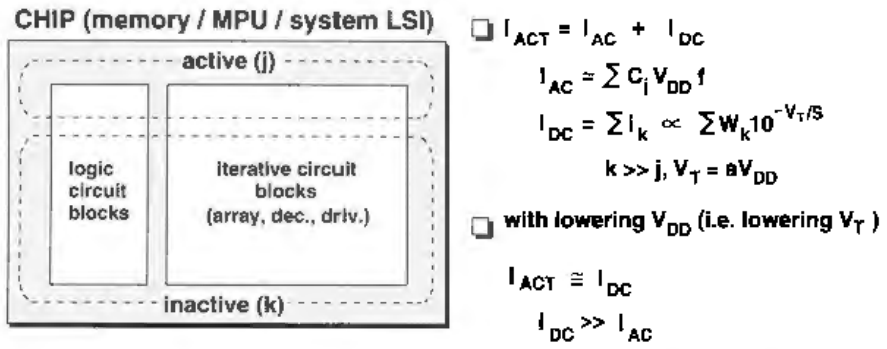


Fig. 8.5. The subthreshold current increase in CMOS LSIs [8.3]

(I_{ACT}). This is because, in addition to the huge difference in the number of relevant circuits, the ac capacitive current from each active circuit decreases with a decrease in V_{DD}, while the subthreshold current from each inactive circuit increases exponentially with a reducing V_{DD}.

This is well exemplified by the DRAMs shown in Fig. 8.6 [8.2]. The I_{ACT} increases with lowering of the operating voltage, reaching 1.2 A for a 16 Gb chip because of the rapid increase in the subthreshold dc current, I_{DC}. The ever-increasing total channel width of the iterative circuit blocks with increasing memory capacity is also responsible for the increase in I_{DC}. Other LSIs,

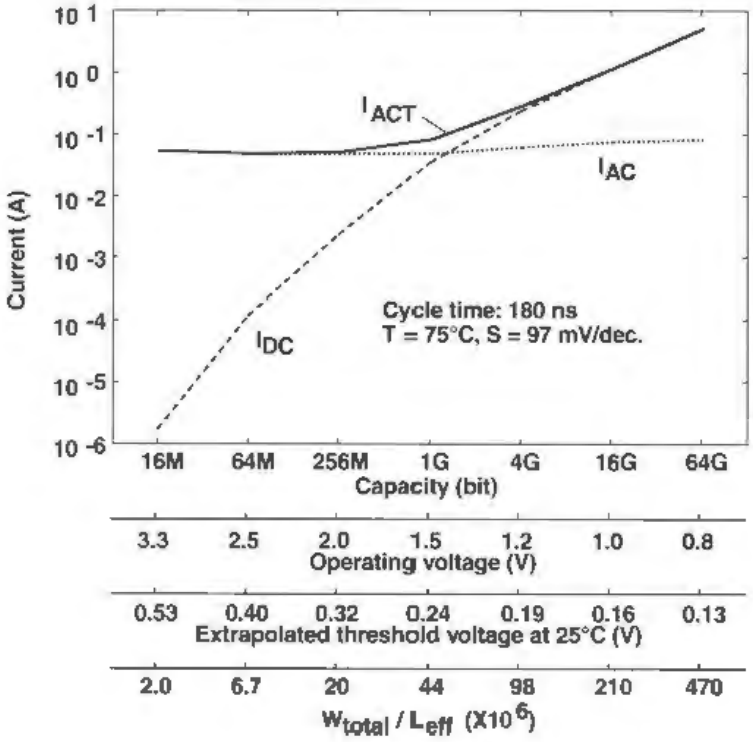


Fig. 8.6. The estimated active current of DRAMs [8.2]

such as the SRAM, Flash memory, the MPU/ASIC, and the system LSI, will suffer more or less from the same situation. Thus, our special concern is how to reduce the current from inactive circuits, especially from iterative circuits such as the array, the row and column decoders, and the drivers.

Fortunately, however, reduction of the subthreshold current for a RAM chip is easier than for other LSIs [8.5]. This stems from the following inherent features of a RAM chip. First, a slower memory cycle time makes current reduction easier. A physically large memory-cell array that occupies over 50% of the chip, a relatively large capacitance and high resistance in the word line and data (or bit) line, and a small memory-cell signal that necessitate succeeding amplification are all responsible for a slower cycle time when compared with MPU/ASIC chips, as discussed in Chap. 6. Moreover, the circuits are active only for a short period within the "long" memory cycle time, allowing an additional control time so as to reduce the subthreshold current. Second, a larger number of iterative circuit blocks contribute to the reduction. A RAM chip incorporates many kinds of iterative circuit blocks, such as the memory-cell array, the row/column decoders and their relevant drivers, sense amplifiers, address buffers, and I/O buffers. However, almost all circuits in each iterative circuit block, except for a selected one, are inactive even during the active period of the chip. This enables simple and effective control in reducing the subthreshold current of each block. Third, the incorporation of input-predetermined logics is more effective in reducing the current. Although a memory cycle starts to randomly select a memory cell using a few external clocks and address signals, the peripheral circuits do not work as random logic within the memory cycle. Thus, the designer can predict which FETs in the chip will cut off not only during the stand-by period, but also during the active period.

It is obvious that reduction of the subthreshold current in a memory-embedded system LSI, in which a large memory block and a large logic block are incorporated, is the most difficult. This is because, for a conventional memory chip with a large memory block and a relatively small logic block in the peripheral circuit, reduction is necessary only for the memory block, with less concern for the logic block that generates a small subthreshold current. This could be managed by utilizing the above-described features of the memory. On the other hand, for a conventional logic chip with a small memory block (such as a SRAM cache) and a large logic block, it is needed only for the logic block, with less concern for the memory block that generates a small subthreshold current. However, the subthreshold currents generated during high-speed logic operations are hard to reduce because of a shortage of sufficient time for reduction control, unlike the memory chip. On the other hand, for the memory-embedded system LSI, the reduction is necessary for both the logic and the memory blocks, which requires the most sophisticated circuit techniques.

8.2.2 Stable Memory-Cell Operation

In addition to the subthreshold-current issue, there are two major issues that are relevant to stable memory-cell operation. They are maintenance of the signal charge (Q_S) and the decrease of the V_T-mismatch between a pair of MOSFETs in the flip-flop circuits [8.6]. This V_T-mismatch is caused by both extrinsic and intrinsic V_T variations.

Maintenance of Q_S. Reducing the power-supply voltage, V_{DD}, inevitably decreases the signal charge (Q_S) of the memory cell, as shown in Fig. 8.7, causing a small cell-signal voltage on the data line. Therefore, the cell operation is susceptible to various kinds of noise, which results in unstable operation [8.5]. Hence, an increase in Q_S is critical in extending the lower limit of V_{DD}. For DRAMs, in addition to a full write operation and a thinner capacitor-insulator, a half-V_{DD} capacitor plate that enables double capacitance, vertical capacitors such as stacked and trench capacitors, and a capacitor-over-data (or -bit) line (COB) structure affording an increased capacitor area have all been especially important in obtaining a larger Q_S in products. The details are given in Chap. 4.

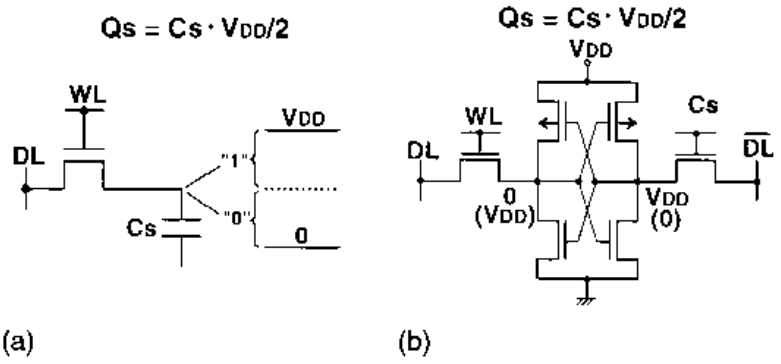


Fig. 8.7. The cell signal charge Q_S versus the supply voltage V_{DD} [8.5]. (**a**) DRAM cell; (**b**) SRAM cell

V_T-Mismatch. The ever-increasing V_T-mismatch between paired MOSFETs will be a limiting factor for future RAM designs. In DRAM design it will be possible to maintain the mismatch of sense amplifiers to some extent through enlargements of MOSFETs with minimum chip-area penalty, as discussed in Chap. 4. In SRAM design, however, the enlargement of a flip-flop SRAM cell is fatal because the chip area is increased unacceptably. Even in the absence of extrinsic variations (implant non-uniformities and channel length/width variations), intrinsic variations in V_T exist due to random microscopic fluctuations of dopant atoms in the extremely small MOSFET channel area. Figure 8.8 shows the standard deviation of intrinsic V_T-mismatch and the largest V_T-mismatch in a SRAM cell [8.21]. The ever-increasing mismatches for three sets of MOSFET pairs in a full CMOS cell are

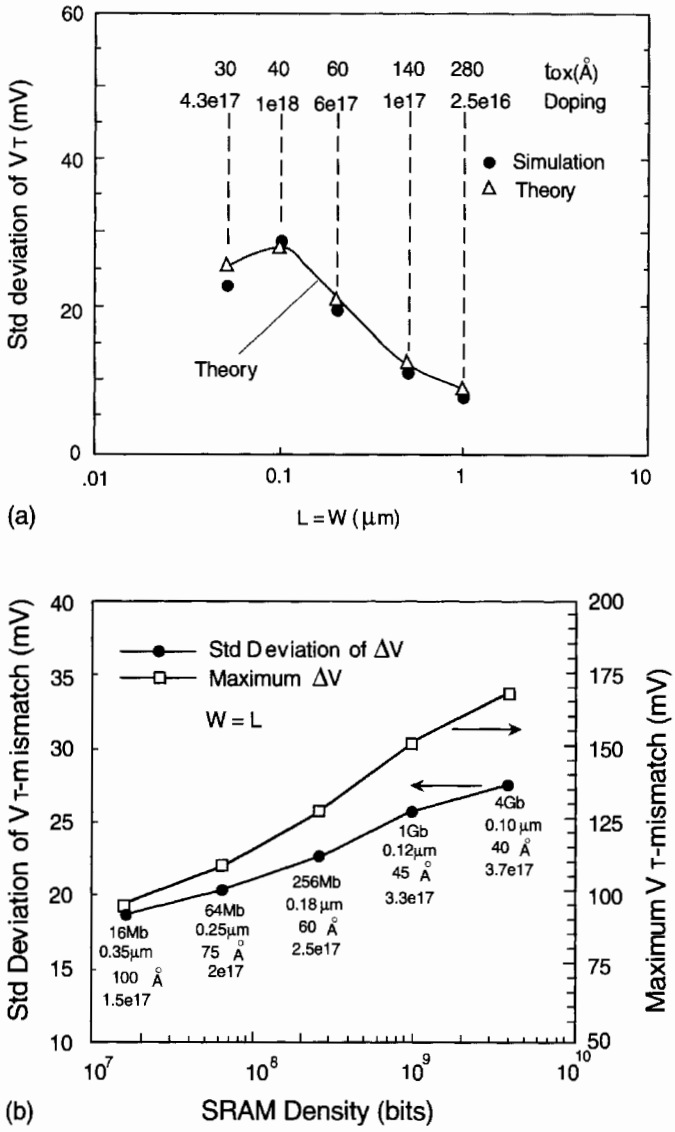


Fig. 8.8. Intrinsic deviations in V_T [8.21]. The gate-oxide thickness (t_{OX}), and doping are listed for each technology. (**a**) The standard V_T deviation; (**b**) the standard deviation of V_T-mismatch and the maximum V_T-mismatch in a SRAM cell for minimum size FETs

anticipated to be one of the most serious issues for ultra-low-voltage designs. Thus, a solution will be scaled MOSFET developments combined with stringent control of MOSFET characteristics, and/or redundancy techniques [8.6] that avoid fatal cells.

8.2.3 Suppression of, or Compensation for, Design Parameter Variations

Suppression and compensation for variations in design parameters such as V_T, channel length, temperature, and V_{DD} are keys to achieving ultra-low-voltage

LSIs. Extrinsic device-parameter variations, however, are unavoidably introduced during volume production. Even a fixed variation increases chip-to-chip speed variations with a reduction in V_{DD}. Unfortunately, the miniaturization of FETs increases extrinsic and intrinsic variations, causing unexpectedly large speed variations at ultra-low values of V_{DD}. Unregulated battery power supply makes the design more complicated, causing further speed variations. Figure 8.9 shows an example of the speed variations of a DRAM chip, assuming $\Delta V_T = \pm 0.15\,V$ and $\Delta L = \pm 0.1\,\mu m$ for FETs of 0.6 µm channel length (L) and 15 nm gate-oxide thickness (t_{OX}). Here, the memory capacity, the chip area, and all dimensions except those of the FETs are assumed to be fixed. A voltage setting of $V_{DD} = 3\,V$ and $V_T = 0.45\,V$ allows a normalized access time spread of 0.7–1.5 according to two combinations of $\Delta V_T/\Delta L$; that is, $-0.15\,V/-0.1\,\mu m$ and $+0.15\,V/+0.1\,\mu m$. Another voltage setting of $V_{DD} = 1\,V$ and $V_T = 0.15\,V$, however, increases the spread from 1.6 to 5.6 with a nominal value of 3. Obviously, the reduction in V_{DD} not only degrades the speed, but also the speed spread for fixed design-parameter variations. If a high-performance FET of $L = 0.3\,\mu m$ and $t_{OX} = 7.5\,nm$ is used, almost the same nominal access time could be obtained at $V_{DD} = 1\,V$ as at $V_{DD} = 3\,V$. The speed spread of 0.4–2.3 expanded by the same ΔV_T and ΔL can be narrowed remarkably to 0.7–1.5, if both ΔV_T and ΔL are scaled down to 0.5. Note that in addition to the temperature increase, a decrease in V_T of about 0.1 V increases the subthreshold current ten-fold. In addition to the extrinsic V_T and L variations above, intrinsic V_T variations become large across a chip, which enhances the statistical variations in the MOSFET drain current (I_{dsat}) and speed. Figure 8.10 shows projected ranges of V_T and I_{dsat} variations on a chip [8.21]. The range of V_T is expected to roughly double as devices are scaled down from 0.5 µm to 0.1 µm and the number of devices

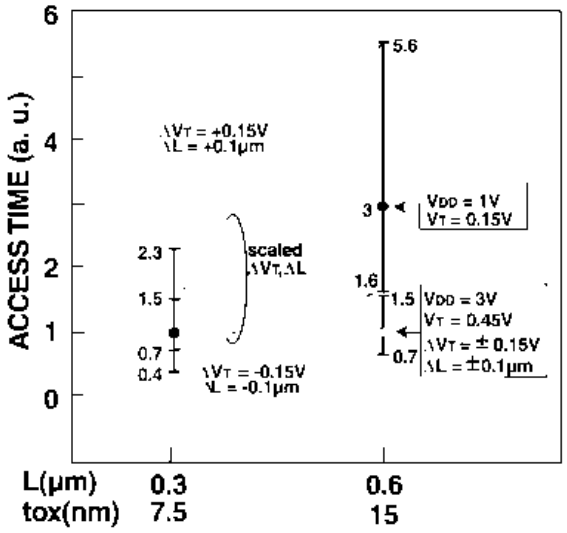


Fig. 8.9. Speed variation for design-parameter variations [8.5]

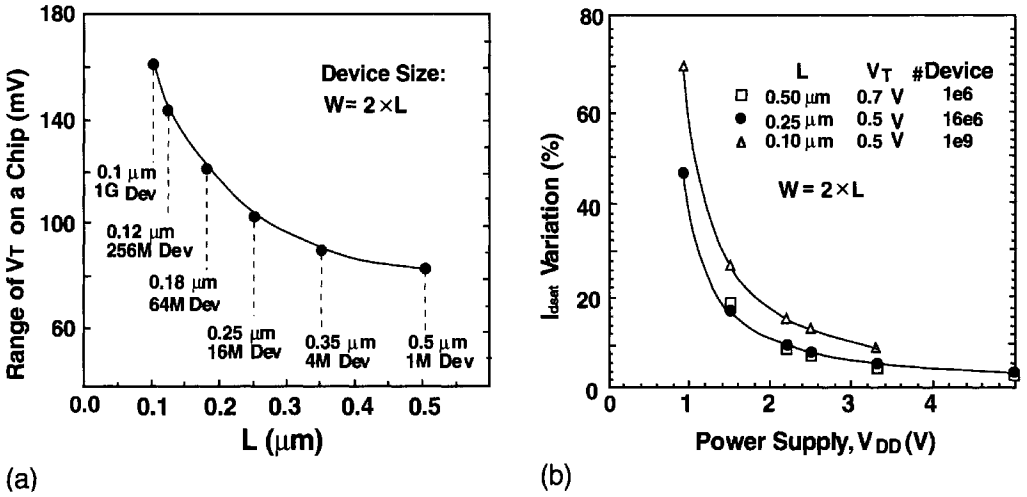


Fig. 8.10. The projected range of V_T (**a**) and variation in I_{dsat} (**b**) on a chip [8.21]. The gate-oxide thickness and doping for a given technology are the same as those listed in Fig. 8.8

increases. The corresponding I_{dsat} variation due to doping fluctuations is less than 10% for 0.5 μm and 0.25 μm FETs with $V_{DD} \geq 2.5$ V, while it is about 30% for a chip with one billion 0.1 μm FETs at $V_{DD} = 1.5$ V and $V_T = 0.5$ V.

In addition to the stringent control of channel length, a shallow-junction MOSFET, which is formed by reducing the ion-implantation energy and process temperature, reduces the variations in V_T and in the offset voltage of sense amplifiers. Innovative MOSFETs for suppressing the intrinsic device-parameter variations are indispensable. Compensation circuits against design-parameter variations are also important. One solution may be internal operation-voltage control to track the variations, in which on-chip voltage generators play an important role, as illustrated by substrate bias control for a microprocessor [8.63].

8.2.4 Power-Supply Standardization

Single-Power Supply. RAM chips for general-purpose use have needed a single power supply while retaining power-supply standardization [8.23], in spite of the two power supplies of recent MPU/ASIC and memory-embedded system LSI chips, which are exemplified by 3.3 V (= V_{DDQ}) for the I/O and 2.5 V (= V_{DD}) for the internal core circuit (Fig. 8.11). In the course of lowering the operating voltage of general-purpose RAMs, three-power-supply operation of 12 V (V_{DD}), 5 V (V_{cc} only for the I/O interface circuit), and −5 V (V_{BB} for substrate bias) changed to single power-supply operation of 5 V V_{DD} in the 64 Kb generation, and then to 3.3 V in the 64 Mb generation. On-chip voltage generators, as explained in Chap. 5, such as V_{BB} generators, voltage down-converters, voltage up-converters, half-V_{DD} generators, and reference-

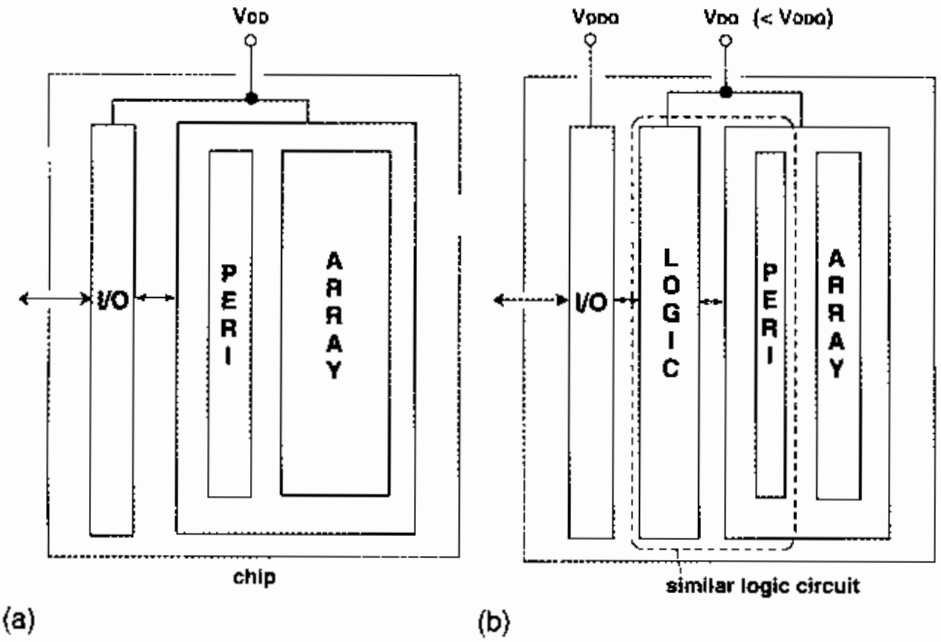


Fig. 8.11. Examples of applications of a low-voltage memory circuit. PERI, peripheral circuit of memory; LOGIC, large-scale logic. (**a**) Single power supply (general-purpose-use memory); (**b**) dual power supply (MPU, ASIC, and memory-embedded system LSI)

voltage generators, have contributed to single-V_{DD} operation. In single-V_{DD} operation, the choice of a standard V_{DD} is one of the most important, and serious concerns because V_{DD} is always closely related to chip performance. However, the recent excessive rapid down-scaling of CMOS devices and the strong demand for battery operation are making standardization difficult. In fact, V_{DD} is still controversial in the 1 Gb DRAM generation, although many attempts at low-voltage (2 V to 0.5 V) operation have been made.

In the long run, on-chip voltage generators will continue to be important to achieve single power-supply operation and to standardize the power supply. All generators require a high conversion efficiency, precise control, and trimming of internal voltages to compensate for the above-described variations in the design parameter. Their power consumptions are another concern. For example, a higher boost ratio, as explained in Chap. 5, for a voltage up-converter consumes a higher power because of its lower conversion efficiency. Thus, low-power/low-voltage analog circuits are expected to become increasingly important. Using these generators, a gradual transition toward external or internal power supplies of sub-V levels seems inevitable in terms of ever-decreasing devices, as shown in Fig. 8.12 [8.4]. Thus a landmark will be passed at around 1 V, which is suitable for one-cell battery operation. At the 0.5 V level, even one-solar-cell operation may be possible.

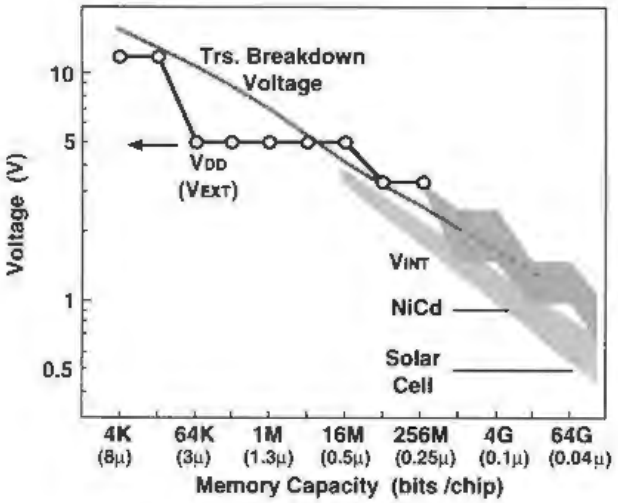


Fig. 8.12. The standard power-supply voltage, V_{DD}, of DRAMs [8.4]. V_{INT}, internal supply voltage

Dual-Power Supply. Embedded RAM technology for system LSIs (Fig. 8.11) is a real challenge, even in the low-voltage era. In system LSIs, the dual power-supply scheme as used in recent MPU/ASIC chips will continue to exist. Although the scheme is inferior to the single power-supply scheme in terms of ease of use, it achieves low power and high-speed operations caused by scaled-down devices in the internal core circuit while allowing the I/O interface to be matched with the high-voltage interface of low-end LSI chips. However, in order to ensure device reliability at the I/O and high speed in the internal circuit, the scheme requires two kinds of gate-oxide thickness of MOSFETs, thick oxide for the I/O circuit and thin oxide for the internal circuit, or the insertion of stress-release MOSFETs for the I/O circuit [8.59–8.61]. A high V_T of the I/O circuit enables not only the use of traditional CMOS circuit techniques, which can avoid hazardous designs to reduce the subthreshold current, but also the use of MOSFETs without a substrate (or well) bias. Thus, the issue of the subthreshold current is confined to the internal core circuit. On the contrary, the low-V_T I/O circuit that is necessary in a low-voltage single power-supply scheme may result in instabilities such as the undershoot problem at I/O pins and CMOS latch-up, as discussed in Chap. 3.

8.3 Ultra-Low-Voltage DRAM Circuits

In this section, reduction of the subthreshold current, especially for the DRAM, is discussed. The discussion covers logic-oriented designs, because they would be applicable to peripheral circuits of RAM chips and memory-embedded system LSIs.

Table 8.1. Summary of concepts for low-voltage high-speed circuits[a]

Threshold voltage V_T	Usage of V_T	Controls for MOSFET			Well stucture
Fixed V_T	High V_T	G-boost			Double
	Low/high V_T	Multi-V_T			Double
	low V_T	G–S	G–S offset drive		Double
	Low V_T	back-bias	G–S self-bias		Double
Variable V_T	Low V_T (active)	W–S back-bias	W-control	Chip	Triple
				Block	Triple
				Circuit	Double
	High V_T (stand-by)			MOSFET	(SOI)
			S-control		Double
			W/S-control		Double

[a]G, gate; S, source; W, well (i.e. sub).

The subthreshold-current reduction circuits [8.5] proposed so far can roughly be categorized as fixed-V_T circuits and variable-V_T circuits, as shown in Table 8.1. The fixed-V_T circuits are further categorized as gate (G) boosting circuits, multi-V_T circuits, and gate–source (G–S) back-biasing circuits. Note that the fixed high V_T would eventually provide a limitation on an ultra-low-V_{DD} high-speed active operation. A variable V_T can be achieved by well(or substrate)–source (W–S) back-biasing through well control, source control, or well–source control, so that V_T in the inactive period is higher than it is in the active period. This offers fast operation during the active period while suppressing the leakage current during the inactive (stand-by, or sleep) period. There are four kinds of well controls, depending on the scale of the circuits whose wells are simultaneously controlled: simultaneous control of the well of the whole chip, control of the well of a circuit block, control of the well of a circuit, and control of the well of a MOSFET. In general, well controls of the whole chip and block need a triple-well structure. Note that a variable V_T is also effectively realized by the gate–source back-biasing, despite a fixed low V_T.

Figure 8.13 illustrates the concepts behind the variable-V_T approaches. Obviously, the substrate–source direct connection of PMOS and NMOS during active period achieves high speed, with the resulting low V_T. The subthreshold current during the stand-by period is reduced by the gate (G)–source (S) back-bias or the substrate (SUB or well)–source (S) back-bias. The G–S back-bias realizes an effectively high V_T by changing the source voltage

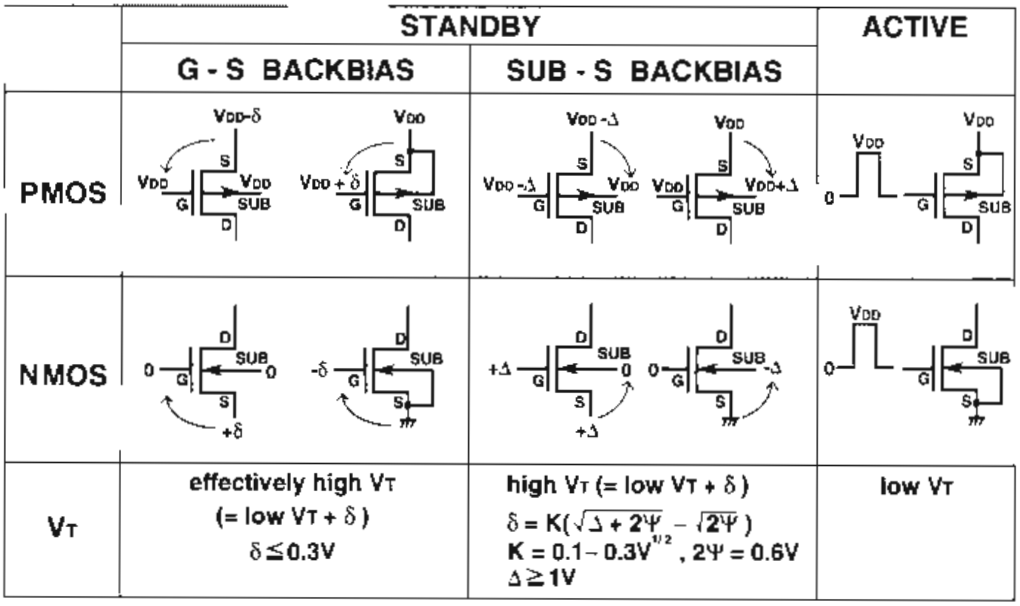


Fig. 8.13. The concepts behind the variable-V_T approaches for reducing the stand-by subthreshold current of low-V_T MOSFETs

from V_{DD} to $V_{DD} - \delta$ for a PMOS, and from 0 V to δ for an NMOS, or by changing the gate voltage from V_{DD} to $V_{DD} + \delta$ for a PMOS, and from 0 V to $-\delta$ for an NMOS. In practice, δ can be as small as less than 0.3 V, because even a small δ greatly reduces the subthreshold current. On the other hand, the SUB–S back-bias realizes a high V_T by changing the source voltage by Δ with a fixed substrate voltage, or by changing the substrate voltage by Δ with a fixed source voltage. To realize the same high V_T (i.e. the same δ) as in the G–S back-bias, Δ must be quite large. The necessary Δ is derived from the well-known equation

$$\delta = K\left(\sqrt{\Delta + 2\psi} - \sqrt{2\psi}\right) .$$

For example, a δ of 0.2 V that enables a two-order reduction in the subthreshold current for MOSFET with $S = 0.1\,\mathrm{V/dec.}$ requires a Δ as large as 2.5 V for $K = 0.2\,\mathrm{V}^{1/2}$, as shown in Fig. 8.14 [8.66]. The resulting large Δ may cause an excessive stand-by power due to p-n-junction leakage [8.65]. Note that the G–S back-bias caused by changing the source voltage by δ also raises the V_T. However, the raising effect is negligible, because of a small δ.

The following details are for the concepts shown in Table 8.1.

8.3.1 Gate Boosting Circuit

The scheme aims at obtaining a high speed by boosting the gate of a high-V_T MOSFET. It is particularly useful for sense amplifiers: the speed of a sense amplifier is lowered significantly when the operating voltage is reduced,

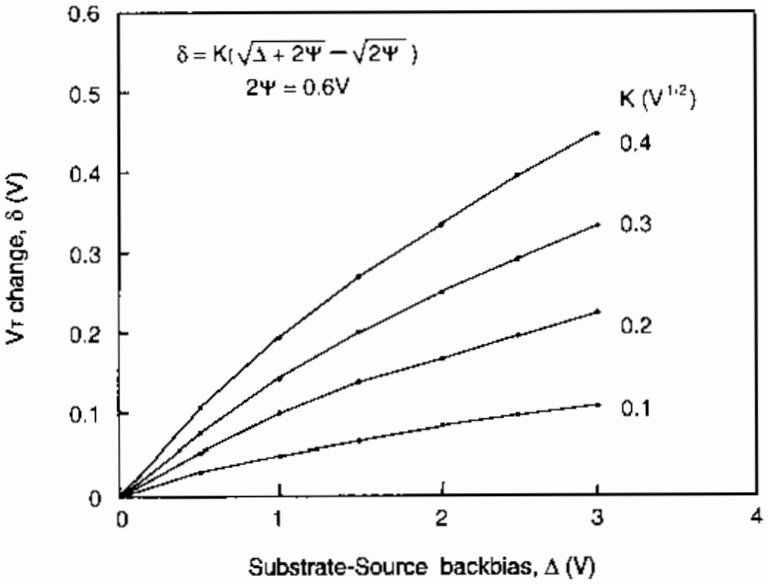


Fig. 8.14. Change in V_T versus substrate-source back-bias [8.66]

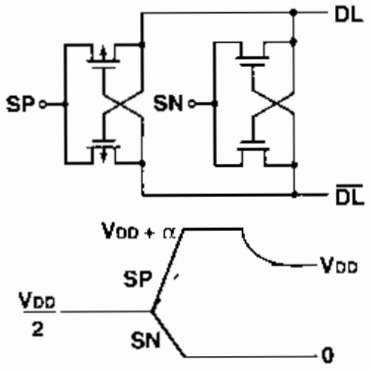


Fig. 8.15. A low-voltage overdrive seuse amplifier [8.35]

because the effective gate–source voltage of the MOSFETs is around $V_{DD}/2$ in the initial stage of cell-signal simplification. The overdrive (node boost) scheme shown in Fig. 8.15 [8.35] offsets the reduction in speed despite a high-V_T. The common-source node is boosted in the initial stage to accelerate the amplification speed of the PMOS amplifier and then driven to V_{DD} after an adequate difference is established between the data-line pair.

8.3.2 The Multi-V_T Circuit

This cuts off the leakage path with a high-V_T MOSFET while using a low-V_T MOSFET for the main signal path during the active period. The following are typical applications for the power switch and logic circuits. Figure 8.16a shows a switched power-supply inverter with a static level holder [8.17]. This is

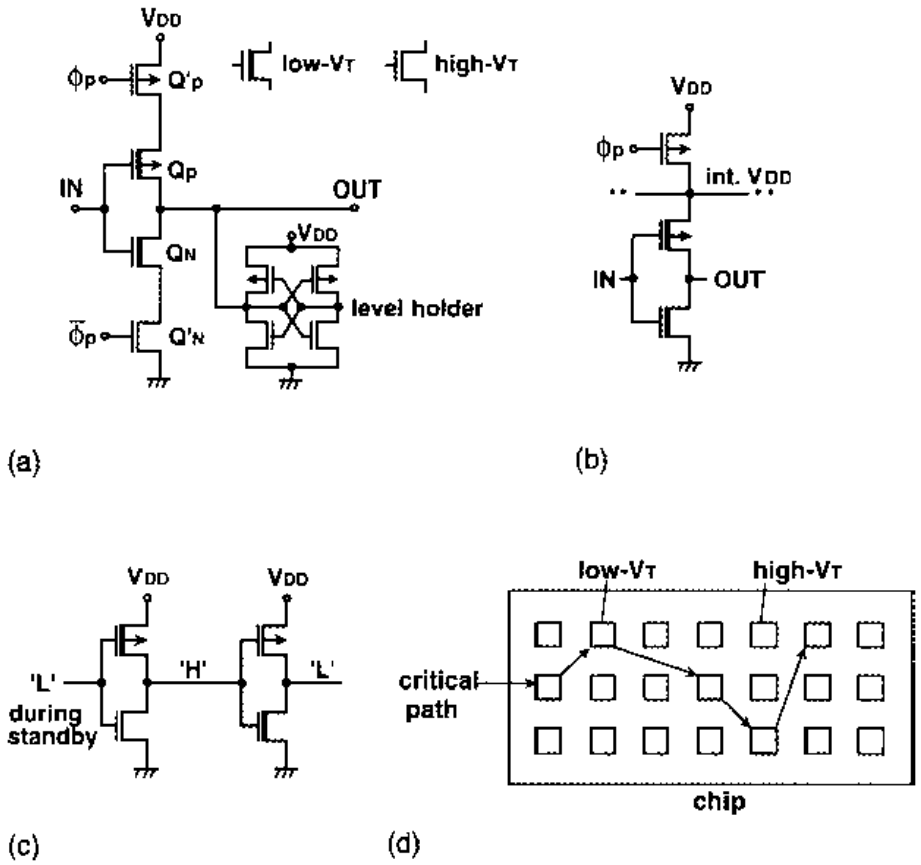


Fig. 8.16. Multi-V_T circuits [8.17, 8.19]. (**a**) A power switch with a level holder; (**b**) a shared power switch; (**c, d**) logic circuits

useful for some applications in which the input voltage is not predictable. The power supply of the CMOS circuit is controlled by the FET switches Q'_N and Q'_p. The V_T values of all FETs except Q_N and Q_p are so high that the subthreshold current is negligble. As soon as the input level has been evaluated at high speed as a result of the low V_T of Q_N and Q_p, and the resulting output is held in the holder, the switches are turned off. Consequently, the output level is maintained without a subthreshold current. The area of the level holder can be minimized, since it only plays a role in holding the level. The level holder and Q'_N can be eliminated if the output level does not need to be held. In this case, the power switch can be shared by many internal low-V_T circuits without any subthreshold current [8.19], as shown in Fig. 8.16b. However, the internal V_{DD} node and all the output nodes of the circuits are finally discharged, requiring a long recovery time and a large charging current and power. A multi-V_T circuit combined with a dynamic V_T circuit [8.28] that refreshes the slightly degraded voltage of an internal V_{DD} line at a fixed interval holds the data without a special holding circuit. However, the interval strongly depends on V_T and its variations. Therefore, the power switch is limited to applications that accept a very slow speed.

Figure 8.16c shows an application to a logic circuit [8.18]. A high V_T prevents current leakage in the stand-by mode. Figure 8.16d shows another example of a multi-V_T scheme actually used in a 480 MHz RISC microprocessor [8.22]. A low V_T is used only for the critical paths. The scheme is useful to some extent for stand-by and active current reduction. If we assume that the total W of the critical paths is 10% of the total W of the chip, the low V_T and high V_T are 0.21 V and 0.31 V, respectively, and S is 0.1 V/dec., the scheme reduces the current to one-fifth of the current for uniform use of a low V_T. Obviously, a large difference in V_T makes the scheme more useful. However, it could create a racing problem; that is, a pulse-timing imbalance between low-V_T and high-V_T circuits. In the multi-V_T schemes described above, the high V_T eventually restricts ultra-low-voltage operation, although $V_{DD} = 1$ V at most could be managed. This is because the conductance of the high-V_T FETs decreases as V_{DD} approaches the high V_T, causing the additional area and power to compensate for the decrease in conductance.

8.3.3 The Gate–Source Back-Biasing Circuit

Two schemes of gate–source offset driving and gate–source self-back-biasing are well known.

Gate–Source Offset Driving. Figure 8.17 shows a gate–source offset driving scheme applied to the cell FET [8.5]. Note that δ in Fig. 8.13 corresponds to ΔV_{DL} and ΔV_{WL}. The boosted sense ground (BSG) shown in Fig. 8.17b features the lowest data-line voltage raised by ΔV_{DL} to create a back-bias for non-selected cell FETs. The subthreshold current flow is cut off even for a low V_T as long as the sum of ΔV_{DL} and V_T is larger than the minimum V_T (V_{T0}) that achieves a negligibly small subthreshold current. Obviously, V_{T0} is the value of V_T that satisfies t_{REFmax} in each successive generation (see Fig. 4.17). The negative word line (NWL) shown in Fig. 8.17c works in a similar way. Figure 8.18 shows comparisons of the voltage relationships between cell-driving schemes. Here, V_T is chosen to be 1 V for the conventional

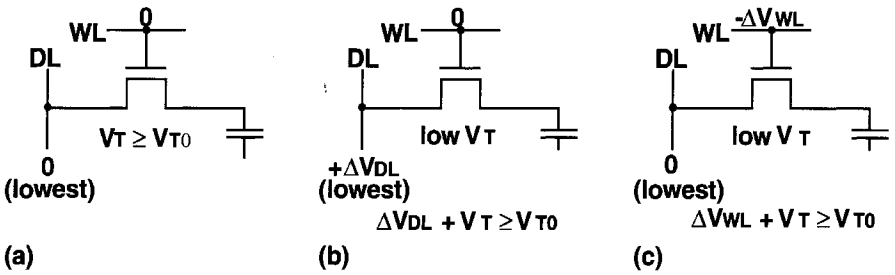


Fig. 8.17. Gate–source back-biasing schemes applied to the DRAM cell [8.5]. V_{T0}, the minimum V_T necessary for preventing subthreshold current flow under the data-line "L" disturbances. (**a**) Conventional; (**b**) boosted sense ground (BSG); (**c**) negative word line (NWL)

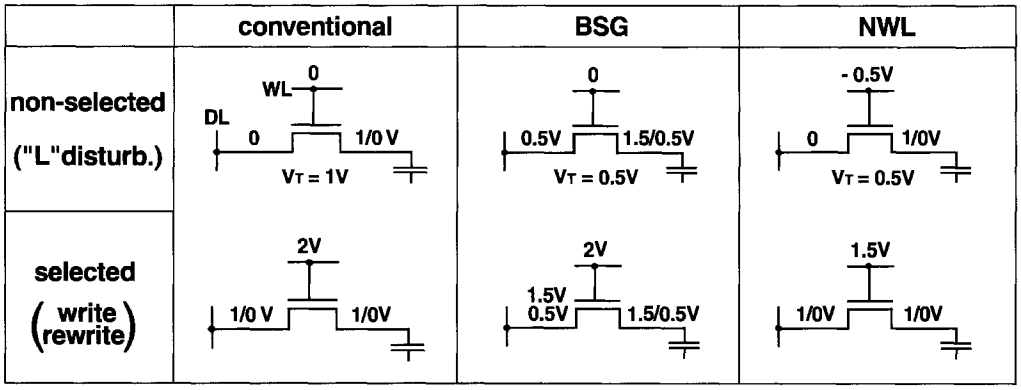


Fig. 8.18. Comparisons between cell-driving schemes with assumptions of $K = 0$, $V_{T0} = 1\,\mathrm{V}$, and a cell storage voltage of 1 V

scheme and 0.5 V for BSG and NWL. Obviously, the low V_T accepted by BSG and NWL eventually reduces the p–n-junction leakage current of the cell, due to the smaller doping concentration of the storage node [8.58]. For the selected cell, the gate-oxide stress voltage of the conventional scheme is the highest, while for non-selected cells it is the lowest. As for the p–n-junction stress voltage, NWL is lowest, enabling less leakage current. Regarding the word-driver design, NWL is favorable due to a smaller boost ratio, but it requires a negative word-line voltage instead. Recently, an NWL scheme was used in an 8 Mb embedded DRAM [8.64]. In BSG the design of a ΔV_{DL} generator is as difficult as that of an on-chip voltage down-converter (VDC), as described in Chap. 5, because it must sink a large data-line discharging current.

Figure 8.19 shows another example of gate–source offset driving [8.24] applied to a CMOS inverter. This design enables the use of a low V_T value by offsetting the source level of the driver by the difference between the high and the low values of V_T. Thus, it achieves high-speed switching due to a reduced signal swing, while keeping the subthreshold current sufficiently low. Consequently, it is suitable for a bus driver with heavy loading capacitance, although it needs two on-chip voltage-generators for V_{DL} and V_{SL}.

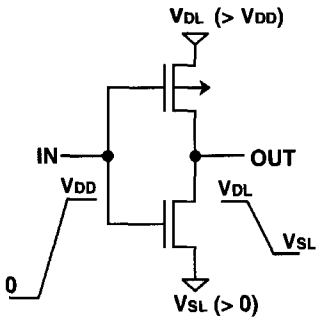


Fig. 8.19. Offset driving [8.24]

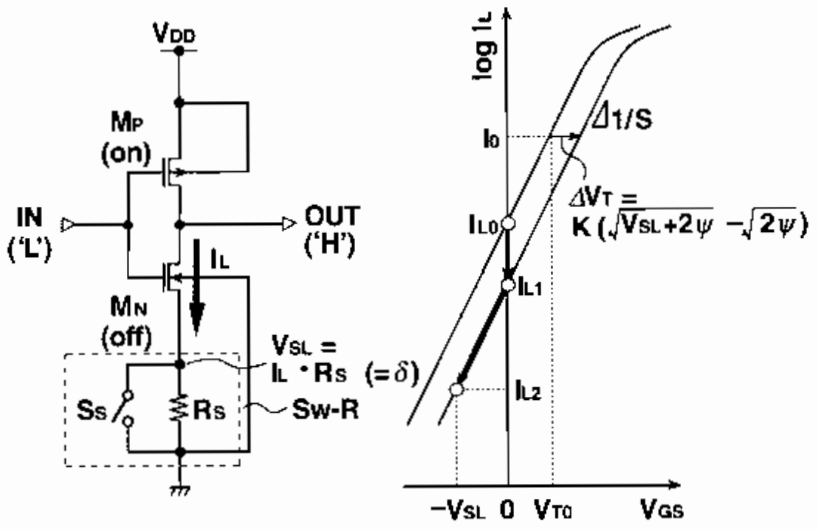


Fig. 8.20. The principle of the gate–source self-back-biasing scheme due to a switched source impedance circuit [8.16]. δ is the same as in Fig. 8.13

Gate–Source Self-Back-Biasing. Figure 8.20 shows the principle of a gate-source self-back-biasing scheme [8.16], using a switched-source-impedance (SSI) scheme. It features a switch, S_s, and a resistor, R_s, which are connected in parallel and inserted at the source of the NMOS transistor M_N, the back gate of which is connected to the ground. In order to achieve concurrent high-speed operation and a low stand-by current, S_s is on in active mode, while it is off in stand-by mode. In stand-by mode, the subthreshold current I_L through the resistor R_s raises the source voltage V_{SL} to δ $(= I_L \cdot R_s)$, creating gate–source back-biasing. Note that δ is the same as the one in Fig. 8.13. The current is reduced through the following two mechanisms. First, the back-gate bias of $-\delta$ enhances V_T by ΔV_T, and the current is reduced from I_{L0} to I_{L1}. Second, the gate–source voltage of M_N becomes negative, $-\delta$, and the current is further reduced from I_{L1} to I_{L2}. The subthreshold current is reduced by 3–4 decades with a δ of 0.3 V. Note that negative feedback through R_s provides immunity against V_T-fluctuations, which become larger with device down scaling. Only one low-V_T FET can realize the switched impedance, because R_s is regarded as the leakage resistance of S_s. This scheme is also applicable to other logic gates as long as the input voltage is predictable. Fortunately, almost all RAM chip node voltages are predictable, as discussed before.

Figure 8.21 shows typical circuit configurations for realizing the SSI scheme. Circuit (a) consists of a high-V_T MOSFET, which works as a switch, and a pure resistor. Circuit (b) comprises only one low-V_T FET (Q_S) that can realize a switched impedance. Here, W_N and W_S are the channel widths of Q_N and Q_S, and V_{TS} and V_{TN} are their constant-current V_T values, defined at I_0. The current-reduction effect of the impedance can be formulated [8.1] if the substrate-bias effect is neglected.

In circuit (a), the Q_N-current (I_L) is expressed as follows:

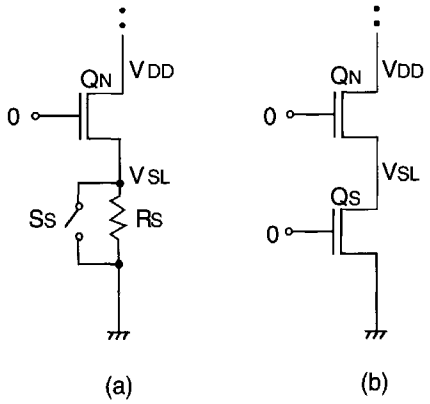


Fig. 8.21. Switched-source impedances [8.1]

$$I_{\mathrm{L}}(V_{\mathrm{SL}})\frac{I_0}{W_0} = W_{\mathrm{N}} \cdot 10^{-(V_{\mathrm{SL}}+V_{\mathrm{TN}})/S} \tag{8.3}$$

$$I_{\mathrm{L}}(V_{\mathrm{SL}}) = V_{\mathrm{SL}}/R_{\mathrm{S}} \; ; \tag{8.4}$$

$$\therefore \; \frac{V_{\mathrm{SL}}}{R_{\mathrm{S}}} = \frac{I_0}{W_0} W_{\mathrm{N}} \cdot 10^{-(V_{\mathrm{SL}}+V_{\mathrm{TN}})/S} \tag{8.5}$$

Thus, the current-reduction ratio (γ) is given by the ratio of $I_{\mathrm{L}}(V_{\mathrm{SL}})$ to $I_{\mathrm{L}}(V_{\mathrm{SL}} = 0)$, as

$$\gamma = 10^{-V_{\mathrm{SL}}/S} \; . \tag{8.6}$$

The sensitivity of I_{L} to V_{TN} variation is obtained by differentiating (8.5) with respect to V_{TN}, as

$$\frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}V_{\mathrm{TN}}} = -\frac{\ln 10}{S}\frac{1}{1+(\ln 10/S)V_{\mathrm{SL}}} \; . \tag{8.7}$$

The sensitivity of I_{L} to S variation is given by differentiating both sides of (8.5) with respect to S, as

$$\frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}S} = \frac{\ln 10}{S}\frac{V_{\mathrm{SL}}+V_{\mathrm{TN}}}{S}\frac{1}{1+(\ln 10/S)V_{\mathrm{SL}}} \; . \tag{8.8}$$

The sensitivity to temperature is expressed as

$$\frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}T} = \frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}V_{\mathrm{TN}}}\frac{\mathrm{d}V_{\mathrm{TN}}}{\mathrm{d}T} + \frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}S}\frac{\mathrm{d}S}{\mathrm{d}T} + \frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}R_{\mathrm{S}}}\frac{\mathrm{d}R_{\mathrm{S}}}{\mathrm{d}T} \; . \tag{8.9}$$

From (8.4) and by differentiating (8.5) with respect to R_{S}, we can obtain

$$\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}R_{\mathrm{S}}} = \frac{1}{R_{\mathrm{S}}}\frac{\mathrm{d}V_{\mathrm{SL}}}{\mathrm{d}R_{\mathrm{S}}} - \frac{V_{\mathrm{SL}}}{R_{\mathrm{S}}^2} \; , \tag{8.10}$$

$$\frac{\mathrm{d}V_{\mathrm{SL}}}{\mathrm{d}R_{\mathrm{S}}} = \frac{SV_{\mathrm{SL}}}{R_{\mathrm{S}}(S+V_{\mathrm{SL}}\ln 10)} \; . \tag{8.11}$$

Thus, by substituting (8.7), (8.8), (8.10) and (8.11) for (8.9), we obtain the following equation:

$$\frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}T} = -\frac{\ln 10}{S}\frac{\frac{\mathrm{d}V_{\mathrm{TN}}}{\mathrm{d}T} - \frac{V_{\mathrm{SL}}+V_{\mathrm{TN}}}{T} + I_{\mathrm{L}}\frac{\mathrm{d}R_{\mathrm{S}}}{\mathrm{d}T}}{1+(\ln 10/S)V_{\mathrm{SL}}} \,. \tag{8.12}$$

In circuit (b), the currents through $\mathrm{Q_S}$ and $\mathrm{Q_N}$ are expressed as

$$I_{\mathrm{L}}(\mathrm{Q_S}) = \frac{I_0}{W_0} W_{\mathrm{S}} \cdot 10^{-V_{\mathrm{TS}}/S} \,, \tag{8.13}$$

$$I_{\mathrm{L}}(\mathrm{Q_N}) = \frac{I_0}{W_0} W_{\mathrm{N}} \cdot 10^{-(V_{\mathrm{SL}}+V_{\mathrm{TN}})/S} \,, \tag{8.14}$$

if $V_{\mathrm{SL}} \gg kT/q$. The source voltage (V_{SL}) varies so that $I_{\mathrm{L}}(\mathrm{Q_N})$ equals $I_{\mathrm{L}}(\mathrm{Q_S})$, which is a constant current determined only by V_{TS} and W_{S}. Thus, by equalizing both currents we can obtain the following:

$$V_{\mathrm{SL}} = (V_{\mathrm{TS}} - V_{\mathrm{TN}}) + \frac{S}{\ln 10} \ln \frac{W_{\mathrm{N}}}{W_{\mathrm{S}}} \,. \tag{8.15}$$

The current-reduction ratio (γ) is given using the same expression as (8.6). The sensitivities of I_{L} to V_{T} and temperature variations are given from (8.2), (8.14), and (8.15), as

$$\frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}V_{\mathrm{TN}}} = -\frac{\ln 10}{S} \,, \tag{8.16}$$

$$\frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}S} = \frac{\ln 10}{S}\frac{V_{\mathrm{TS}}}{S} \,; \tag{8.17}$$

$$\therefore \quad \frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}T} = \frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}V_{\mathrm{TN}}}\frac{\mathrm{d}V_{\mathrm{TN}}}{\mathrm{d}T} + \frac{1}{I_{\mathrm{L}}}\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}S}\frac{\mathrm{d}S}{\mathrm{d}T}$$

$$= -\frac{\ln 10}{S}\left(\frac{\mathrm{d}V_{\mathrm{TN}}}{\mathrm{d}T} - \frac{V_{\mathrm{TS}}}{T}\right) \,. \tag{8.18}$$

Here, an assumption of $\mathrm{d}V_{\mathrm{TS}}/\mathrm{d}V_{\mathrm{TN}} = 1$ (i.e. same variation) is made. Note that V_{SL} is zero if $V_{\mathrm{TS}} = V_{\mathrm{TN}}$ and $W_{\mathrm{S}} = W_{\mathrm{N}}$, revealing no suppression effect. Actually, however, V_{SL} is raised from 0 V. This is because the subthreshold current has a slight dependence on the drain–source voltage (V_{DS}), unless V_{DS} is somewhat larger than kT/q. If V_{SL} is comparable to kT/q, the $\mathrm{Q_S}$ current is slightly smaller than that expressed by (8.13). Therefore, even the insertion of $\mathrm{Q_S}$, whose characteristics are same as those of $\mathrm{Q_N}$, reduces the subthreshold current [8.25].

The following are distinguishing features of both circuits. The current is reduced exponentially with increasing V_{SL}, as expressed in (8.6) – that is, with increasing $R_{\mathrm{S}}W_{\mathrm{N}}$ for circuit (a) – and $V_{\mathrm{TS}} - V_{\mathrm{TN}}$ and/or $W_{\mathrm{N}}/W_{\mathrm{S}}$ for circuit

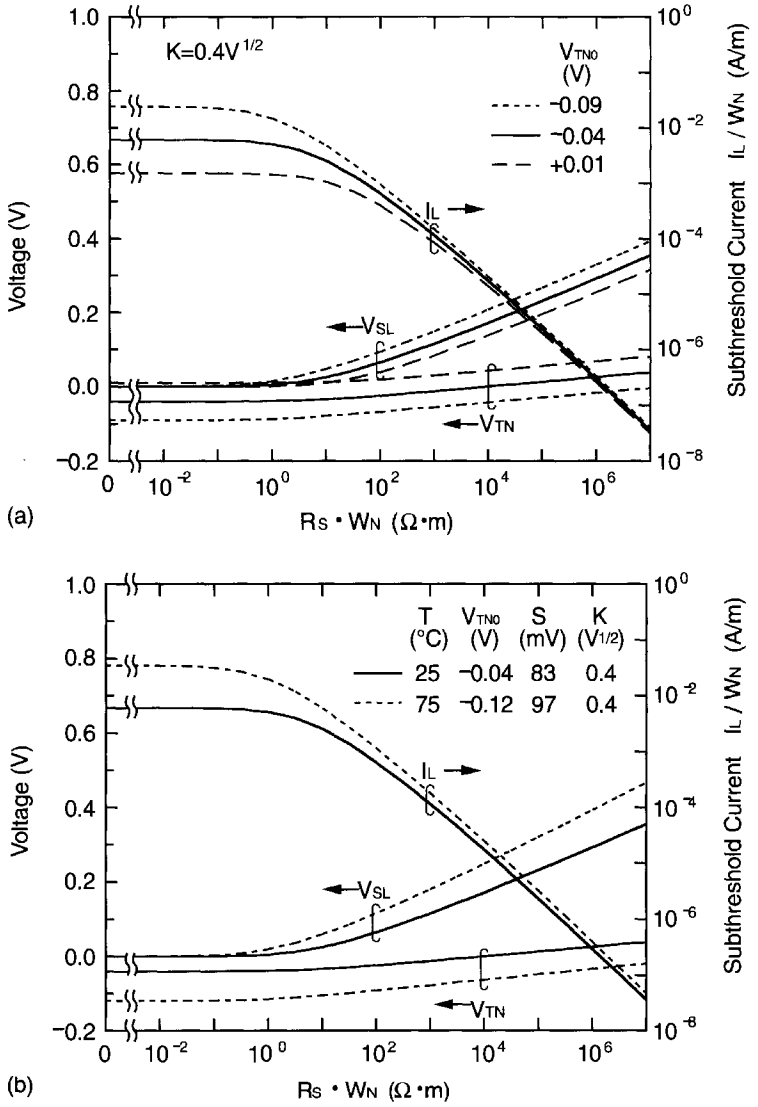


Fig. 8.22. The reduction of the subthreshold current due to source resistance [8.16]. Dependences on V_T (**a**) and temperature (**b**)

(b), as shown by (8.5) and (8.15). In addition, it is obvious that the current of circuit (a) is less sensitive to variations for V_T, S, and temperature, when comparing between (8.7) and (8.16), (8.8) and (8.17), and (8.12) and (8.18). The advantages of circuit (a) are enhanced by a larger V_{SL}. Note that the current variation of circuit (a), expressedusing (8.12), could be cancelled if $dR_S/dT > 0$ because of $dV_T/dT < 0$. Here, $dR_S/dT = 0$ for poly-Si resistors that are usually used in LSIs.

Figure 8.22 shows analyses [8.16] that take the substrate-bias effect into consideration. Figure 8.22a depicts the relationship between the normalized subthreshold current (I_L/W_N) and $R_S W_N$ with a parameter of V_{TN0}, which is the constant current V_T without the substrate-bias effect. Here, $I_0/W_0 =$

2×10^{-3} A/m, $S = 83$ mV/dec., and $V_{TN0} = -0.04 \pm 0.05$ V are assumed. It is obvious that the current reduces remarkably with an increase in $R_S W_N$, that is, an increase in R_S for a fixed W_N, or an increase in W_N for a fixed R_S. The increased V_{SL} and V_{TN} are responsible for this decrease. Consequently, a V_{SL} as small as about 0.3 V allows current reductions of 3–4 decades. In addition, the current is less sensitive to variations in V_{TN} as V_{SL} becomes large due to a negative feedback effect given by R_S, as expected. Figure 8.22b shows the temperature characteristics. In the case of $R_S = 0$, a 5.7-fold current increase is developed for a temperature increase from 25 °C to 75 °C, because of the decreased V_{TN0} and the increased S. At a $R_S W_N$ value of $10^4 \Omega \cdot$m, however, the current increases by only 1.45 times, as a result of a suppressed temperature effect.

Figure 8.23 shows variations of circuit (a) in Fig. 8.21. Case (a) is an extreme one, with $R_S = \infty$, while case (b) is another one in which the source is fixed at a certain voltage (V'), so as to cut the subthreshold current during the inactive period. The former might be hazardous in some designs due to the resulting floating node. The latter can completely cut off the FET through simple control of the switch, if $V' = V_{DD}$. In the case of a heavy source-capacitance, however, the necessary time and power dissipation involved in charging and discharging of the capacitance with a large voltage swing of V_{DD} become serious.

Figure 8.24 shows various applications of the scheme [8.16]. The scheme shown in Fig. 8.20 is only effective when the voltage levels of the IN and OUT terminals are low ("L") and high ("H"), respectively, as shown in Fig. 8.24a. If OUT is at a low level, the switched impedance must be inserted at the source of PMOS transistor M_P, as shown in Fig. 8.24b, because M_P is in the subthreshold region. Note that, in both cases, full-swing output voltages are available despite the source impedances. If OUT is at a tristate (high impedance), however, switched impedances are necessary at both sources, as shown in Fig. 8.24c. This is because the OUT voltage level is determined by another circuit with which the output terminal is shared. The switched

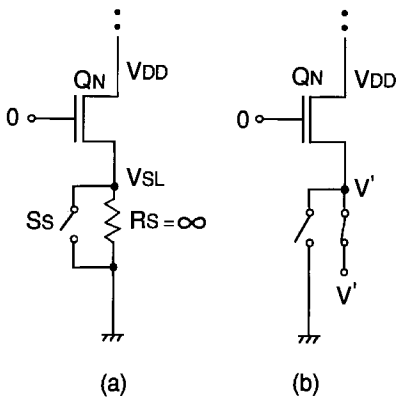


Fig. 8.23. Extreme cases of the SSI scheme

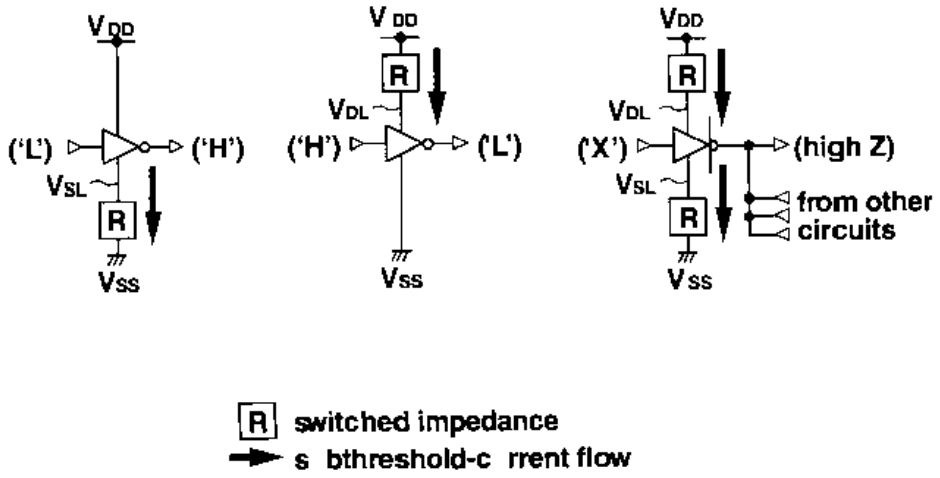


Fig. 8.24. Variations of CMOS circuits using the SSI scheme [8.16]: (a) Inverter with low output; (b) inverter with high input; (c) clocked inverter

impedance can be shared by other inverters to minimize the area penalty. Figure 8.25 shows a typical example of a switched impedance [8.16] shared with a number of circuits. An application to an inverter chain is shown. This reduces the stand-by subthreshold current at an L input. Note that in active mode the operating currents of the inverters do not overlap, although all inverters operate. This "time share" capability allows switched impedances to be small in size.

A voltage degradation at a floating node, if any, caused by the subthreshold current, must be suppressed. A typical example is word drivers to which a floating raised voltage (V_{DH}) from an on-chip voltage up-converter comprising a charge pump is applied. If the total subthreshold current exceeds the current provided by the converter, the V_{DH} level is degraded, and the driver fails to perform a full write to the cell. This inconvenience tends to occur because each driver MOSFET in a word-driver block has an inherently large channel width for high speed, which creates a large subthreshold cur-

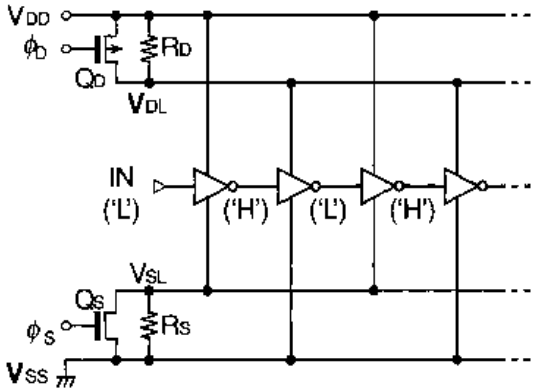


Fig. 8.25. The sharing of switched-source impedances for an inverter chain [8.16]

rent. Besides, the charge pump in the converter provides the V_{DH} node with a poor current, as discussed in Chap. 5. Once the V_{DH} level is degraded, the recovery time is long, due to the poor driving capability of the converter and a heavy floating node capacitance, exemplified by over 10 pF for a 256 Mb design. Figure 8.26 shows a switched impedance shared with decoded word drivers [8.15, 8.17] to suppress the stand-by subthreshold current. A low-V_TP-ch switching FET (Q_S) inserted between the V_{DH} power-supply line (PSL) and the common-source terminal of the driver FETs (the Q_S) is the switched impedance, which works as a current-limiting device. In the active period, the successive operation of selection and word-line driving is done after the PSL is connected to V_{DH} by turning on Q_S. Here, the Q_S channel width (W_S) can be reduced to an extent comparable to the Q channel width (W_D) without degrading the speed, since only one of the n driver transistors is turned on. Just after the stand-by period starts, with the turning off of Q_S, the total subthreshold current, nI, causes a voltage drop (V_{SL}) at the power line, because Q_S acts as an impedance. As a result, a voltage drop creates a gate–source back-bias to each PMOS driver transistor, so that the current is reduced. The current reduction ratio, γ, and V_{SL} are expressed by using (8.6) and (8.15), as

$$\gamma = 10^{-V_{SL}/S} ,$$
$$V_{SL} = (V_{TS} - V_{TD}) + \frac{S}{\ln 10} \ln \frac{nW_D}{W_S} . \tag{8.19}$$

Obviously, the current is drastically reduced because W_S is comparable to W_D and the number of driver transistors, n, is large. For example, γ is as small as 1.5×10^{-3}, and V_{SL} is 0.254 V, with assumptions of $W_S/W_D = 5$, $n = 256$,

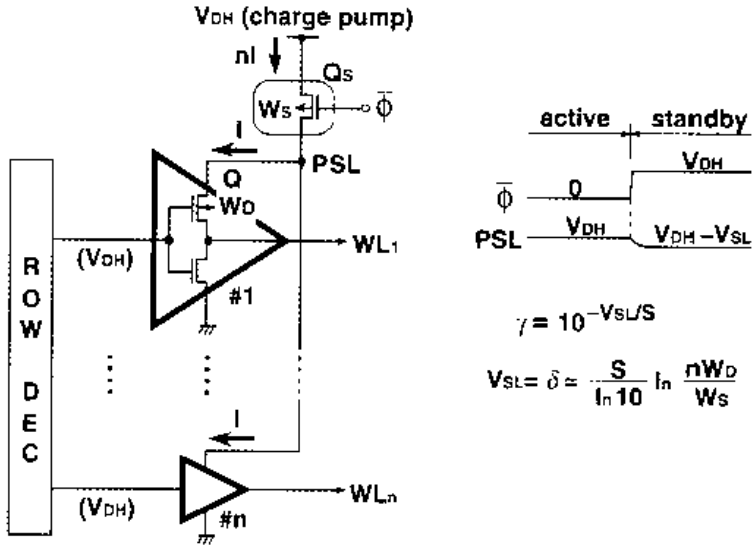


Fig. 8.26. The application of the SSI scheme to word drivers [8.15, 8.17]

$S = 90\,\mathrm{mV/dec.}$, $V_{TS} = 0.2\,\mathrm{V}$, $V_{TD} = 0.1\,\mathrm{V}$. It has been reported [8.15] that this scheme allows the subthreshold current of a 256 Mb chip to be reduced from 219 μA to 6 μA at room temperature. A small V_{SL} enables the high-speed recovery (2–3 ns) of the PSL node back to the V_{DH} level in the transition from the stand-by mode to the active mode. Note that if Q_S has a sufficiently high V_T value, the PSL node is discharged to zero, which implies a slow recovery time, and increased charging current and power.

Figure 8.27 shows another example of a 256 Mb chip using the SSI scheme [8.26]. The chip is composed of an NMOS memory-cell array (MA), a memory-array control circuit block (MA CTL), including iterative circuits such as row/column decoders and row(i.e. word)/column drivers, and I/O and peripheral circuits. Various internal supply voltages are generated, with an external single V_{DD} of 2.5 V. A V_{DH} of 3.8 V is for word the line, a V_{DL} of 2 V is for the data line, and a V_{BB} of −1 V is for the substrate of the memory-cell array. The internal voltages were also utilized as the substrate voltages of specific MOSFETs in the MA CTL to raise their V_T values and realize multi-V_T without additional masks, as explained below. The substrates (i.e. wells) of the I/O and peripheral logic circuits are not back-biased: the wells of the PMOSFETs and the NMOSFETs are fixed to V_{DD} (2.5 V) and V_{SS} (0 V) to realize high-speed active operation with a low V_T and to avoid minority-carrier injection at the I/O pins and CMOS latch-up. The resulting subthreshold current is still small, because of a small-scale logic circuit. Figure 8.28 shows a device cross-section of the 256 Mb chip [8.26]. The I/O and peripheral circuit are formed with a double-well structure on

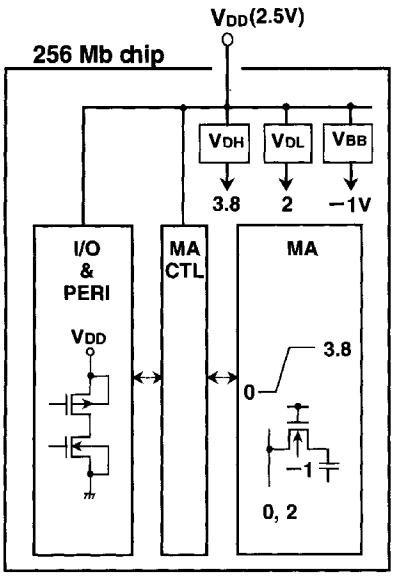


Fig. 8.27. 256 Mb chip architecture using the SSI scheme [8.26]. MA, memory array; MA CTL, MA control circuit (X/Y dec. and driver etc.); I/O & PERI, I/O and periphery

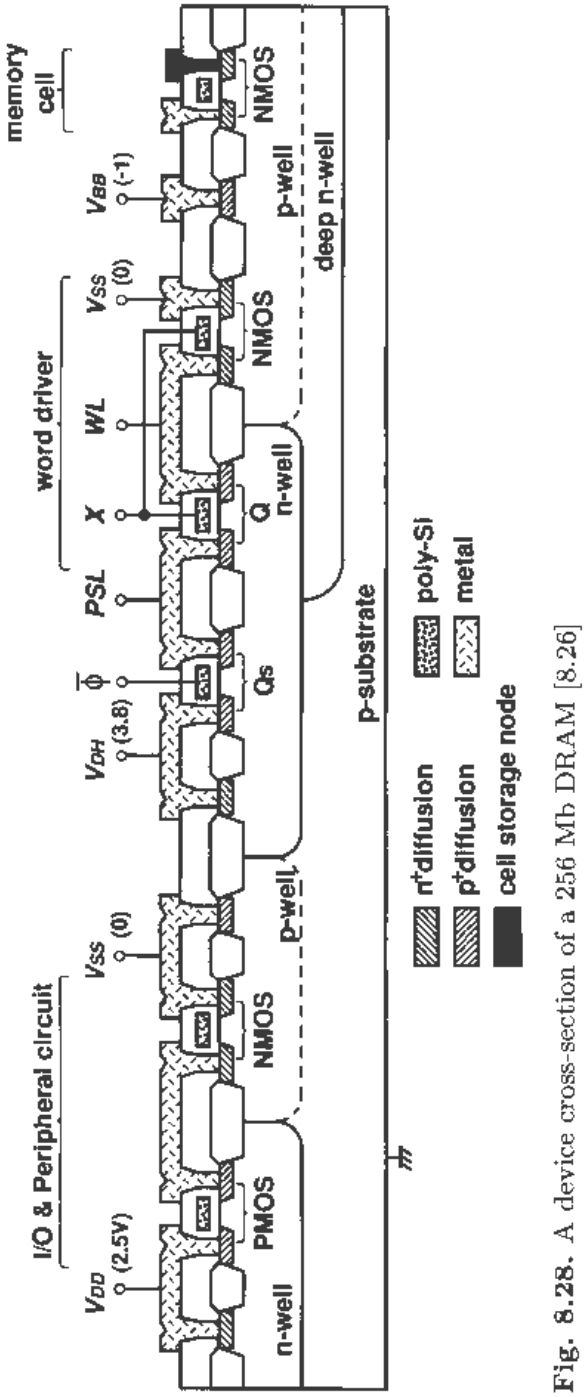


Fig. 8.28. A device cross-section of a 256 Mb DRAM [8.26]

Table 8.2. Charateristics of MOSFETs[a]

	V_{SUB-S}	T_j =75 °C (stand-by)	T_j =105 °C (operation)
PMOS	0 V	0.03 V	0.11 V
V_T	− 0.62 V	− 0.10 V	− 0.02 V
NMOS	0 V	− 0.02 V	− 0.07 V
V_T	− 0.80 V	0.14 V	0.11 V

[a]Worst condition, $V_T = V_{GS}$ at $I_{DS} = 10\,nA/15\,\mu m$.

a grounded p-substrate. The wells of the PMOSFETs and NMOSFETs are fixed to V_{DD} and V_{SS}, as usual. On the other hand, the memory-array control circuit, such as the word drivers and the memory-cell array, are formed with a triple-well structure. The n-well of the PMOSFET is fixed at a raised voltage V_{DH}, while the p-well of the NMOSFET is fixed at a negative voltage V_{BB}, to ensure stable operation of the memory-cell array. The deep n-well is necessary to isolate the V_{BB} p-well from the grounded p-substrate. Table 8.2 shows the characteristics of MOSFETs. Weak depletion characteristics at no substrate-source back-bias enable high-speed active operation of the I/O and peripheral circuits. When back-biases are supplied, the MOSFETs change to low-V_T enhancement MOSFETs. Here, the subthreshold current of the MA CTL must be reduced, because it dominates the stand-by chip current. Figure 8.29 shows the major circuits that are relevant to the subthreshold current in the MA CTL. A hierarchical word-line architecture (see Fig. 3.54), composed of main word lines (MWLs) and subword lines (SWLs), is adopted. In the conventional design in which a low V_T, no substrate (i.e. well) bias, and no SSI scheme are used, there are three major sources of subthreshold current in stand-by mode. They are the column (Y select line, YL) drivers, the main word drivers (WDs), and the row selection (RX) drivers. In order to reduce the subthreshold current, more attention was paid to the PMOSFETs, since the PMOS subthreshold current is usually much larger than the NMOS subthreshold current. In fact, the S-factor and the total channel width involving the subthreshold current were 118 mV/dec. (at 105 °C) and 868 mm for the PMOSFETs in the MA CTL, and 97 mV/dec. and 468 mm for the NMOSFETs. Thus, the SSI scheme was applied to the PMOSFETs. Moreover, the substrate-source back-biasing scheme utilizing the above internal supply voltages was applied to further reduce the PMOS current with an increased V_T, although it was not applied to the PMOSFETs that operate on V_{DH} because a further raised power supply was not available. As for the NMOSFETs, the subthreshold current was only sufficiently confined by application of substrate back-bias.

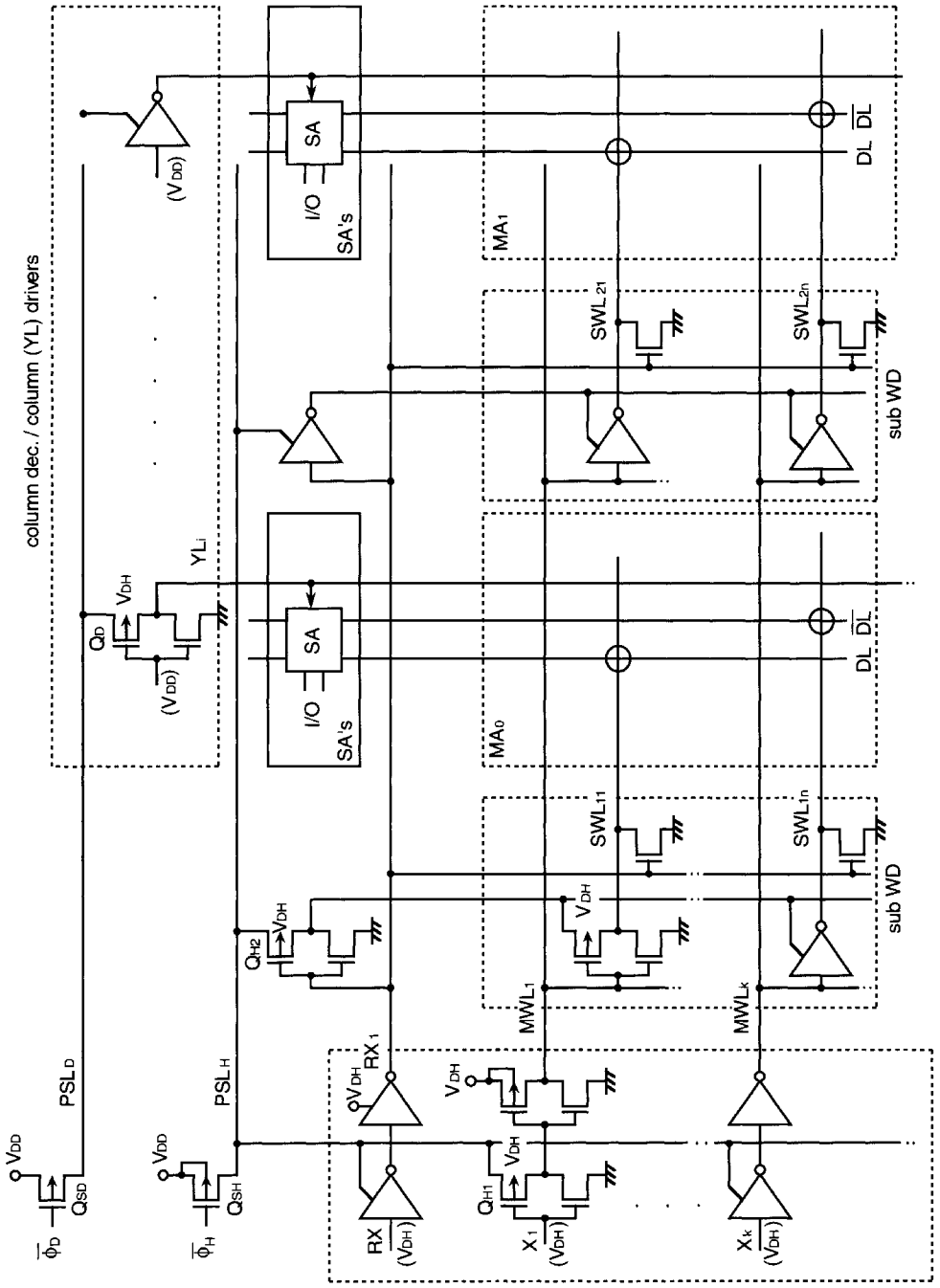


Fig. 8.29. A memory-array control circuit block (MA CTL) and its voltage supplying scheme [8.26]. All NMOSFETs in the MA CTL are back-biased at V_{BB} ($-1\,V$). A memory array is composed of 256 word lines and 256 pairs of data lines

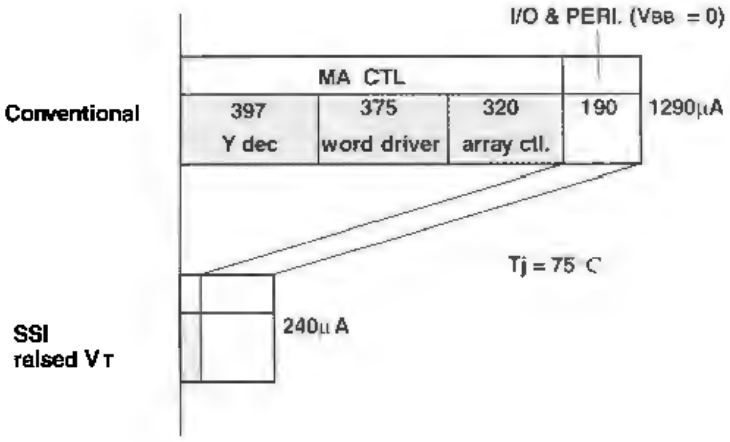


Fig. 8.30. The stand-by current of a 256 Mb DRAM [8.26]

The subthreshold current of the column driver is reduced by using V_{DH}-well-biased PMOSFETs and an SSI that is validated by a PSL_D-line voltage that decreases after turning off Q_{SD} with high-level Φ_D. Fortunately, the circuit can operate on V_{DD} because the maximum data-line voltage is limited to 2 V, allowing application of V_{DD} to the column selection line (YL). The leakage currents of the main word driver (Q_{H1}) and the RX driver (Q_{H2}) are also reduced in a similar manner, although the circuits operate on V_{DH}. The SSI, combined with an increased V_T reduces the stand-by current of the 256 Mb DRAM [8.26] from 1.3 mA to 240 μA, as shown in Fig. 8.30. Eventually, the subthreshold current of the I/O and peripheral circuits dominates the stand-by chip current.

The subthreshold current in active mode is another concern for iterative circuit blocks, although it can be reduced in stand-by mode by the circuit described above. After one selected word line is activated, all of the drivers

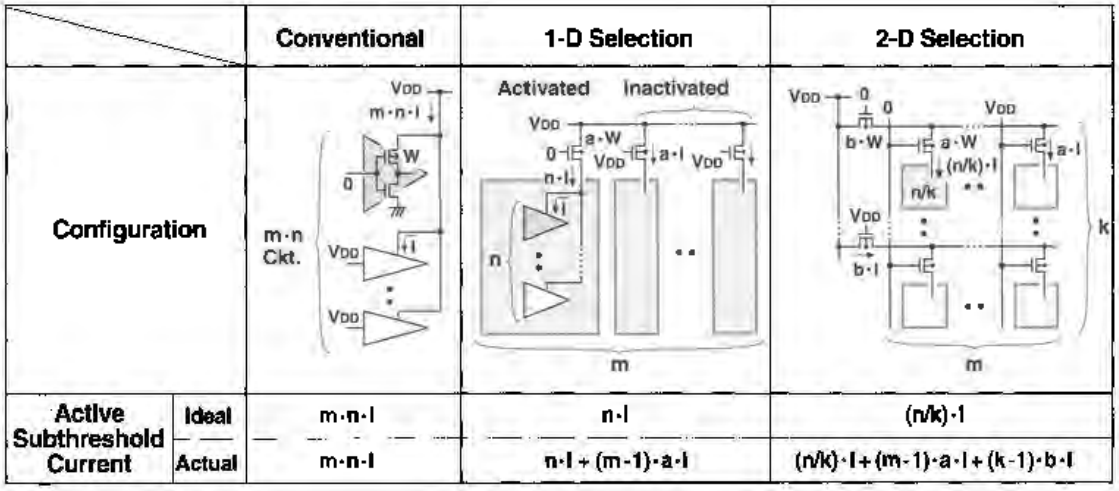


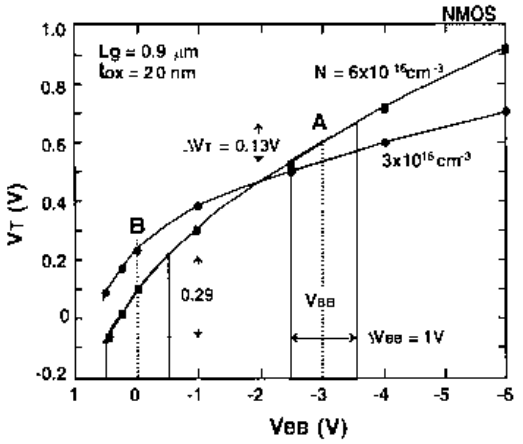
Fig. 8.31. The partial activation of a multidivided power line [8.17, 8.27]

are sources of subthreshold current, which eventually dominates the total active current. This is overcome by partial activation of the multidivided power line [8.17, 8.27], as shown in Fig. 8.31. One-dimensional (1-D) power-line selection features a selective power supply to part of the circuit block by dividing it into m sub-blocks, each consisting of n circuits. The operation is performed by turning on a switch corresponding to a selected (activated) sub-block, while the others remain off. All the non-selected (inactivated) sub-blocks substantially have no subthreshold current, since the same voltage relationship as in stand-by mode in Fig. 8.26 is established in each sub-block. This reduces the current to nI with an m-fold reduction. For further reduction of the current, two-dimensional (2-D) selection is effective. In this configuration, a circuit block having $m \cdot n$ circuits is divided into $m \cdot k$ sub-blocks, m in a row and k in a column. The subthreshold current is reduced, in inverse proportion to the number of sub-blocks, to $(n/k) \cdot I$ with an $m \cdot k$-fold reduction.

8.3.4 The Well Control Circuit

Variable V_T schemes through well control are becoming increasingly important. However, careful attention should be paid to instabilities involved in the scheme, that are expected from the long history of DRAM development. It is well known that the DRAM has been the only one large-volume production LSI using a substrate bias voltage (with a substrate–source voltage difference of V_{BB}) which is supplied from an on-chip V_{BB} generator. However, the DRAM has been lucky because both the setting of a deep V_{BB} of about -2 to -3 V and a sufficiently high V_T of about 0.5 V have enabled stable chip operation.

Figure 8.32 shows the relationship between V_T and V_{BB}. The V_T of an NMOSFET increases as V_{BB} becomes deeper. The tendency is enhanced by



	active		noise		inactive	
	V_{BB}	V_T	$\frac{\Delta V_T}{\Delta V_{BB}}$	$\frac{\Delta V_T}{V_T}$	V_{BB}	V_T
static V_T A (STD DRAM)	deep (-3)	high (0.5)	small	small	deep (-3)	high (0.5)
variable V_T A↔B (R&D)	shallow (0)	low (0.1)	large	large	deep (-3)	high (0.5)

Fig. 8.32. The relationship between V_T and V_{BB} [8.5]

increasing the substrate doping concentration, N. In the standard DRAM, V_T is static (fixed) because of the same V_{BB} for both active and inactive periods. Moreover, both the setting of a deep V_{BB} of $-3\,V$ and a high V_T of $0.5\,V$ contribute to stable chip operation. A deep V_{BB} of $-3\,V$ causes a small change in V_T (ΔV_T) despite a change in V_{BB} (ΔV_{BB}, i.e. noise) of as much as $1\,V$. Even for a large N, ΔV_T is $0.13\,V$. Thus, a small ΔV_T and a high V_T result in a small ΔV_T-to-V_T ratio, causing stable operation despite a V_{BB} noise. In the variable V_T scheme, V_{BB} is changed so that shallow V_{BB} (point B) is for the active period, and a deep V_{BB} (point A) is for the inactive period. Thus, we can realize fast operation with a low V_T in the active period, and a low subthreshold current with a high V_T in the inactive period. However, the operation (at $V_{BB} = 0\,V$) in the active period is susceptible to V_{BB} noise. Figure 8.33 illustrates the substrate-noise generation mechanism [8.62]. A V_{BB} of $0\,V$ during active operation means no difference between the substrate and source voltages. A voltage difference (i.e. substrate noise), however, is actually developed during high-speed active operation, because the substrate line and the source line are separated and thus the voltages coupled to each line can be different. Note that in a modern DRAM, as in MPUs and ASICs, the tight connection between the source (that is, a power line of V_{DD} or $0\,V$) and the well in each MOSFET never creates a source to well voltage difference, thus ensuring a fixed V_{BB} of $0\,V$ throughout the chip. Obviously, the noise causes a large V_T variation because V_T drops more sharply with a shallower substrate bias. This would result in unstable active operation, especially at low V_{DD} (i.e. low V_T), with a large ΔV_T-to-V_T ratio.

A large change in V_T, which is the key to the scheme, can be attained by switching V_{BB} from a deeper V_{BB} during the stand-by period to a shallower V_{BB} during the active period with a larger ΔV_{BB} and/or a large K, as shown in Fig. 8.14. Note that ΔV_{BB} corresponds to Δ in Fig. 8.13. Here,

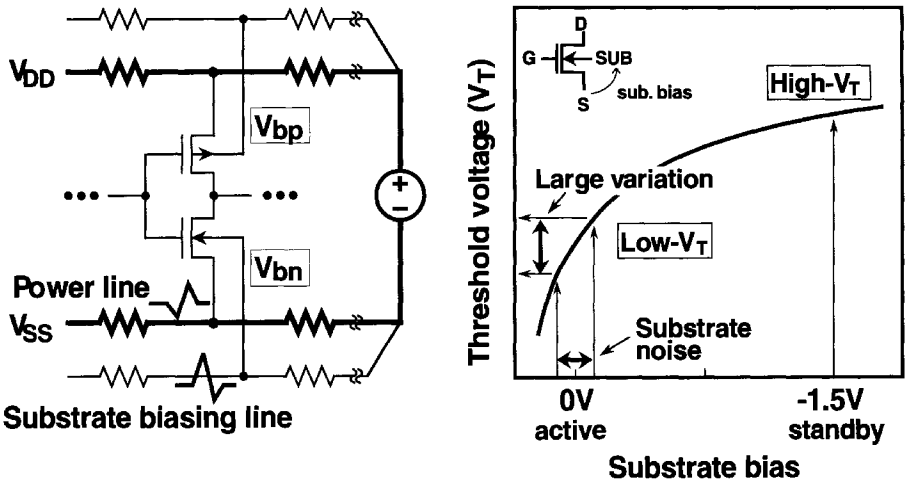


Fig. 8.33. Substrate noise [8.62]

the shallowest V_{BB}, including the above substrate noise, must not exceed about $V_{BB} = +0.6\,V$. Otherwise, the p–n-junction formed at the substrate and source is forward-biased. This restriction makes a shallow V_{BB} setting to below 0 V during the active period hazardous. A large ΔV_{BB} requires an additional high external supply voltage, and this fails to realize single power-supply operation at low voltages. Although a larger K is desired, the realization of this is controversial, because K tends to be smaller with smaller FETs [8.65]. A large K also necessitates a higher word-line bootstrapping voltage and causes a slower speed for the AND logic circuit, as explained in Fig. 2.5.

Power-on CMOS latch-up and/or the rush current tend to be enhanced by a lower V_T (even for an enhancement MOSFET in some cases) that is developed at a shallow V_{BB} of around 0 V during power-on. This is because the heavy substrate capacitance is slowly charged up from a floating 0 V to a floating V_{BB} (for example, $-3\,V$) due to the poor current-driving cability of an on-chip V_{BB} generator that comprises a charge pump, as explained in Chap. 5.

There have been four proposals: control of the common well (substrate) for the whole chip, so as to simultaneously change all V_T values in the chip, control of only the common well of a certain circuit block in the chip, control of each well of the individual circuit, and control of the well of the individual MOSFET. It should be noted that common-well control cannot manage the subthreshold current in active mode, because high-speed control of the heavy well capacitance is difficult. On the other hand, well controls of the circuit block, the individual circuit, and the individual MOSFET could reduce the subthreshould current even in active mode because of lighter well capacitances.

Well Control of the Chip. Figure 8.34 shows a V_{BB} driving scheme [8.29] for a 0.3 μm CMOS LSI. To obtain a low V_T of 0.1 V in the active mode and a high V_T of 0.5 V in the stand-by mode, the V_{BB} of the p-well is changed from $-0.5\,V$ to $-3.3\,V$ while the V_{BB} of the n-well is changed from 1.4 V to 4.2 V to establish the same bias condition for $V_{DD} = 0.9\,V$. The large change in V_T of 0.4 V comes from a large FET (i.e. large K) of 0.3 μm and a large V_{BB} swing, as explained previously. The V_{BB} in the active period is well regulated by low-frequency pumping controlled by leakage current monitors (LCMs) after completing high-speed well discharge due to high-frequency pumping. In the sleep (stand-by) mode high-frequency pumping again discharges the well to a deep V_{BB}. At the beginning of the transition from sleep to active mode, a substrate injector consisting of a CMOS inverter quickly charges up the well to around $-0.5\,V$. In principle, V_{BB} driving can inherit the traditional circuit and design methodology, because only V_{BB} control is required. In addition to the I_{DDQ} test, adjustment of the chip-to-chip V_T variations by an appropriate V_{BB} setting can be achieved. The V_{BB} response time, however, is as slow as a few hundreds of μs. This is because on-chip V_{BB} generators (Chap. 5) cannot

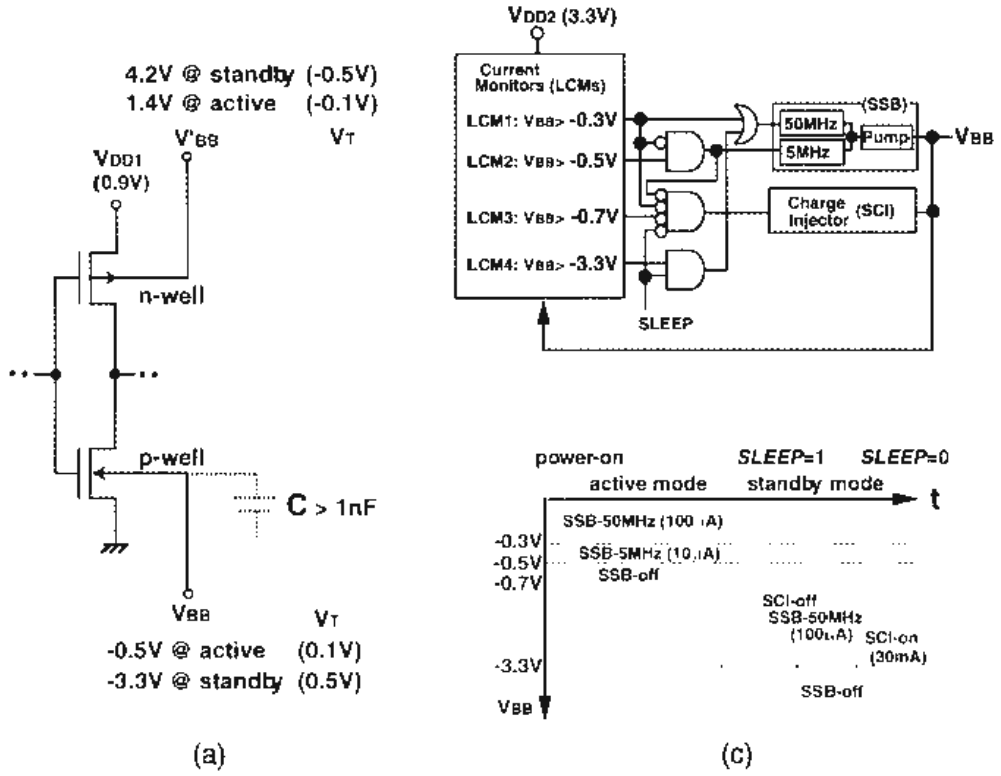


Fig. 8.34. The dynamic V_T scheme due to well driviug [8.29]. (**a**) The concept; (**b**) V_{BB} driving circuit; (**c**) V_{BB} response

quickly drive a well capacitance that is heavier than 1 nF with a V_{BB} swing as large as 2.8 V. Thus, this V_{BB} control is not suitable for the reduction of the subthreshold current in the active period, which requires a high-speed V_{BB} control. A large V_{BB} swing also needs an additional external supply voltage of 3.3 V, which excludes it from a single voltage supply scheme. The ac and dc instabilities of V_{BB}, which originate from a floating substrate (well), which may be caused, as discussed before.

Figure 8.35 depicts another example of well-driving (i.e. a switched substrate-impedance scheme) applied to an actually designed dual-power supply (i.e. 1.8 V, 3.3 V) 0.2 μm 200 MHz microprocessor [8.62] that incorporates millions of transistors. In order to reduce the well–source voltage difference of each MOSFET as much as possible during the active period, the source line and the well line are connected at every 200 μm by turning on high-V_T (0.45 V) switch cells. In stand-by mode the cells are turned off to supply deep well-biases of 3.3 V and -1.5 V (from an on-chip generator) to the PMOSFETs and NMOSFETs, respectively. Thus, coupled with the power supply and substrate bias grids to ensure uniform voltages throughout the chip, high-speed stable active-operation with a low V_T of 0.15 V can be obtained. The stand-by current is reduced from 1.3 mA to 47 μA due to a high V_T of about 0.3 V. The active to stand-by mode transition time, which is the

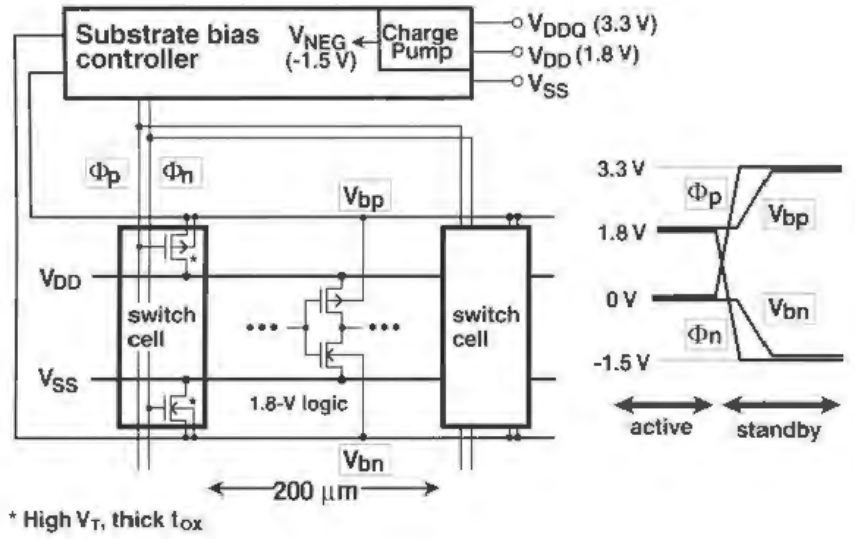


Fig. 8.35. A substrate-voltage switching scheme [8.62]

time needed for the charge pump to drive the wells, is about 50 μs, while the stand-by to active mode transition time is only 370 ns. Such slow responses reduce the surge currents at the transitions. The area penalty of the switch cells was about 5% of the chip size.

Well Control of a Circuit Block. Figure 8.36 shows a common well-driving scheme [8.33] applied to a sense-amplifier block that suffers from an inherently slow speed, as explained previously. This scheme, combined with a triple-well structure, achieves a low V_T without well-bias during amplification, so that the speed is enhanced. A high V_T is attained by applying enough bias just after the sensing/restoring operation. Here, the NMOSFETs are located in a p-well isolated from the p-type substrate by an n-well and a deep n-well, while the PMOSFETs are located in the n-well. In the equalizing

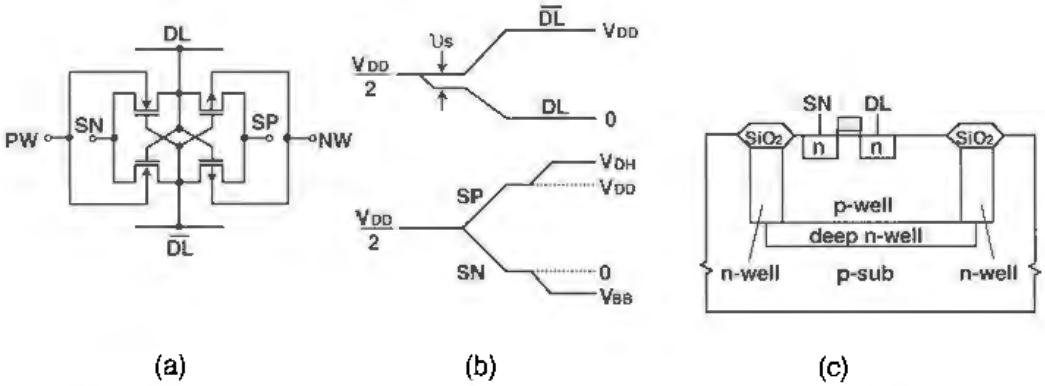


Fig. 8.36. Sense-amplifier well driving [8.33]. (a) The circuit; (b) timing; (c) a cross-section of the NMOS

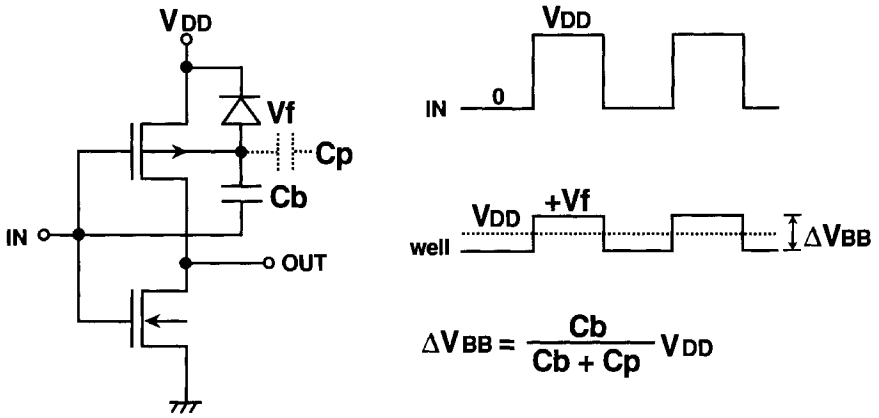


Fig. 8.37. Capacitor-coupled well driving [8.34]

period the wells return to the half-V_{DD} level. The resultant low V_T boots the current so that the data lines are quickly equalized.

Well Control of an Individual Circuit. Figure 8.37 shows well-driving [8.34] for an individual circuit. It features capacitor-coupled driving in which the MOSFET well is dynamically connected to the gate by a capacitor, so that V_T is automatically adjusted during operation. Hence it enables a much higher current when turned on (effectively a low V_T) and a much lower leakage current when turned off (effectively a high V_T). A diode is used to discharge C_b. Automatic V_T control is important in designing a complicated chip, since all the design issues involving the subthreshold current are confined to the individual circuit level. However, there are drawbacks of an increased loading capacitance for the previous stage, and an area penalty due to an additional capacitor.

Well Control of an Individual MOSFET. It is ideal if the well of an individual MOSFET can be controlled independently. This is realized by dynamic V_T (DTMOS) by utilizing a SOI structure in which the gate is directly connected to the well. Its operating voltage, however, is strictly limited to less than 0.8 V, as discussed later.

8.3.5 The Source Control Circuit

The V_T can be changed by changing the MOSFET source voltage under a fixed well-bias voltage. Figure 8.38 shows source-driving schemes [8.30, 8.31]. In the stand-by (sleep) mode, the common source voltage of NMOSFETs in the core circuit is raised, while that of PMOSFETs is lowered to increase V_T. In Fig. 8.38a this is accomplished by controlling the gate voltages of the output MOSFETs (Q_N, Q_P) of the power management unit. The design parameter, temperature, and voltage variations are automatically compensated by precisely controlling the gate voltages, which is accomplished by the use of a CMOS delay line, a phase detector, and charge pumps in the power

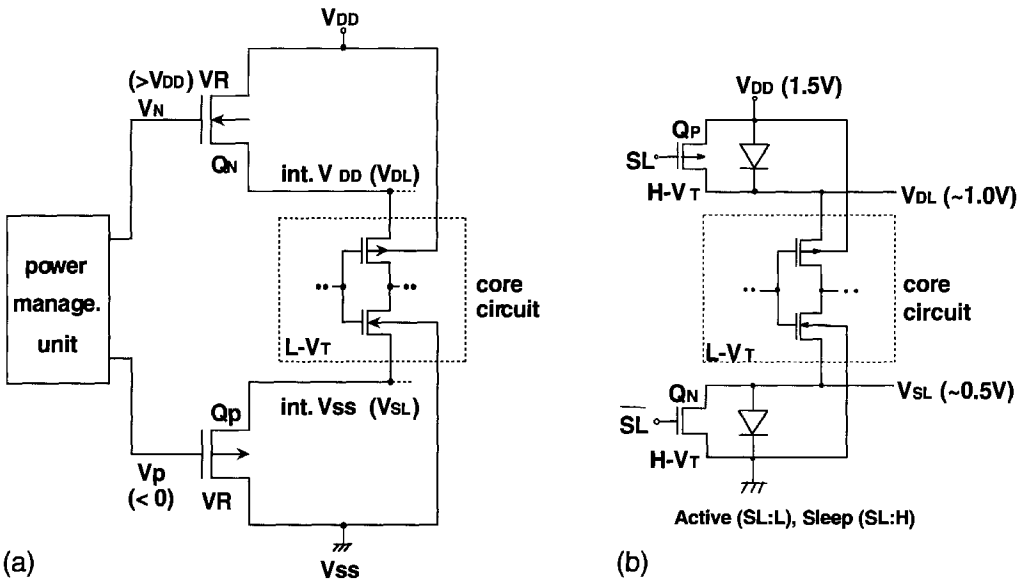


Fig. 8.38. Source driving schemes with a fixed well-bias voltage [8.30, 8.31]. The source voltages of the MOSFETs in the internal circuit are controlled by a power management unit (**a**) and diodes (**b**)

management unit. However, a spike source noise, developed when operating the internal core circuit, would be a major obstacle to stable operation. This is because Q_N and Q_P cannot manage a large current without an area penalty. The slow response time of internal power lines when they are switched, as in V_{BB} driving, is another problem. In addition, V_{DD} must be higher than the internal supply voltage, thus limiting low-V_{DD} operation of the chip. In Fig. 8.38b, low-V_T MOSFETs in the core circuit are deeply biased by a diode drop, so as to raise their V_T values.

8.3.6 The Well and Source Control Circuit

Figure 8.39 shows a combination of well- and source-driving [8.32], although it needs two external power supplies. A sufficiently low V_T, realized by the absence of gate–source back-bias in the active mode, achieves a high speed even at 1 V. In stand-by (sleep) mode the potentials of V_{DL} and V_{SL} begin to rise due to the subthreshold current that flows from V_{DD2} to ground through the core circuit. This rise stops when the currents of both diodes (D_1 and D_2) become equal to the subthreshold current of the core circuit. Consequently, V_{DL} and V_{SL} are around 2 V and 1 V, respectively. Then, the V_T of every transistor in the internal circuit increases, because their substrates are reverse-biased by around 1 V.

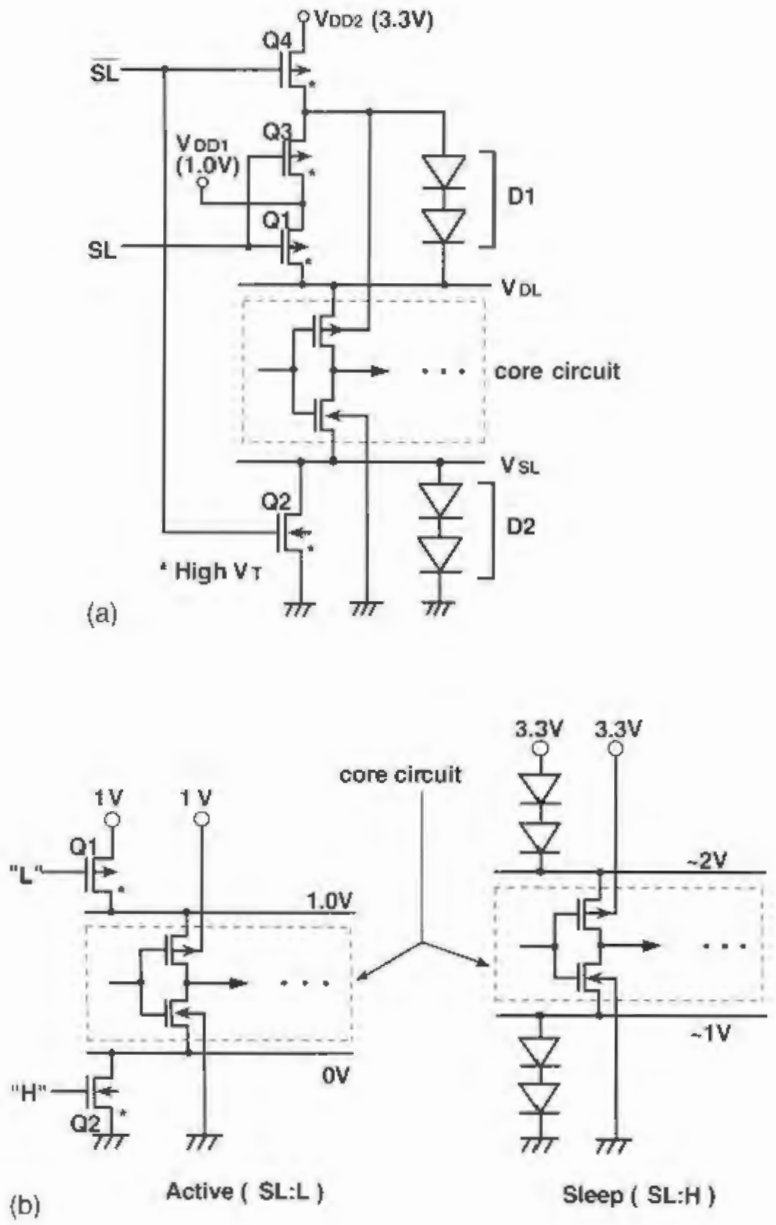


Fig. 8.39. An auto-backgate-controlled multi-V_T CMOS circuit [8.32]. (a) The actual circuit; (b) an equivalent circuit for active and sleep modes

8.4 Ultra-Low-Voltage SRAM Circuits

In ultra-low-voltage operation, the SRAM array is a major concern in terms of both the subthreshold current and V_T mismatch. It is the largest channel-width block dominating the subthreshold current of the chip, unlike the situation in DRAMs. In addition, it has the largest number of flip-flop circuits, whose operations are sensitive to the V_T mismatch of paired MOSFETs.

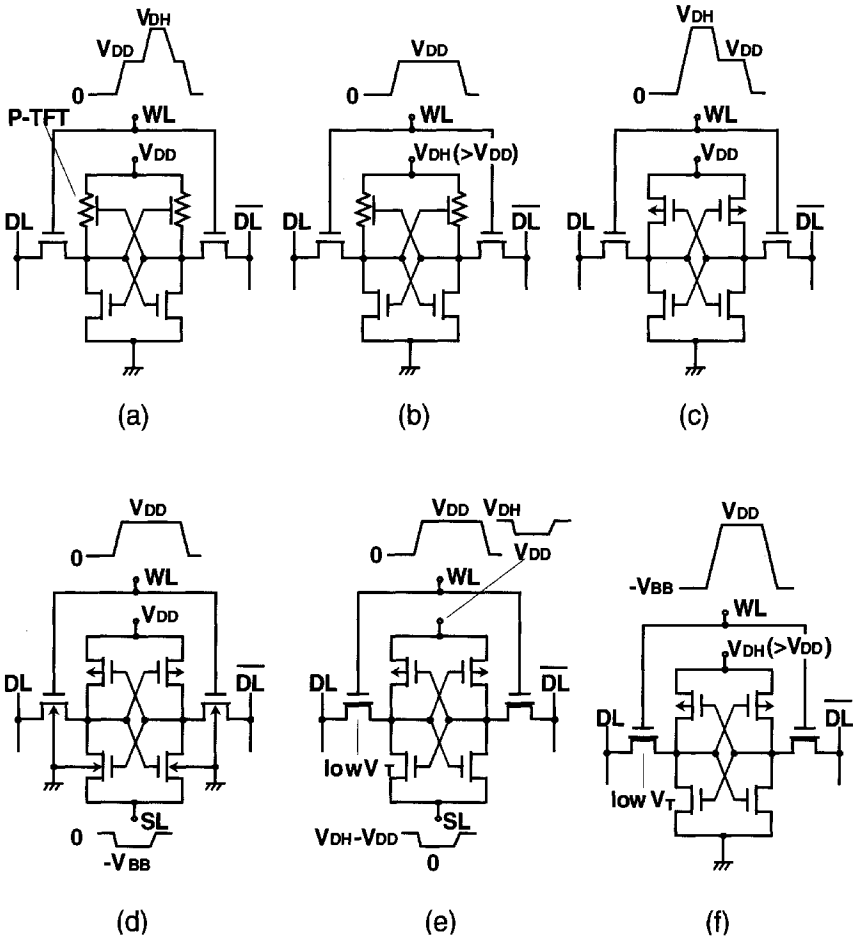


Fig. 8.40. Various SRAM cell-driving schemes. (**a**) Two-step word-voltage [8.36]; (**b**) raised $V_{DD}(= V_{DH})$ [8.37]; (**c**) step-down boosted word line [8.38]; (**d**) negative source line [8.39]; (**e**) offset source line [8.40]; (**f**) boosted storage node [8.41]

Note that the current in the remaining circuit blocks could be suppressed with circuits similar to those of DRAMs, previously described.

A high V_T is needed for cell driver FETs to suppress a huge subthreshold current, as discussed before. In principle, a high V_T is also necessary for cell-transfer FETs, to avoid a leakage current flowing to either of the data lines. A high V_T, however, decreases the cell voltage-margin as the cell supply voltage (V_{DD}) approaches V_T, limiting the minimum V_{DD} of the chip. Figure 8.40 shows various cell-driving schemes to overcome this limitation. Two-step word-driving [8.36] enables a high-speed write operation directly from the data lines by boosting the word line to eliminate the V_T drop of the transfer FETs. The poor driving capability of TFT cell loads makes the above word-driving operation necessary despite a write-speed penalty caused by the necessity for a preceding read operation. The raised dc voltage [8.37], generated by an on-chip voltage up-converter to supply the cell loads, allows the storage node voltage of the cell to rise quickly during a write operation

because of the increased TFT conductance. Both schemes, however, suffer from a slow read operation because of the high V_T of the transfer FETs. The step-down boosted word line [8.38] offers high-speed operation as well as low power. However, additional delay in boosting, and variations in the amplitude and duration of the boosted pulse caused by process variations, are involved. The negative source-line scheme [8.39] also achieves high speed with a reduced V_T in cell driver FETs and a boost effect for the cell-transfer FETs. However, a heavy capacitance, which almost equals the data-line capacitance multiplied by the number of selected cells, established at the source line, prevents single-V_{DD} operation. This is because an on-chip negative-voltage generator comprised of charge-pumping circuits can never cope with such a heavy capacitance. The offset source-line scheme [8.40] solves the heavy capacitance issue, requiring an on-chip voltage up-converter instead of the above negative-voltage generator. In the above cells, the minimum V_{DD} may be around 1 V in a practical design in which soft errors, V_T variations, and V_T mismatch – in addition to a high V_T – are considered. In particular, a V_T mismatch that continues to increase between paired FETs in a cell along with FET miniaturization may limit low-voltage developments in the future, as discussed previously. Hence the boosted storage-node scheme [8.41] is effective in lowering the minimum V_{DD} down to less than 0.5 V, with the help of an on-chip voltage up-converter. A low V_T for transfer FETs combined with negative word-line biasing also helps high-speed operation. However, the challenge is to achieve a sufficiently high voltage generation at low power. Obviously, voltage generation is not needed if a higher externally supplied voltage becomes available.

Figure 8.41 shows well-driving along the selected word line of an SRAM array [8.42] which is similar to that of a DRAM (Fig. 8.36). The n- and p-well voltages are dynamically changed to V_{DD} (0.5–1.0 V) and V_{SS} (0 V), respectively, when the memory cells are activated. When unselected, the well voltages recover to V_{NWELL} (about $2V_{DD}$) and V_{PWELL} (about $-V_{DD}$). Thus, the V_T of the activated cells becomes low (~0 V) for fast operation, while that of the non-activated cells is high (~0.25 V) for a low subthreshold current. The scheme features a low stress voltage to gate oxide, which differs from the schemes that use a boosted power supply. Instead, in addition to a small signal charge of the non-activated memory cells, high stress voltages applied to the well junctions, and an area penalty, which is caused by an additional well for each cell and each well driver, are design concerns.

The dynamic data-line load [8.36, 8.39, 8.41, 8.56] that is particular to DRAMs is best in terms of power, although the static (ratio) load has been common in SRAMs. The voltage swing on the data line, however, must be sufficiently reduced. The write power is greatly reduced by a small-signal differential write operation, combined with half-V_{DD} data-line precharge [8.39]. In this operation, a small signal written into the cell is amplified by turning on a cell feedback loop after isolation of the data lines. Control of word-

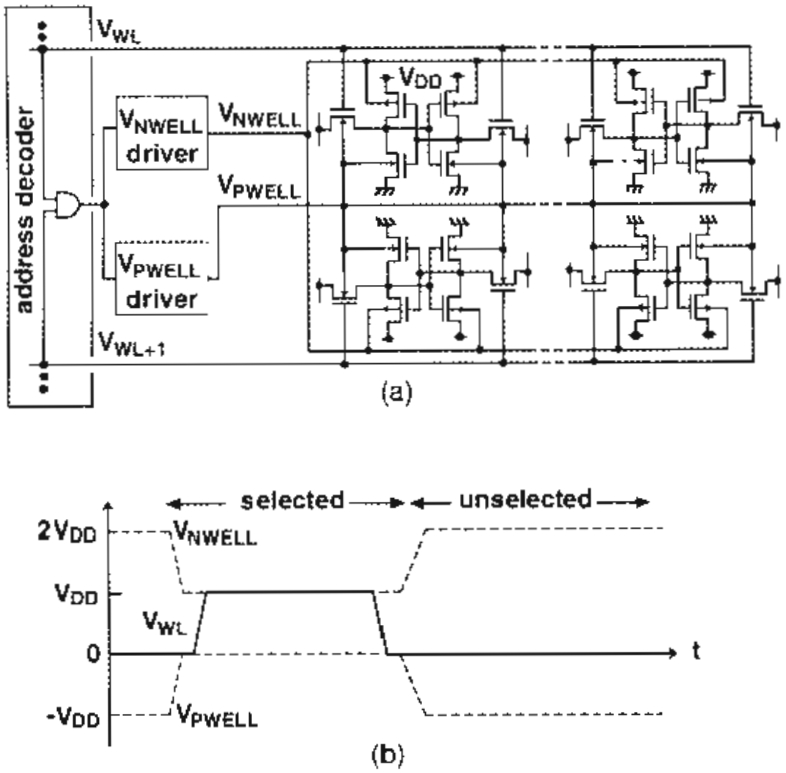


Fig. 8.41. Well driving along the selected word line [8.42]

line replica feedback [8.57] precisely limits the read signal amplitude against various design-parameter variations.

8.5 Ultra-Low-Voltage SOI Circuits

SOI CMOS technology may meet the requirements of ultra-low-voltage operation and/or gigabit DRAMs better than the bulk CMOS technology [8.43, 8.44] discussed thus far. The small junction areas of the source and drain that are completely isolated by the SiO_2 layer, as shown in Fig. 8.42, reduce the junction capacitance, the leakage current, and the critical charge of a soft error. The small body (substrate) of each FET, that is formed between only the thick SiO_2 layer and the gate area, results in a small subthreshold swing (a small S-factor), fewer short-channel effects, and a reduction in the back-gate-bias effect [8.45]. Less capacitance in the SOI body favors dynamic V_T control, which is attained by body control, and this lowers the operating voltage. In addition, the resulting FET structure is simple and free from latch-up. Nevertheless, some problems still remain unsolved. These are: the floating body effect of a DRAM cell FET, which degrades data retention and soft error characteristics; the effect of the FET structure, which inherently increases the thermal resistance; and the formation of a damage-free structure

and the low-cost preparation of a thick SiO_2 layer. The following is a summary of state-of-the-art SOI circuit technology.

Figure 8.43 shows a 3.3 V 0.6 μm floating body 64 Kb DRAM test chip [8.43] which was fabricated on a SIMOX (Separation by Implanted OXygen) wafer. Oxygen was implanted at an energy of 190 KeV and at a dose of 1.8×10^{18} /cm^2, and high-temperature annealing was performed at 1320 °C for 6 hours. The reduced C_D/C_S, which arose from reducing the junction capacitance in the data line, increased the read signal voltage by 25%. The access time was also improved by about 35% at 3 V, due to the reduced junction capacitance and back-gate-bias effect in the peripheral circuits. It has also been reported that, at 2 V supply voltage, an SOI CMOS SRAM operates twice as fast as its bulk CMOS counterpart [8.46]. A body-bias control

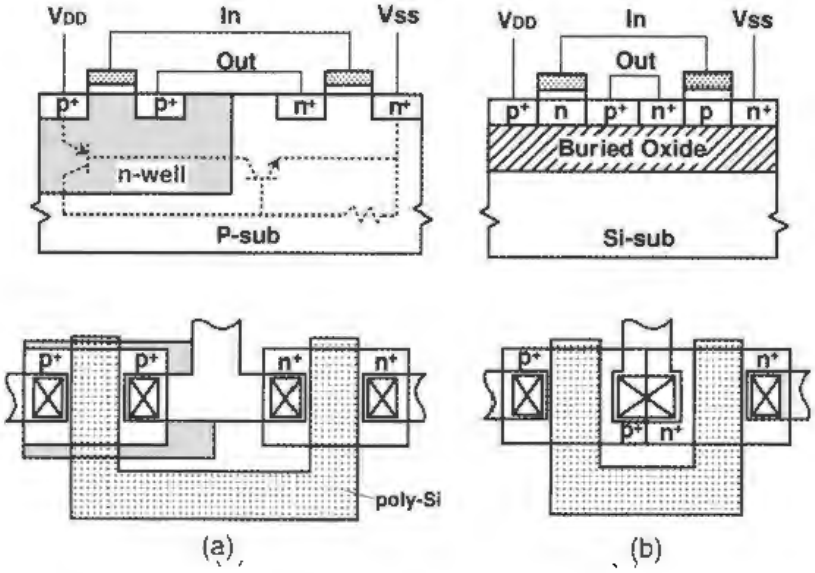
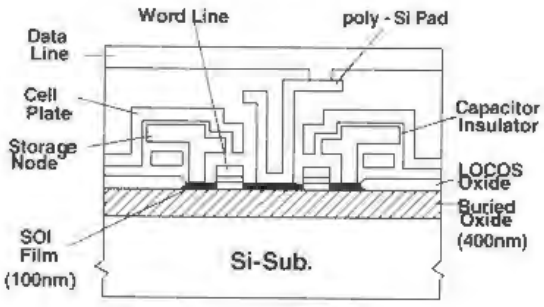


Fig. 8.42. SOI CMOS versus bulk CMOS. (a) A bulk CMOS inverter; (b) a SOI CMOS inverter



		SOI	BULK
Cb/Cs (128cells/DL)		9 (216/24fF)	12 (288/24fF)
cell signal	3.0V	150 mV	115 mV
	1.5V	58 mV	38 mV
access time	4.0V	52 ns	70 ns
	3.0V	66 ns	98 ns
	2.3V	85 ns	—
active current (3V,260ns cycle)		1.1mA	3.2mA

Fig. 8.43. A stacked-capacitor SOI DRAM cell and experimental results of a 64 Kb chip compared with a bulk CMOS counterpart [8.43]. Memory cell: 5.12 μm^2, $V_{tn} = 0.8$ V, $L_n = 0.7$ μm. Peripheral circuit: $V_{Tn}/V_{Tp} = 0.43\,\text{V}/-0.6\,\text{V}$, $L_n/L_p = 0.6/0.6$ μm

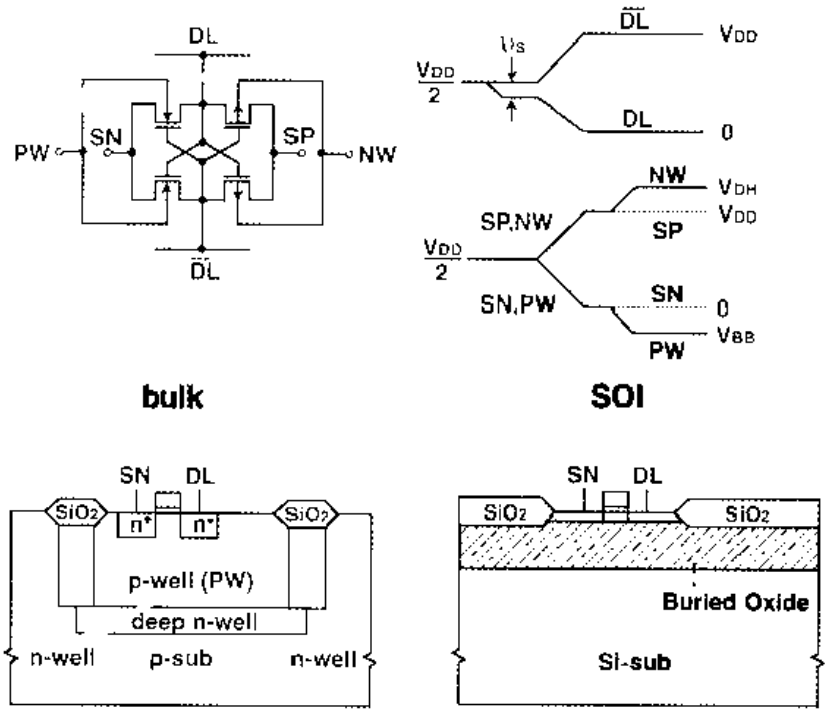


Fig. 8.44. Body-bias control of a sense amplifier [8.47, 8.48]

technique to achieve a variable V_T further lowered the minimum operating voltage, down to 1.5 V or less [8.47, 8.48]. Figure 8.44 shows SOI body control for a scuse amplifier compared with bulk CMOS body control. When the sense amplifier is activated by applying a negative-going pulse to SN and a positive-going pulse to SP, both the NMOS body (PW) and the PMOS body (NW) are also driven. This condition of no body-bias causes a low V_T. Just after a sensing operation, sufficicent bias is provided by additional NW and PW pulses, so that a high V_T is obtained for a low subthreshold current. In the bulk CMOS, this is realized by a triple-well structure in which NMOSFETs, for example, are located in a p-well isolated from the p-type substrate by an n-well and deep n-well. Thus, the p-well has a heavy capacitauce, preventing high-speed, low-power p-well driving. On the contrary, the SOI structure solves the problem because of a small body. Figure 8.45a shows an NMOSFET for a hypothetical 1.5 V 4 Gb DRAM [8.48]. The FET is partially depleted, and thus V_T can be controlled by the body-bias voltage. Note that even when the voltage differenco (V_{bs}) betweeu the body and source is $+0.5$ V and the p–n junction is thus forward-biased, the current flow is negligible. Figure 8.45b shows the body-bias control logic applied to peripheral circuits. During a stand-by cycle, the p-body V_{pb} is set to -1.5 V and V_T is set to 0.65 V. The variable V_T scheme can also be applied to the sense amplifier, as in the bulk CMOS previously described. The expected current of the peripheral circuits was approximately 1/20th that of the bulk, and the access time of the SOI was 35% faster than that of the conventional bulk. Such high performance

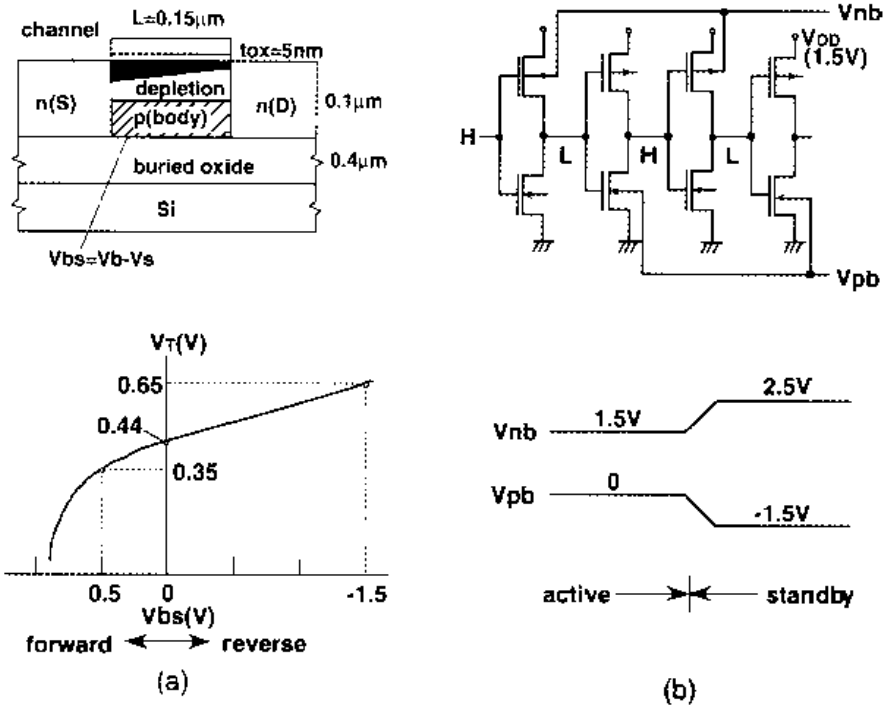


Fig. 8.45. The body-bias control technique applied [8.48]. (**a**) V_T controlled by V_{bs}; (**b**) logic circuits

has been verified using an actual 0.5 μm CMOS/SIMOX 16 Mb DRAM [8.47]. Body-bias control attained a high-speed access time of 46 ns even with an operating voltage as low as 1 V. The body-current clamper, which suppresses the body current caused by body-bias overshooting, and a negative word-line voltage scheme, contribute to stable memory-chip operation.

The high-speed control of body-bias in the logic is difficult, since their bodies are connected, and the resulting body capacitance and resistance are consequently heavy despite the SOI structure. A dynamic-V_T MOSFET (DTMOS) built on an SOI [8.49] may solve this problem, because the control is confined to an individual circuit, as shown in Fig. 8.37. In it, the body is connected to the gate, as shown in Fig. 8.46, and it has a low V_T when the FET is turned on for a high current. The DTMOS can even achieve 0.5 V operation [8.50] in the feedback buffer in the figure, but its operating voltage is strictly limited to less than around 0.8 V due to the forward bias of the body-source p–n junction. This drawback is overcome by the power-switch FET (Qp) [8.51] shown in Fig. 8.47a. As a result of forward bias at the p–n diode, a large leakage current flows from the body to the gate if the V_{DD} is over 0.8 V, which is the diode built-in potential, as shown in Fig. 8.47b. The insertion of a reverse-biased low-V_T MOS diode between the body and the gate permits a higher V_{DD}, as shown in Fig. 8.47c. The low-V_T diode is one-tenth the Qp size. For $V_{DD} > 0.8$ V, the diode clamps the forward bias of the Qp p–n diode, suppressing the gate-leakage current. Thus, in the active

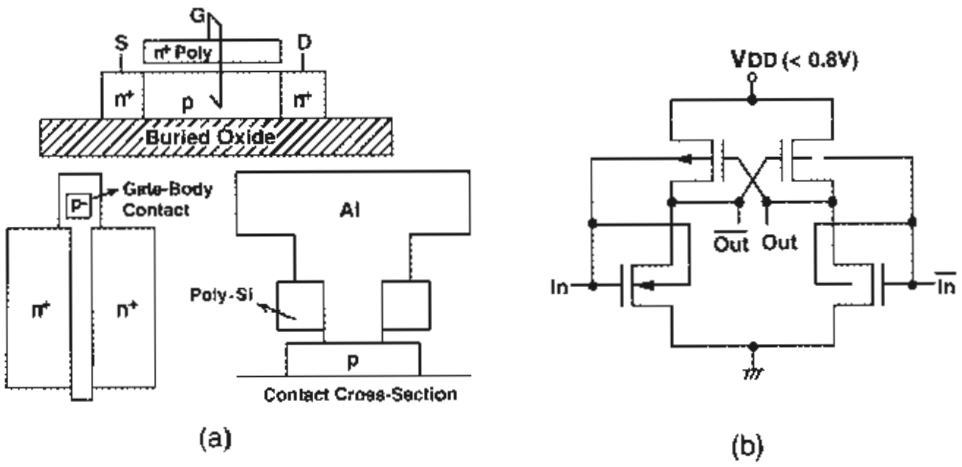


Fig. 8.46. The DTMOS concept featuring a body tied to the gate [8.49] and its application to a 0.5 V PMOS feedback buffer [8.50]. (**a**) Connection; (**b**) application

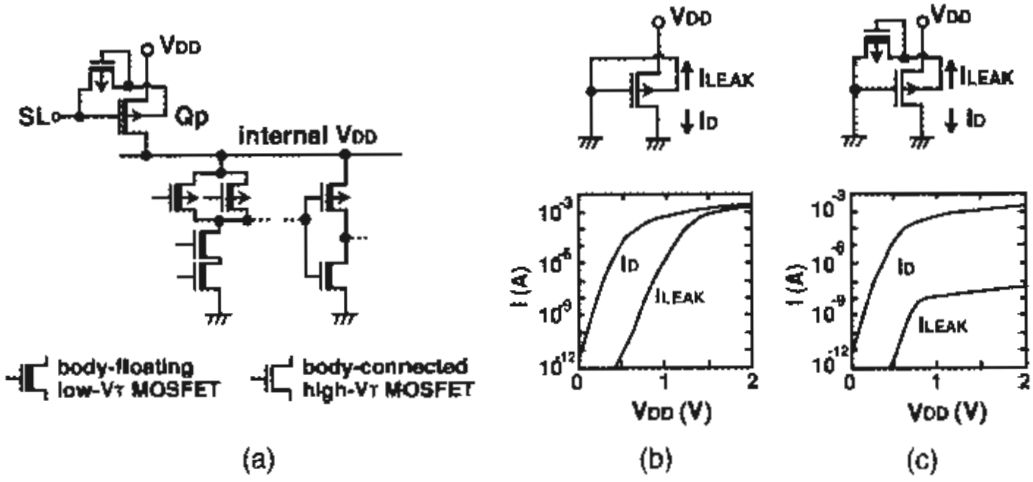


Fig. 8.47. A SIMOX-multi-V_T CMOS circuit applied to a power-switch MOSFET [8.51]. (**a**) The circuit; (**b**) the DTMOS; (**c**) the MOS with a diode

period at low-level SL, Qp provides V_{DD} to the internal V_{DD} line with small leakage current; while in the stand-by period at high-level SL, Qp completely isolates the internal V_{DD} line from the V_{DD} line without any subthreshold current.

The major problem in the SOI structure is closely related to the floating body effect in the DRAM cell transistor: the instability of the floating body potential degrades the data-retention characteristics and soft-error immunity. Figure 8.48 shows the degradation mechanism for the data-retention characteristics [8.43, 8.52]. The data-retention time is defined as the time until the cell voltage stored by the write operation decays to almost half V_{DD}. This decay stems from the stored charges (holes) lost by p–n junction leakage ① at the cell storage node. The resulting accumulated holes at the body raise the

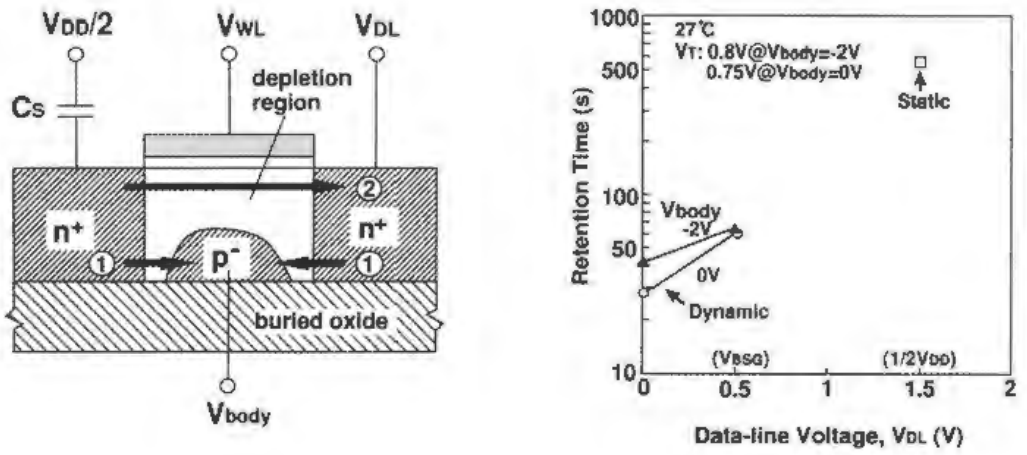


Fig. 8.48. The data-retention characteristics of a SOI DRAM cell [8.52]. The V_{DL} of 1.5 V ($= 1/2V_{DD}$) is for the static mode, while the V_{DL} values of 0 V and 0.5 V are for the dynamic mode and the BSG scheme, respectively

body potential due to its small capacitance, causing a reduced V_T of the cell transistor, which is the origin of the subthreshold leakage current ②. In the static mode, an SOI DRAM cell achieves a superior data-retention time of 550 s at 27 °C, which is six times longer than that of a bulk memory cell [8.53]. In this mode the cell transistor is completely cut off despite the reduced V_T, because the data line remains at a high level of a half V_{DD} (the precharge voltage of the data line) and the word line is maintained at a low level of 0 V during retention. Thus, the long retention time comes from the small area of the p-n junction. Unfortunately, the dynamic mode that keeps the data-line

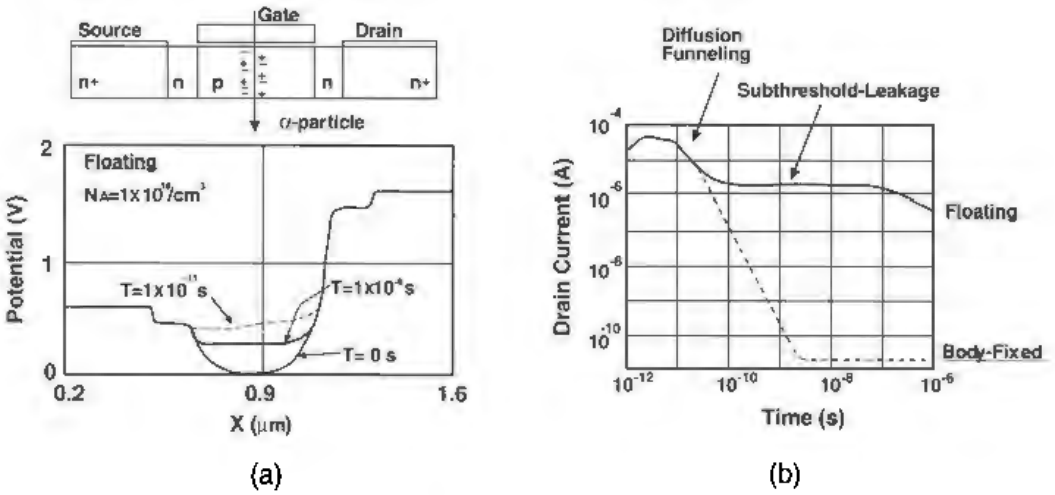


Fig. 8.49. The potential increase (**a**) and resulting subthreshold leakage current (**b**) caused by an α particle incident on the body region of an SOI transistor [8.43]. $L = 0.5\,\mu$m, $W = 2\,\mu$m, $t_{SOI} = 100$ nm, and $t_{SiO_2} = 400$ nm

voltage at 0 V shortens the data-retention time to 42 s, as shown in Fig. 8.48. This results from the subthreshold current caused by the reduced V_T. The boosted sense-ground (BSG) scheme described previously, which raises the lowest data-line voltage from 0 V to 0.5 V, improves the characteristics to some extent. The body-refresh scheme [8.54], combined with the BSG, can produce a long data-retention time, although additional refresh operations for the body node are necessary.

The floating body also degrades the soft-error immunity of MOSFETs, not only in DRAM and SRAM cells [8.43, 8.55], but also in peripheral circuits. Figure 8.49 shows variations in simulated body potentials and leakage currents over time after an α-particle incidence for floating and body-fixed SOI MOSFETs. Electrons which are generated in the floating body diffuse to the source and drain, while holes remain in the floating body region and raise the potential. This potential increase causes a large continuous subthreshold current. The introduction of body contacts [8.43] to suppress body-potential change is very effective in peripheral circuits. For memory cells, however, it increases the memory cell area, especially in DRAMs.

References

Chapter 1

[1.1] L.M. Terman, "Memory at ISSCC", ISSCC Commemorative Supplement to the Digest of Technical Papers, pp. 91–111, Feb. 1993.

[1.2] H. Sunami, "Coping with Memory-Cell Miniaturization by Using New Materials", Nikkei Microdevices, pp. 144–157, Dec. 1997.

[1.3] T. Murotani et al., "A 4-Level Storage 4 Gb DRAM", ISSCC Dig. Tech. Papers, pp. 74–75, Feb. 1997.

[1.4] R.H. Dennard, "Field-Effect Transistor Memory", U.S. Patent 3387286, June 4, 1968.

[1.5] F.M. Wanlass, C.T. Sah, "Nanowatt Logic Using Field-Effect Metal-Oxide Semiconductor Triodes", ISSCC Dig. Tech. Papers, pp. 32–33, Feb. 1963.

[1.6] D.F. Benchkowsky, "A Fully-Decoded 2048-bit Electrically-Programmable MOS ROM", ISSCC Dig. Tech. Papers, pp. 80–81, Feb. 1971.

[1.7] Nikkei Electronics, No. 694, pp. 83–102, July 14, 1997.

[1.8] Final Worldwide Semiconductor Market Share, 1995 and 1998, and Worldwide Memory Market Share, 1993 to 1995, Dataquest.

[1.9] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).

[1.10] Calculated with data for 64 Mb DRAMs presented at ISSCC91, ISSCC92, ESSCIRC 92 and ESSCIRC 93.

[1.11] Calculated with data for a 16 Mb CMOS SRAM presented at ISSCC 92.

[1.12] T. Takeshima et al., "A 3.3 V Single-Power-Supply 64 Mb Flash Memory with Dynamic Bit-Line Latch Programming Scheme", ISSCC Dig. Tech. Papers, pp. 148–149, Feb. 1994.

[1.13] H.K. Burke, G.J. Michon, "Charge Pump Random-Access Memory", ISSCC Dig. Tech. Papers, pp. 16–17, Feb. 1972.

[1.14] W. Martino, B.F. Croxon, "The Inverting Cell Concept for MOS Dynamic RAMs", ISSCC Dig. Tech. Papers, pp. 12–13, Feb. 1972.

[1.15] W.M. Regitz, J. Karp, "A Three-Transistor-Cell, 1024-Bit, 500 NS MOS-RAM", ISSCC Dig. Tech. Papers, pp. 42–43, Feb. 1970.

[1.16] J.A. Karp et al., "A 4096-Bit Dynamic MOS RAM", ISSCC Dig. Tech. Papers, pp. 10–11, Feb. 1972.

[1.17] K. Itoh, IEEE J. Solid-State Circuits **25**(3), 778 (1990).

[1.18] T. Masuhara et al., IEICE Trans. **E74**(1), 130 (1991).

[1.19] Y. Nakagome, K. Itoh, IEICE Trans. **E74**(4), 779 (1991).

[1.20] K. Itoh, "Reviews and Prospects of Deep Sub-Micron DRAM Technology", SSDM91, Extended Abstracts, pp. 468–471, Aug. 1991.

[1.21] K. Itoh et al., IEEE Proc. **83**(4), 524 (1995).

[1.22] K. Itoh et al., IEEE J. Solid-State Circuits **32**(5), 624 (1997).
[1.23] K. Itoh, "Ultralow-Voltage Memory Circuits", VLSI'97, Tutorial, Gramado (Brazil), Aug. 1997.
[1.24] K. Itoh et al., Electrochem. Soc. Proc. **98**(1), 350 (1998).
[1.25] H. Masuda et al., IEEE Trans. Electron. Devices **ED-27**(8), 1607 (1980).
[1.26] H. Masuda et al., IEEE J. Solid-State Circuits **SC-15**(5), 846 (1980).
[1.27] Y. Kawamoto et al., "A 1.28 μm^2 bit-line shielded memory cell technology for 64 Mb DRAMs", Symp. VLSI Technol. Dig. Tech. Papers, pp. 13–14, June 1990.
[1.28] Y. Nakagome et al., IEEE J. Solid-State Circuits **26**(4), 465 (1991).
[1.29] M. Takada, T. Enomoto, IEICE Trans. **E74**(4), 827 (1991).
[1.30] K. Ishibashi, IEICE Trans. **E79-C**(6) 724 (1996).
[1.31] M. Minami et al., "A 6.93-μm^2 n-Gate Full CMOS SRAM Cell Technology with High-performance 1.8 V Dual-Gate CMOS for Peripheral Circuits", Symp. VLSI Technol. Dig. Tech. Papers, pp. 13–14, June 1995.
[1.32] K. Ishibashi et al., "A 300 MHz 4 Mb Wave-Pipeline CMOS SRAM Using a Multi-Phase PLL", ISSCC Dig. Tech. Papers, pp. 308–309, Feb. 1995.
[1.33] F. Masuoka et al., "A New Flash EEPROM Cell using Triple Polysilicon Technology", IEDM Tech. Dig., pp. 464–467, Dec. 1984.
[1.34] F. Masuoka et al., IEICE Trans. **E74**(4), 868 (1991).
[1.35] P. Pavan et al., IEEE Proc. **85**(8), 1248 (1997).
[1.36] H. Kume, Oyo Buturi **65**(11), 1114, (1996) (in Japanese).
[1.37] V.N. Kynett et al., "An in-system reprogrammable 256 K CMOS Flash memory", ISSCC Dig. Tech. Papers, pp. 132–133, Feb. 1988.
[1.38] S. Haddad et al., Electron Device Letters, **11**(11), 514 (1990).
[1.39] F. Masuoka et al., "New ultra high density EPROM and Flash EEPROM cell with NAND structure cell", IEDM Dig. Tech. Papers, pp. 552–555, 1987.
[1.40] H. Onoda et al., "Anovel cell structure suitable for a 3 V operation, sector erase FLASH memory", IEDM Dig. Tech. Papers, pp. 599–602, 1992.
[1.41] H. Kume et al., "A 1.28 μm^2 contactless memory cell technology for a 3 V-only 64 Mbit EEPROM", IEDM Dig. Tech. Papers, pp. 991–993, 1992.
[1.42] M. Baur et al., "A Multilevel-Cell 32 Mb Flash Memory", ISSCC Dig. Tech. Papers, pp. 132–133, Feb. 1995.
[1.43] *Nonvolatile Semiconductor Memory Technology*, W.D. Brown, J.E. Brewer, Editors, IEEE PRESS, 1997.
[1.44] K. Itoh et al., "VLSI Memory Technology: Current Status and Future Trends", ESSCIRC'99 Dig. Tech. Papers, pp. 3–10, Sept. 1999.
[1.45] Kirihata et al., "A 390 mm^2 16 Bank 1 Gb DDR SDRAM with Hybrid Bitline Architecture", ISSCC99 Dig. Tech. Papers, pp. 420–421, Feb. 1999.
[1.46] S. Takase, N. Kushiyama, "A 1.6 GB/s DRAM with Flexible Mapping Redundancy Technique and Additional Refresh Scheme", ISSCC99 Dig. Tech. Papers, pp. 410–411, Feb. 1999.
[1.47] O. Takahashi et al., "1 GHz Fully Pipelined 3.7 ns Address Access Time 8 k × 1024 Embedded DRAM Macro", ISSCC Dig. Tech. Papers, pp. 396–397, Feb. 2000.
[1.48] H. Nambu et al., "A 550-ps Access, 900-MHz, 1-Mb ECL-CMOS SRAM", Symp. VLSI Circuits, June 1999.
[1.49] For example, SRAM chips in ISSCC Dig. Tech. Papers, Feb. 1999 and 2000.
[1.50] A. Nozoe et al., "A 256 Mb Multilevel Flash Memory with 2 MB/s Program Rate for Mass Storage Applications", ISSCC99 Dig. Tech. Papers, pp. 110–111, Feb. 1999.
[1.51] K. Imamiya et al., "A 130 mm^2 256 Mb NAND Flash with Shallow Trench Isolation Technology", ISSCC99 Dig. Tech. Papers, pp. 112–113, Feb. 1999.

[1.52] Byung-Gil Jeon et al., "A 0.4 μm 3.3 V 1T1C 4 Mb Nonvolatile Ferroelectric RAM with Fixed Bit-line Reference Voltage Scheme and Data Protection Circuit", ISSCC Dig. Tech. Papers, pp. 272–273, Feb. 2000.

Chapter 2

[2.1] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).
[2.2] S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (John Wiley & Sons, New York 1981).
[2.3] M.I. Elmasry, Ed., *Digital MOS Integrated Circuits 2* (IEEE Press, New York 1992).
[2.4] P.E. Allen, D.R. Holberg, *CMOS Analog Circuit Design*, (Holt and Winston, Inc.).
[2.5] E.H. Nicollian, J.R. Brews, *MOS Physics and Technology* (Wiley, New York 1982).
[2.6] T. Masuhara et al., IEICE Trans. **E74**(1), 130 (1991).
[2.7] A. Ishitani et al., IEICE Trans. Electron. **E76-C**(11), 1564 (1993).
[2.8] R.L.M. Dang, N. Shigyo, IEEE Electron Device Letters **EDL-2**(8), 196 (1981).
[2.9] Y. Kagenishi et al., "Low Power Self Refresh Mode DRAM with Temperature Detecting Circuits", 1993 Symp. VLSI Circuits Dig. Tech. Papers, pp. 43–44, May 1993.
[2.10] H.B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, (Addison-Wesley Publishing Company, Inc. 1990).
[2.11] K. Kimura et al., IEEE J. Solid-State Circuits **SC-21**(3), 381 (1986).
[2.12] H. Masuda et al., IEEE Trans. Electron. Devices **ED-27**(8), 1607 (1980).
[2.13] T. Kawahara et al., IEEE J. Solid-State Circuits **27**(4), 589 (1992).
[2.14] Y. Ushiku et al., "A Three-Level Wiring Capacitance Analysis for VLSIs using A Three-Dimensional Simulator", IEDM88, pp. 340–343 (1988).
[2.15] K. Hinode, J. Japan Institute of Light Metals **41**(9), 614 (1991).
[2.16] E. Hamdy, A. Mohsen, "Characterization and Modeling of Transient Latchup in CHMOS Technology", IEDM83 Dig. Tech. Papers, pp. 172–173 (1983).
[2.17] O.J. McAteer, *Electrostatic Discharge Control* (McGraw-Hill, Inc. 1990).
[2.18] R.H. Dennard et al., IEEE J. Solid-State Circuits **SC-9**, 256 (1974).
[2.19] P. Chatterjee et al., IEEE Electron Device Lett. **EDL-1**, 220 (1980).
[2.20] S. Okazaki et al., Appl. Surf. Sci. **70/71**, 603 (1993).
[2.21] K. Itoh, in *Low Power Design Methodologies*, J.M. Rabaey, M. Pedram, Editors (Kluwer, Norwell, MA 1995).

Chapter 3

[3.1] K. Itoh et al., IEEE J. Solid-State Circuits **32**(5), 624 (1997).
[3.2] K. Itoh et al., "VLSI memory technology: current status and future trends", ESSCIRC99 Dig. Tech. Papers, pp. 3–10, Sept. 1999.
[3.3] K. Itoh et al., IEEE Proc. **83**(4), 524 (1995).
[3.4] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).

[3.5] C.N. Ahlquist et al., "A 16 K dynamic RAM", ISSCC76 Dig.Tech.Papers, pp. 128–129, Feb. 1976.
[3.6] S.S. Eaton et al., "A 100 ns 64 K dynamic RAM using redundancy techniques", ISSCC81 Dig. Tech. Papers, pp. 84–85, Feb. 1981.
[3.7] K. Kimura et al., IEEE J. Solid-State Circuits **SC-21**(3), 381 (1986).
[3.8] R.C. Foss et al., "Application of a high-voltage pumped supply for low-power DRAM", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 106–107, 1992.
[3.9] K. Itoh et al., "A single 5 V 64 K dynamic RAM", ISSCC80 Dig. Tech.Papers, pp. 228–229, Feb. 1980.
[3.10] T. Mano et al., "Circuit technologies for 16 Mb DRAMs", ISSCC87 Dig. Tech. Papers, pp. 22–23, Feb. 1987.
[3.11] K. Itoh et al., "An experimental 1 Mb DRAM with on-chip voltage limiter", ISSCC84 Dig. Tech. Papers, pp. 282–283, Feb. 1984.
[3.12] K. Arimoto et al., "A 60 ns 3.3 V 16 Mb DRAM", ISSCC89 Dig.Tech. Papers, pp. 244–245, Feb. 1989.
[3.13] S. Fujii et al., IEEE J. Solid-State Circuits **24**(5), 1170 (1989).
[3.14] P.R. Schroeder, R.J. Proebsting, "A 16 K×1 Bit dynamic RAM", ISSCC77 Dig. Tech. Papers, pp. 12–13, Feb. 1977.
[3.15] R. Taylor, M. Johnson, "A 1 Mb CMOS DRAM with a divided bitline matrix architecture", ISSCC85 Dig.Tech. Papers, pp. 242–243, Feb. 1985.
[3.16] M. Inoue et al., "A 16 Mb DRAM with an open bit-line architecture", ISSCC88 Dig. Tech. Papers, pp. 246–247, Feb. 1988.
[3.17] T. Fujii et al., "A 90 ns 256 K×1 b DRAM with double level al technology", ISSCC83 Dig.Tech. Papers, pp. 226–227, Feb. 1983.
[3.18] K. Noda et al., "A boosted dual word-line decoding scheme for 256 Mb DRAMs", Symp. VLSI Circuits, Dig.Tech. Papers, pp. 112–113, June 1992.
[3.19] T. Sugibayashi et al., "A 30 ns 256 Mb DRAM with multi-divided array structure", ISSCC93 Dig.Tech. Papers, pp. 50–51, Feb. 1993.
[3.20] D.J. Lee et al., "A 35 ns 64 Mb DRAM using on-chip boosted power supply", Symp.VLSI Circuits, Dig.Tech. Papers, pp. 64–65, 1992.
[3.21] T. Nakano, Y. Akasaka, *ULSI DRAM Technology* (Science Forum, Tokyo 1992) (in Japanese).
[3.22] K.M. Hardee, R. Sud, IEEE J. Solid-State Circuits **SC-16**(5), 435 (1981).
[3.23] H. Ozaki et al., IEICE **J71-C**(8), 1156 (1988) (in Japanese).
[3.24] K. Matsui et al., "A study on drive method of rowdecoder", Proc. IEICE Spring Conf., C-623, 1992 (in Japanese).
[3.25] Y. Kubota et al., "Reduction of wordline noises by decoded-pulldown circuits", Proc. IEICE Spring Conf., C-626, 1992 (in Japanese).
[3.26] Y. Oowaki et al., "A 33 ns 64 Mb DRAM", ISSCC Dig. Tech. Papers, pp. 114–115, Feb. 1991.
[3.27] D. Takashima et al., IEEE J. Solid-State Circuits **27**(4), 603 (1992).
[3.28] K. Sato et al., IEEE J. Solid-State Circuits **26**(11), 1556 (1991).
[3.29] S.M. Yoo et al., IEEE J. Solid-State Circuits **28**(4), 499 (1993).
[3.30] G. Kitsukawa et al., IEEE J. Solid-State Circuits **25**(5), 1102 (1990).
[3.31] K. Komatsuzaki et al., "Circuits techniques for a wide word I/O path 64 Meg DRAM", Symp.VLSI Circuits, Dig. Tech. Papers, pp. 133–134, 1991.
[3.32] P. Gillingham et al., IEEE J. Solid-State Circuits **26**(8), 1171 (1991).
[3.33] D. Galbi et al., "A 33 ns 64 Mb DRAM with master-wordline architecture", ESSCIRC'92 Dig. Tech. Papers, pp. 131–134, 1992.
[3.34] K. Kimura et al., IEEE J. Solid-State Circuits **SC-22**(5), 651 (1987).
[3.35] G. Kitsukawa et al., IEEE J. Solid-State Circuits **SC-22**(5), 657 (1987).
[3.36] T. Kawahara et al., IEEE J. Solid-State Circuits, **26**(11), 1530 (1991).
[3.37] A. Tanabe et al., IEEE J. Solid-State Circuits **27**(11), 1525 (1992).

[3.38] T. Ooishi et al., "A well-synchronized sensing/equalizing method for sub-1.0 V operating advanced DRAMs", Symp. VLSI Circuits, Dig.Tech. Papers, pp. 81–82, May 1993.
[3.39] T. Yamada et al., "A 64 Mb DRAM with meshed power line and distributed sense-amplifier driver", ISSCC91 Dig. Tech. Papers, pp. 108–109, Feb. 1991.
[3.40] H. Hidaka et al., IEEE J. Solid-State Circuits **27**(7), 1020 (1992).
[3.41] H. Miyamoto et al., "A 32 ns 64 Mb DRAM with extended second metal line architecture", ESSCIRC'93 Dig. Tech. Papers, pp. 41–44, Sept. 1993.
[3.42] S. Watanabe et al., "BiCMOS circuit technology for high speed DRAMs", Symp.VLSI Circuits, Dig. Tech.Papers, pp. 79–80, 1987.
[3.43] T. Nagai et al., "A 17 ns 4 Mb CMOS DRAM using direct bit-line sensing technique", ISSCC91 Dig. Tech. Papers, pp. 58–59, Feb. 1991.
[3.44] G. Kitsukawa et al., IEICE **J75-C-2**(1), 17 (1992).
[3.45] Y. Nakagome et al., IEEE J. Solid-State Circuits **26**(4), 465 (1991).
[3.46] K. Kimura et al., IEICE Trans. **J68-C**(12), 1006 (1985) (in Japanese).
[3.47] M. Taguchi et al., "A 40 ns 64 Mb DRAM with current-sensing data-bus amplifier", ISSCC91 Dig. Tech. Papers, pp. 112–113, Feb. 1991.
[3.48] Y. Tsukikawa et al., "Shared read gate architecture suitable for high speed DRAMs", Proc. IEICE Spring Conf., C-631, 1992 (in Japanese).
[3.49] Y. Takano et al., "A study of data transfer scheme for DRAM", Proc. IEICE Spring Conf., C-634, 1993 (in Japanese).
[3.50] K. Sato et al., "A 20 ns static column 1 Mb DRAM in CMOS technology", ISSCC85 Dig. Tech. Papers, pp. 254–255, Feb. 1985.
[3.51] H.L. Kalter et al., IEEE J. Solid-State Circuits **25**, 1118 (1990).
[3.52] M. Taniguchi et al., IEEE J. Solid-State Circuits **SC-16**, 492 (1981).
[3.53] K. Nogami et al., IEEE J. Solid-State Circuits **SC-21**, 662 (1986).
[3.54] K. Sawada et al., "Self-aligned refresh scheme for VLSI intelligent dynamic RAMs", Symp. VLSI Technology Dig.Tech. Papers, pp. 85–86, May 1986.
[3.55] Y. Miyamoto et al., "Study of new refresh method for low data retention current", Proc. IEICE Spring Conf., C-638, 1993 (in Japanese).
[3.56] K. Kenmizaki et al., "A 36 μA 4 Mb PSRAM with quadruple array operation", Symp.VLSI Circuits, Dig. Tech. Papers, pp. 79–80, 1989.
[3.57] Y. Konishi et al., IEEE J. Solid-State Circuits **25**(5), 1112 (1990).
[3.58] T. Kawahara et al., "A charge recycle refresh for Gb-scale DRAMs in file applications", Symp.VLSI Circuits, Dig. Tech. Papers, pp. 41–42, May 1993.
[3.59] Cenker et al., IEEE Trans. Electron Devices **ED-26**, 853 (1979).
[3.60] Mano et al., IEEE J. Solid-State Circuits **SC-17**, 726 (1982).
[3.61] T. Smith et al., IEEE J. Solid-State Circuits **SC-16**, 506 (1981).
[3.62] K. Shimohigashi et al., "Redundancy techniques for dynamic RAMs", Proc. 14th Conf. Solid State Devices, pp. 63–67, Aug. 1982.
[3.63] M. Horiguchi, "Redundancy Techniques for High-Density DRAMs", Proc. International Conf. on Innovative Systems in Silicon, pp. 23–29, 1997.
[3.64] U.S. Patent 5 631 862, U.S. Patent 5 734 617.
[3.65] M. Kumanoya et al., IEEE J. Solid-State Circuits **SC-20**, 909 (1985).
[3.66] S. Fujii et al., IEEE J. Solid-State Circuits **SC-18**, 441 (1983).
[3.67] M. Horiguchi et al., IEEE J. Solid-State Circuits **26**(1), 12 (1991).
[3.68] G. Kitsukawa et al., IEEE J. Solid-State Circuits **28**, 1105 (1993).
[3.69] K. Arimoto et al., IEEE J. Solid-State Circuits **25**, 11 (1990).
[3.70] R. Hori et al., IEEE J. Solid-State Circuits **SC-19**, 634 (1984).
[3.71] K. Itoh, IEEE J. Solid-State Circuits **25**, 778 (1990).
[3.72] Sugibayashi et al., IEEE J. Solid-State Circuits **28**, 1092 (1993).
[3.73] K. Sasaki et al., IEEE J. Solid-State Circuits **24**, 1219 (1989).
[3.74] H. Yamauchi et al., IEEE J. Solid-State Circuits **28**, 1084 (1993).

[3.75] K. Ishibashi et al., IEEE J. Solid-State Circuits **29**, 411 (1994).
[3.76] M. Asakura et al., "A hierarchical bit-line architecture with flexible redundancy and block compare test for 256 Mb DRAM", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 93–94, May 1993.
[3.77] T. Kirihata et al., IEEE J. Solid-State Circuits **31**, 558 (1996).
[3.78] K. Furutani et al., "A board level parallel test and short circuit failure repair circuit for high-density, low-power DRAMs", Symp.VLSI Circuits, Dig. Tech. Papers, pp. 70–71, June 1996.
[3.79] Asakura et al., "A 34 ns 256 Mb DRAM with boosted sense-ground scheme", ISSCC, Dig. Tech. Papers, pp. 140–141, Feb. 1994.
[3.80] R. Sud, K.C. Hardee, "Redundancy", Electronics, July 28, 1981, pp. 121–126.
[3.81] K. Kokkonen et al., "Redundancy techniques for fast static RAMs", ISSCC81 Dig. Tech. Papers, pp. 88–81, Feb. 1980.
[3.82] B. Keeth, "Redundancy approaches for maximum yield", Memory Design and Evolution, Symp. VLSI Circuits, June 1998.
[3.83] Y. Inoue et al., "An 85 ns 1 Mb DRAM in a plastic DIP", ISSCC'85 Dig. Tech. Papers, pp. 238–239, Feb. 1985.
[3.84] M. Horiguchi et al., IEEE J. Solid-State Circuits **26**(1), 12 (1991).
[3.85] K. Arimoto et al., IEEE J. Solid-State Circuits **25**, 11 (1990).
[3.86] J. Inoue et al., "Parallel testing technology for VLSI memories", Proc. Int. Test Conf., pp. 1066–1071, 1987.
[3.87] K. Arimoto et al., ISSC91 J. Solid-State Circuits **24**, 1184 (1989).
[3.88] S. Mori et al., "A 45 ns 64 Mbit DRAM with a Merged Match-Line Test Architecture", ISSCC91 Dig. Tech. Papers, pp. 110–111, Feb. 1991.
[3.89] T. Sugibayashi et al., "A distibrutive serial multi-bit parallel test scheme for large capacity DRAMs", *Symp. VLSI Circuits*, Dig. Tech. Papers, pp. 63–64, 1993.
[3.90] Y. Nakagome, K. Itoh, IEICE Trans., **E74**(4), 799 (1991).
[3.91] T. Takashima et al., IEEE J. Solid-State Circuits **25**, 903 (1990).

Chapter 4

[4.1] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).
[4.2] K. Itoh et al., Electrochemical Society Proc. **98-1**, 350 (1998).
[4.3] K. Itoh, "Ultralow-Voltage Memory Circuits", VLSI '97 Tutorial, Gramado (Brazil) (1997).
[4.4] T.C. May, M.H. Woods, "A New Physical Mechanism for Soft Errors in Dynamic Memories", Proc. Reliab. Physics Symp., pp. 33–40, April 1978.
[4.5] K. Itoh, in *Low Power Design Methodologies*, J.M. Rabaey, M. Pedram, Editors (Kluwer, Norwell, MA 1995).
[4.6] K. Shimotori et al., "A 100 ns 256 K DRAM with Page-Nibble Mode", ISSCC Dig. Tech. Papers, pp. 228–229, Feb. 1983.
[4.7] M. Koyanagi et al., "Novel High Density, Stacked Capacitor MOS RAM", IEDM Tech. Dig., pp. 348, 1978.
[4.8] H. Sunami et al., IEEE Trans. Electron Devices **ED-31**, 746 (1984).
[4.9] N.C.C. Lu, "Advanced Cell Structure for Dynamic RAMs", IEEE Circuits and Devices Mag., pp. 27–36, Jan. 1989.
[4.10] T. Ema et al., IEICE Trans. Electron. **E76-C**(11), 1564 (1993).
[4.11] S. Kimura et al., IEEE Trans. Electron Devices **37**(3), 737 (1990).

[4.12] M. Sakao et al., "A Capacitor-Over-Bit-Line (COB) Cell with a Hemispherical-Grain Storage Node for 64 Mb DRAMs", IEDM Ext. Abst., pp. 655–658, Dec. 1-1990.
[4.13] A. Ishitani et al., IEICE Trans. Electron. **E76-C**(11), 1564 (1993).
[4.14] T. Kuroiwa et al., Jpn. J. Appl. Phys. **33-1**(9B), 5187 (1994).
[4.15] K. Koyama et al., "A Stacked Capacitor with $(Ba_xSr_{1-x})TiO_3$ for 256 M DRAM", IEDM Ext. Abst., pp. 823–826, Dec. 1991.
[4.16] K. Itoh et al., Proc. IEEE **83**(4), 524 (1995).
[4.17] K. Itoh et al., IEEE J. Solid-State Circuits **32**(5), 624 (1997).
[4.18] K. Ohyu et al., Jpn. J. Appl. Phys. **28**(6), 1041 (1989).
[4.19] D.S. Yaney et al., IEEE Trans. Electron Devices **ED-26**, 10 (1979).
[4.20] K. Takeuchi et al., "Experimental Characterization of a-Induced Charge Collection Mechanism for Megabit DRAM Cells", Symp. VLSI Tech. Dig. Tech. Papers, pp. 99, May 1987.
[4.21] Y. Konishi et al., IEEE J. Solid-State Circuits **24**, 35 (1989).
[4.22] M. Aoki et al., IEEE J. Solid-State Circuits **23**(5), 1113 (1988).
[4.23] H. Hidaka et al., IEEE J. Solid-State Circuits **24**(1), 21 (1989).
[4.24] M. Yoshida et al., "Scaled Bit Line Capacitance Analysis using a Three-Dimensional Simulator", Symp. VLSI Tech. Dig. Tech. Papers, pp. 66–67, 1985.
[4.25] K.U. Stein, A. Sihling, E. Doering, "Storage Array and Sense/Refresh Circuit for Single-Transistor Memory Cells", ISSCC Dig. Tech. Papers, pp. 56–57, Feb. 1972.
[4.26] K. Itoh, IEEE J. Solid-State Circuits **25**(3), 778 (1990).
[4.27] H. Masuda et al., IEEE J. Solid-State Circuits **SC-15**(5), 846 (1980).
[4.28] K. Shimotori et al., Trans. IECE **J61-C**(6), 399 (1978).
[4.29] Y. Nagayama et al., Trans. IECE **J65-C**(7), 522 (1982).
[4.30] J.J. Barnes, J.U. Chan, IEEE J. Solid-State Circuits, **SC-15**(5), 831 (1980).
[4.31] P.R. Schroeder, R.J. Proebsting, "A 16 K × 1 Bit Dynamic RAM", ISSCC Dig. Tech. Papers, pp. 12–13, Feb. 1977.
[4.32] J.M. Lee et al., "A 80 ns 5 V-only 16 K dynamic RAM", ISSCC Dig. Tech. Papers, pp. 142–143, Feb. 1979.
[4.33] S. Fujii et al., "A 50 μA Standby 1 MW × 1 b/256 KW × 4 b CMOS DRAM", ISSCC Dig. Tech. Papers, pp. 266–267, Feb. 1986.
[4.34] K. Itoh et al., "An Experimental 1 Mb DRAM with On-Chip Voltage Limiter", ISSCC Dig. Tech. Papers, pp. 282–283, Feb. 1984.
[4.35] S. Fujii et al., "A 45 ns 16 Mb DRAM with Triple-Well Structure", ISSCC Dig.Tech. Papers, pp. 248–249, Feb. 1989.
[4.36] R. Kraus, K. Hoffmann, IEEE J. Solid-State Circuits **24**(4), 895 (1989).
[4.37] H. Geib et al., IEEE J. Solid-State Circuits **27**(7), 1028 (1992).
[4.38] Y. Watanabe et al., "Offset Compensating Bit-Line Sensing Scheme for High Density DRAMs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 116–117, 1992.
[4.39] E.J. Sprogis, Proc. IEEE 1991. Int. Conference on Microelectronic Test Structures **4**(1), 103 (1991).
[4.40] S. Suzuki, M. Hirata, IEEE J. Solid-State Circuits **SC-14**(6), 1066 (1979).
[4.41] T. Furuyama et al., "A New Sense Amplifier Technique for VLSI Dynamic RAM's", IEDM Ext. Abst., pp. 44–47, Dec. 1981.
[4.42] L.G. Heller et al., IEEE J. Solid-State Circuits **SC-11**(5), 596 (1976).
[4.43] L.G. Heller, "Cross-Coupled Charge-Transfer Sense Amplifier", ISSCC Dig. Tech. Papers, pp. 20–21, Feb. 1979.
[4.44] M. Aoki et al., IEEE J. Solid-State Circuits **24**(4), pp. 889–894, Aug. 1989.
[4.45] M. Aoki et al., Trans. IEICE, **J73-C-II**(5), 310 (1990).

[4.46] M. Aoki et al., "A 1.5 V DRAM for Battery-Based Applications", ISSCC Dig. Tech. Papers, pp. 238–239, Feb. 1989.
[4.47] T. Nakano et al., "A Sub 100 ns 256 Kb DRAM", ISSCC Dig. Tech. Papers, pp. 224–225, Feb. 1983.
[4.48] S.S. Eaton et al., "A Sub 100 ns 64 K Dynamic RAM using Redundancy Techniques", ISSCC Dig.Tech. Papers, pp. 84–85, Feb. 1981.
[4.49] H. Kawamoto et al., "A 288 Kb CMOS Pseudo SRAM", ISSCC Dig. Tech. Papers, pp. 276–277, Feb.1984.
[4.50] R.I. Kung et al., "A Sub 100 ns 256 K DRAM in CMOS III Technology", ISSCC Dig. Tech. Papers, pp. 278–279, Feb.1984.
[4.51] S. Suzuki et al., IEEE J. Solid-State Circuits **SC-19**(5), pp. 624–627, Oct. 1984.
[4.52] S. Saito et al., IEEE J. Solid-State Circuits, **SC-20**(5), 903 (1985).
[4.53] J.Y. Chan et al., IEEE J. Solid-State Corcuits **SC-15**(5), pp. 839–846, 1980.
[4.54] Y. Takemae et al., "A 1 Mb DRAM with 3-Dimensional Stacked Capacitor Cells", ISSCC Dig. Tech. Papers, pp. 250–251, Feb. 1985.
[4.55] N.C.C. Lu et al., IEEE J. Solid-State Circuits **SC-20**(6), pp. 1272–1276, 1985.
[4.56] A. Koike, *Key issues in manufacturing of Giga era* (VLSI Technology Workshop Digest 1996).
[4.57] T. Hamamoto et al., "Well Concentration: A Novel Scaling Limitation Factor Derived from DRAM Retention Time and Its Modeling", IEDM Ext. Abst., pp. 915–918, Dec. 1995.
[4.58] M. Kojima et al., "Optimization of Giga-bit DRAM Cell Transistors by Channel and Drain Engineering ", SSDM Ext. Abst., pp. 36–37, Hiroshima, 1998.
[4.59] H. Suzuki et al., "Trap Assisted Leakage Mechanism of 'worst' Junction in Giga-bit DRAM Using Negative Word-Line Voltage", SSDM Ext. Abst., pp. 32–33, Hiroshima, 1998.
[4.60] W.R. McKee et al., "Cosmic Ray Neutron Induced Upsets as a Major Contributor to the Soft Error Rate of Current and Future Generation DRAMs", Int. Rel. Phys. Symp., pp. 1–6, April 1996.
[4.61] S. Satoh et al., "Scaling Law for Secondary Cosmic-Ray Neutron-Induced Soft Errors in DRAMs", SSDM Ext. Abst., pp. 40–41, Hiroshima, 1998.
[4.62] A. Eto et al., "Impact of Neutron Flux on Soft Errors in MOS Memories", IEDM Ext. Abst., pp. 367–370, Dec. 1998.
[4.63] G. Bonner et al., "A Fully Planarized 0.25 mm CMOS Technology for 256 Mbit DRAM and Beyond", Dig. Tech. Papers, 1995. symp. VLSI Technology, pp. 15, 1995.
[4.64] K. Itoh et al., "VLSI Memory Technology: Current Status and Future Trends", ESSCIRC'99 Dig. Tech. Papers, pp. 3–10, Sept. 1999.
[4.65] Y. Ohji et al., "Reliability of Nano-Meter Thick Multi-Layer Dielectric Films on poly-Crystalline Silicon", Tech. Dig. of International Reliability Physics symposium, p. 55, 1987.
[4.66] T. Kisu et al., "A Novel Storage Capacitance Enlargement Structure Using a Double-Stacked Storage Node in STC DRAM Cell", Ext. Abstract, 20th Conf. On Solid State Devices and Materials, p. 581, 1988.
[4.67] I. Asano et al., "1.5 nm Equivalent Thickness Ta_2O_5 High-k Dielectric with Rugged Si Suitable for Mass Production of High Density DRAMs", IEDM Tech. Dig., p. 755, 1998.
[4.68] H. Shinriki et al., IEEE Trans. Electron Devices, **ED-37**, 1939 (1990).
[4.69] H. Watanabe et al., "A New Cylindrical Capacitor using Hemisoherical Grained Si (HSG-Si) for 256 Mb DRAMs", IEDM Tech. Dig., p. 259, 1992.

[4.70] K. Ono et al., "(Ba, Sr)TiO_3 Capacitor Technology for Gbit-Scale DRAMs", IEDM Tech. Dig., p. 803, 1998.
[4.71] K. Sunouchi et al., "Process Integration for 64 M DRAM using an Asymmetrical Stacked Trench Capacitor (AST) Cell", IEDM Tech. Dig., p. 647, 1990.
[4.72] H. Ishiuchi et al., "Embedded DRAM Technologies", 1995 Symp. VLSI Technology, p. 33, 1997.

Chapter 5

[5.1] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).
[5.2] K. Itoh et al., Proc. IEEE **83**(4), 524 (1995).
[5.3] K. Itoh et al., IEEE J. Solid-State Circuits **32**(5), 624 (1997).
[5.4] K. Itoh, "Ultralow-voltage memory circuits", VLSI'97 Tutorial, Gramado (Brazil), 1997.
[5.5] R.D. Pashley, A. McCormick, "A 70 ns 1 K MOS RAM", ISSCC Dig. Tech. Papers, pp. 138–139, Feb. 1976.
[5.6] K. Itoh et al., "A single 5 V-only 64 K dynamic RAM", ISSCC Dig. Tech. Papers, pp. 228–229, Feb. 1980.
[5.7] H. Masuda et al., IEEE J. Solid-State Circuits **SC-15**(5), 846 (1980).
[5.8] W.L. Martino et al., IEEE J. Solid-State Circuits **SC-15**(5), 820 (1980).
[5.9] T. Kuroda et al., "A 0.9 V 150 MHz 10 mW 4 mm^2 2-D discrete cosine transform core processor with variable-threshold-voltage scheme", ISSCC Dig. Tech. Papers, pp. 166–167, Feb. 1996.
[5.10] T. Furuyama et al., "A latch-up like new failure mechanism for high density CMOS dynamic RAMs", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 33–34, May 1989.
[5.11] M. Hasegawa et al., "A 256 Mb SDRAM with subthreshold leakage current suppression", ISSCC Dig. Tech. Papers, pp. 80–81, Feb. 1998.
[5.12] K. Shimotori et al., IECE **J64-C**(11), 769 (1981).
[5.13] M.I. Elmasry, *Digital MOS integrated circuits*, p. 23 (IEEE Press, New York 1981).
[5.14] Miyamoto et al., IECE **J75-C-II**(1), 38 (1992).
[5.15] D. Takacs et al., "Static and transient latch-up hardness in n-well CMOS with on-chip substrate bias generator", IEDM 85 Dig. Tech. Papers, pp. 504–508, 1985.
[5.16] Taniguchi et al., IECE **J65-C**(7), 530 (1982).
[5.17] E. Takeda et al., "Hot-carrier effects in submicron VLSIs", Symp. VLSI Technol., Dig. Tech. Papers, pp. 104–105, 1983.
[5.18] E. Takeda et al., IEEE Trans. Electron Devices **ED-29**(4), 611 (1982).
[5.19] T. Horiuchi et al., "Hot-carrier induced degradation of N-MOSFETS in inverter operation", Symp. VLSI Technol., Dig. Tech. Papers, pp. 104–105, 1985.
[5.20] S. Ogura et al., IEEE Trans. Electron Devices **ED-27**(8), 1359 (1980).
[5.21] T. Sakurai et al., IEEE J. Solid-State Circuits **SC-21**(1), 187 (1986).
[5.22] J. Harter et al., "A 60 ns hot electron resistant 4 M DRAM with trench cell", ISSCC88 Dig. Tech. Papers, pp. 244–245, Feb. 1988.
[5.23] K. Nogami et al., "VLSI circuit reliability under ac hot-carrier stress", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 13–14, May 1987.
[5.24] K. Furutani et al., IEICE Trans. **J73-C-II**(5), 302 (1990).

[5.25] T. Nakano, Y. Akasaka, ULSI DRAM technology (in Japanese)", Science Forum, Sept. 1992.
[5.26] S. Kohyama et al., "Non-thermal carrier generation in MOS structures", 11th Conf. Solid-State Devices - Tokyo, p. A-2-2, 1979.
[5.27] A. Mohsen et al., IEEE J. Solid-State Circuits **SC-19**(5), 610 (1984).
[5.28] C. Webb et al., "A 65 ns CMOS 1 Mb DRAM", ISSCC Dig. Tech. Papers, pp. 262–263, Feb. 1986.
[5.29] S. Fujii et al., IEEE J. Solid-State Circuits **24**(5), 1170 (1989).
[5.30] K. Sato et al., "A 20 ns static column 1 Mb DRAM in CMOS technology", ISSCC Dig. Tech. Papers, pp. 254–255, Feb. 1985.
[5.31] Y. Konishi et al., IEEE J. Solid-State Circuits **25**(5), 1112 (1990).
[5.32] Y. Tsukikawa et al., "An efficient back-bias generator with hybrid pumping circuit for 1.5 V DRAMs", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 85–86, 1993.
[5.33] T. Mano et al., "Submicron VLSI memory circuits", ISSCC83 Dig. Tech. Papers, pp. 234–235, Feb. 1983.
[5.34] K. Itoh et al., "An experimental 1 Mb DRAM with on-chip voltage limiter", ISSCC84 Dig. Tech. Papers, pp. 282–283, Feb. 1984.
[5.35] K. Itoh, IEEE J. Solid-State Circuits **25**(3), 778 (1990).
[5.36] K. Itoh, "Reviews and prospects of deep sub-micron DRAM technology", Int. Conf. Solid State Devices and Materials, Ext. Abstr., pp. 468–471, Aug. 1991.
[5.37] M. Aoki et al., "A 1.5 V DRAM for battery-based applications", ISSCC89 Dig. Tech. Papers, pp. 238–239, Feb. 1989.
[5.38] Y. Nakagome et al., IEEE J. Solid-State Circuits **26**(4), 465 (1991).
[5.39] K. Sato et al., IEEE J. Solid-State Circuits **26**(11), 1556 (1991).
[5.40] H. Hidaka et al., IEEE J. Solid-State Circuits **27**(7), 1020 (1992).
[5.41] R.S. Mao et al., "A new on-chip voltage regulator for high density CMOS DRAMs", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 108–109, 1992.
[5.42] Y. Nakagome et al., IEEE J. Solid-State Circuits, **26**(7), 1003 (1991).
[5.43] T. Furuyama et al., "An experimental 4 Mb CMOS DRAM", ISSCC86 Dig. Tech. Papers, pp. 272–273, Feb. 1986.
[5.44] T. Furuyama et al., IEEE J. Solid-State Circuits, **SC-22**(3), 437 (1987).
[5.45] D. Chin et al., IEEE J. Solid-State Circuits **24**(5), 1191 (1989).
[5.46] M. Horiguchi et al., IEEE J. Solid-State Circuits **23**(5), 1128 (1988).
[5.47] M. Horiguchi et al., IEEE J. Solid-State Circuits, **25**(5), 1129 (1990).
[5.48] M. Horiguchi et al., IEEE J. Solid-State Circuits, **26**(11), 1544 (1991).
[5.49] D.S. Min et al., IEEE J. Solid-State Circuits **27**(4), 626 (1992).
[5.50] M. Takada et al., "A 4 Mb DRAM with half internal voltage bit line precharge", ISSCC86 Dig. Tech. Papers, pp. 270–271, Feb. 1986.
[5.51] A. Tanbe et al., IEEE J. Solid-State Circuits **27**(11), 1525 (1992).
[5.52] D. Takashima et al., IEEE J. Solid-State Circuits **27**(4), 603 (1992).
[5.53] H. Tanaka et al., IEICE Trans. Electron. **E75-C**(11), 1333 (1992).
[5.54] P.R. Gray, R.G. Meyer, *Analysis and design of analog integrated circuits,* 2nd ed. (John Wiley, New York).
[5.55] H. Tanaka et al., IEICE Trans. **J75-C-2**(8), 425 (1992).
[5.56] Nakamura et al., "Study of the relation of internal voltage converter and ground noise", Proc. IEICE Spring Conf., Part 5, C-618, 1993.
[5.57] G. Kitsukawa et al., IEEE J. Solid-State Circuits **24**(3), 597 (1989).
[5.58] P.E. Allen, D.R. Holberg, *CMOS analog circuit design* (Holt, Rinehart and Winston, New York).
[5.59] S.M. Yoo et al., IEEE J. Solid-State Circuits **28**(4), 499 (1993).
[5.60] R.A. Blauschild et al., IEEE J. Solid-State Circuits **SC-13**(6), 767 (1978).

[5.61] H. Tanaka et al., "Sub-1-μA dynamic reference voltage generator for battery-operated DRAMs", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 87–88, May 1993.
[5.62] Tsuruda et al., "A tuning method of voltage down converter circuit", Proc. IEICE Spring Conf., Part 5, C-643, 1993.
[5.63] R.C. Foss et al., "Application of a high-voltage pumped supply for low-power DRAM", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 106–107, 1992.
[5.64] D. Lee et al., "A 35 ns 64 Mb DRAM using on-chip boosted power supply", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 64–65, 1992.
[5.65] H. Miyamoto et al., "A 32 ns 64 Mb DRAM with extended second metal line architecture", ESSCIRC93 Dig. Tech. Papers, pp. 41–44, Sept. 1993.
[5.66] S. Fujii et al., "A 50 μA standby 1 MW 1 b/256 KW 4 b CMOS DRAM", ISSCC86 Dig. Tech. Papers, pp. 266–267, Feb. 1986.
[5.67] A.L. Roberts et al., "A 256 K SRAM with on-chip power supply conversion", ISSCC Dig. Tech. Papers, pp. 252–253, Feb. 1987.
[5.68] K. Ishibashi et al., IEEE J. Solid-State Circuits **27**(6), 920 (1992).
[5.69] H.J. Shin et al., "Low-dropout on-chip voltage regulator for low-power circuits", IEEE Symp. Low Power Electronics, Dig., pp. 76–77, 1994.
[5.70] G.W. den Besten, B. Nauta, IEEE J. Solid-State Circuits **33**(7), 956 (1998).
[5.71] T. Ooishi et al., "A mixed-mode voltage-down converter with impedance adjustment circuitry for low-voltage wide-frequency DRAMs", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 111–112, June 1995.
[5.72] H. Neuteboom et al., IEEE J. Solid-State Circuits **32**(11), 1790 (1997).
[5.73] Y. Nakagome, "Voltage regulator design for low voltage DRAMs", Symp. VLSI Circuits, Memory Design Short Course, June 1998.
[5.74] P. Favrat et al., IEEE J. Solid-State Circuits **33**(3), 410 (1998).
[5.75] T. Hamamoto et al., "An efficient charge recycle and transfer pump circuit for low operating voltage DRAMs", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 110–111, June 1996.
[5.76] H. Tanaka et al., IEEE J. Solid-State Circuits **34**(8), 1084 (1999).
[5.77] H. Mizuno et al., "A 18 μA-standby-current 1.8 V 200 MHz microprocessor with self substrate-biased data-retention mode", ISSCC Dig. Tech. Papers, pp. 280–281, Feb. 1999.
[5.78] M. Miyazaki et al., "A 1000-MIPS/W microprocessor using speed-adaptive threshold-voltage CMOS with forward bias", ISSCC Dig. Tech. Papers, pp. 420–421, Feb. 2000.

Chapter 6

[6.1] Special Report: Memory, IEEE SPECTRUM, pp. 4–57, Oct. 1992.
[6.2] D.A. Patterson, J.L. Hennesy, *Computer organization and Design, The hardware/software interface* (Morgan Kaufmann).
[6.3] D.A. Patterson, J.L. Hennesy, *Computer architecture: a quantitative approach* 2nd ed. (Morgan Kaufmann).
[6.4] K. Itoh et al., "Limitations and challenges of multi-gigabit DRAM circuits", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 2–7, June 1996.
[6.5] S. Ohsima, T. Furuyama, IEICE Trans. **E77-C**(8), 1303 (1994).
[6.6] T. Kizaki et al., Hitachi Review **46**(1), 27 (1997).
[6.7] J. Torborg, J.T. Kajiya, "Talisman: commodity realtime 3D graphics for the PC", Computer Graphics Proc., Annual Conference Series, pp. 353–363, Aug. 1996.

[6.8] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).
[6.9] T. Shimizu et al., "A multimedia 32 b RISC microprocessor with 16 Mb DRAM", ISSCC Dig. Tech. Papers, pp. 216–217, Feb. 1996.
[6.10] S.S. Iyer, H.L. Kalter, "Embedded DRAM technology", IEEE Spectrum, pp. 56–64, Apr. 1999.
[6.11] K. Ishibashi, IEICE Trans. Electron. **E79-C**(6), 724 (1996).
[6.12] T. Furuyama et al., "A high random-access-data rate 4 Mb DRAM with pipeline operation", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 9–10, 1990.
[6.13] K. Itoh, in *Low power design methodologies*, J.M. Rabaey, M. Pedram, Editors (Kluwer, Norwell, MA 1995).
[6.14] K. Itoh et al., "VLSI memory technology: current status and future trends", ESSCIRC99 Dig. Tech. Papers, pp. 3–10, Sept. 1999.
[6.15] T. Kirihata et al., "390 mm^2 16 Bank 1 Gb DDR SDRAM with hybrid bitline architecture", ISSCC99 Dig. Tech. Papers, pp. 420–421, Feb. 1999.
[6.16] S. Takase, N. Kushiyama, "1.6 GB/s DRAM with flexible mapping redundancy technique and additional refresh Scheme", ISSCC99 Dig. Tech. Papers, pp. 410–411, Feb. 1999.
[6.17] K. Itoh et al., IEEE J. Solid-State Circuits **32**(5), 624 (1997).
[6.18] H.-J. Yoo et al., "A 150 MHz 8-banks 256 M synchronous DRAM with wave pipelining methods," ISSCC95 Dig. Tech. Papers, pp. 250–251, Feb. 1995.
[6.19] T. Saeki et al., "A 2.5 ns clock access 250 MHz 256 Mb SDRAM with a synchronous mirror delay", ISSCC96 Dig. Tech. Papers, pp. 374–375, Feb. 1996.
[6.20] H. Yoon et al., "A 2.5 V 333 Mb/s/pin 1 Gb double data rate SDRAM", ISSCC99 Dig. Tech. Papers, pp. 412–413, Feb. 1999.
[6.21] Y. Takai et al., "A 250 Mb/s/pin 1 Gb double data rate SDRAM with a bidirectional delay and an inter-bank shared redundancy scheme", ISSCC99 Dig. Tech. Papers, pp. 418–419, Feb. 1999.
[6.22] P. Gillingham, B. Vogley, "SLDRAM: high-performance, open-standard memory", IEEE Micro, pp. 29–39, Nov./Dec. 1997.
[6.23] Y. Nakase et al., IEEE J. Solid-State Circuits **34**(4), 494 (1999).
[6.24] L. Paris et al., "A 800 MB/s 72 Mb SLDRAM with digitally calibrated DLL", ISSCC99 Dig. Tech. Papers, pp. 416–417, Feb. 1999.
[6.25] R. Crisp, "Direct Rambus technology: the new main memory standard", IEEE Micro, pp. 18–28, Nov./Dec. 1997.
[6.26] Y. Konishi et al., IEICE Trans. Electron. **E82-C**(3), 438 (1999).
[6.27] HM5264165D-A60, in Hitachi catalog ADE-203-909A(Z) Rev. 1.0, Dec. 3, 1998.
[6.28] Nikkei Electronics(708), pp. 139–152, Jan. 26, 1998.
[6.29] Nikkei Microdevices, pp. 130–141, Feb. 1998.
[6.30] `http://www.rambus.com`.
[6.31] T. Kimura et al., "64 Mb 6.8 ns random ROW access DRAM macro for ASICs", ISSCC99 Dig. Tech. Papers, pp. 416–417, Feb. 1999.
[6.32] I. Naritake et al., "A 12 ns 8 MB DRAM secondary cache for a 64 b microprocessor", ISSCC99 Dig. Tech. Papers, pp. 420–421, Feb. 1999.
[6.33] O. Takahashi et al., "1 GHz fully pipelined 3.7 ns address access time 8 k × 1024 embedded DRAM macro", ISSCC Dig. Tech. Papers, pp. 396–397, Feb. 2000.

Chapter 7

[7.1] K. Itoh, K. Sasaki, Y. Nakagome, IEEE Proc. **83**(4), 524 (1995).

[7.2] K. Itoh, "Low power memory design", in Low power design methodologies, J.M. Rabaey, M. Pedram, Eds., Kluwer, Norwell, MA, pp. 201–251, Oct. 1995.

[7.3] K. Itoh et al., IEEE J. Solid-State Circuits **32**(5), 624 (1997).

[7.4] K. Itoh, IEEE J. Solid-State Circuits **25**, 778 (1990).

[7.5] Y. Nakagome et al., IEEE J. Solid-State Circuits **26**(3), 465 (1991).

[7.6] K. Satoh et al., "A 4 Mb pseudo-SRAM operating at (2.6±1) V with 3 μA data-retention current", ISSCC Dig. Tech. Papers, pp. 268–269, Feb.1991.

[7.7] K. Ishibashi, IEICE Trans. Electron. **E79-C**(6), 724 (1996).

[7.8] M. Matsumiya et al., "A 15 ns 16 Mb CMOS SRAM with reduced voltage amplitude data bus", ISSCC Dig. Tech. Papers, pp. 214–215, Feb. 1992.

[7.9] Hitachi Memory Catalogs, Aug. 1993 and Aug. 1994.

[7.10] D. Takashima et al., "Noise suppression scheme for giga-scale DRAM with hundreds of I/Os", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 196–197, June 1996.

[7.11] M. Taguchi, IEICE Trans. Electron. **E77-C**(12), 1944 (1994).

[7.12] H.W. Johnson, M. Graham, *High-speed digital design* (Prentice-Hall 1993).

[7.13] K. Kimura et al., IEEE J. Solid-State Circuits **SC-21**, 381 (1986).

[7.14] K. Itoh et al., "An experimental 1 Mb DRAM with on-chip voltage limiter", ISSCC Dig. Tech. Papers, pp. 106–107, Feb. 1984.

[7.15] T. Sugibayashi et al., "A 30 ns 256 Mb DRAM with multi-divided array structure", ISSCC Dig.Tech. Papers, pp. 50–51, Feb. 1993.

[7.16] H. Tanaka et al., IEICE Trans. Electron. **E75-C**(11), 1333 (1999).

[7.17] M. Horiguchi et al., IEEE J. Solid-State Circuits, **26**(11), 1544 (1991).

[7.18] M. Yoshimoto, et al., "A 64 Kb CMOS RAM with divided word line structure", ISSCC Dig. Tech. Papers, pp. 58–59, Feb. 1983.

[7.19] O. Minato, et al., "A 20 ns 64 K CMOS RAM", ISSCC Dig. Tech. Papers, pp. 222–223, Feb. 1984.

[7.20] K. Sasaki, et al., IEEE J. Solid-State Circuits **24**, 1219 (1989).

[7.21] E. Seevinck, "A current sense amplifier for fast CMOS SRAMs", Symp. VLSI Circuit, Dig. Tech. Papers, pp. 71–72, June 1990.

[7.22] K. Sasaki, et al., "A 7 Mb 140 mW CMOS SRAM with current sense amplifier", ISSCC Dig. Tech. Papers, pp. 208–209, Feb. 1992.

[7.23] K. Seno et. al., "A 9 ns 16 Mb CMOS SRAM with offset reduced current sense amplifier", ISSCC Dig. Tech. Papers, pp. 248–249, Feb. 1993.

[7.24] K. Ishibashi et al., "A 6 ns 4 Mb CMOS SRAM with offset-voltage-insensitive current sense amplifiers", Symp. VLSI Circuits, Dig. Tech. Papers, pp. 107–108, June 1994.

[7.25] N. Kushiyama et al., "A 295 MHz CMOS 1 M (×256) embedded SRAM using bidirectional read/write shared sense amps and self-timed pulsed word-line drivers", ISSCC Dig. Tech. Papers, pp. 304–305, Feb. 1995.

[7.26] K. Osada et al., "A 2 ns access, 285 MHz, two-port cache macro using double global bit-line pairs", ISSCC Dig. Tech. Papers, pp. 402–403, Feb. 1997.

Chapter 8

[8.1] K. Itoh, *VLSI Memory Design* (Baifukan, Tokyo 1994) (in Japanese).

[8.2] K. Itoh et al., Proc. IEEE **83**(4), 524 (1995).

[8.3] K. Itoh, "Low power memory design", in Low Power Design Methodologies, J.M. Rabaey, M. Pedram, Eds., Kluwer, Norwell, MA, pp. 201–251, Oct. 1995.

[8.4] K. Itoh et al., IEEE J. Solid-State Circuits **32**, 624 (1997).

[8.5] K. Itoh, "Ultralow-Voltage Memory Circuits", VLSI'97 Tutorial, Gramado (Brazil), Aug. 1997.

[8.6] K. Itoh et al., Electrochem. Soc. Proc. **98-1**, 350 (1998).

[8.7] V.P. Tsividis, *Operation and modeling of the MOS transistor* (McGraw-Hill, New York 1998).

[8.8] S.M. Sze, *Physics of semiconductor devices (2nd ed.)* (John Wiley, New York 1981).

[8.9] F.H. Gaensslen, R.C. Jaeger, "Low temperature microelectronics", Ext. Abstr., 22nd Conf. Solid State Devices and Materials, pp. 353–356, Aug. 1990.

[8.10] W.H. Henkels et al., IEEE J. Solid-State Circuits **26**(11), 1519 (1991).

[8.11] M. Aoki et al., IEEE Trans. Electron. Devices, **36**(8), 1429 (1989).

[8.12] J.-P. Colinge IEEE Electron Device Lett. **7**(4), 244 (1986).

[8.13] M. Aoki et al., "A 1.5 V DRAM for battery-based applications", ISSCC Dig. Tech. Papers, pp. 238–239, Feb.1989.

[8.14] Y. Nakagome et al., "A 1.5 V circuit technology for 64 Mb DRAMs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 17–18, June 1990.

[8.15] G. Kitsukawa et al., IEEE J. Solid-State Circuits **28**(11), 1105 (1993).

[8.16] M. Horiguchi et al., "Switched-source-impedance CMOS circuit for low standby current giga-scale LSIs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 47–48, May 1993.

[8.17] T. Sakata et al., "Subthreshold-current reduction circuits for multi-gigabit DRAMs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 83–84, May 1993.

[8.18] D. Takasima et al., "Stand-by/active mode logic for sub-1 V 1 G/4 Gb DRAMs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 83–84, May 1993.

[8.19] S. Mutoh et al., "1 V high-speed digital circuit technology with 0.5 μm multi-threshold CMOS", IEEE ASIC Conf. Dig. Tech. Papers, pp. 186–189, Sept. 1993.

[8.20] K. Seta et al., "50% active-power saving without speed degradation using stand-by power reduction (SPR) circuit", ISSCC Dig. Tech. Papers, pp. 318–319, Feb. 1995.

[8.21] D. Burnett et al., "Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits", Symp. VLSI Technol. Dig. Tech. Papers, pp. 15–16, 1994.

[8.22] N. Rohrer et al., "A 480 MHz RISC Microprocessor in a 0.12 μm L_{eff} CMOS technology with copper interconnects", ISSCC Dig. Tech. Papers, pp. 240–241, Feb. 1998.

[8.23] K. Itoh, IEEE J. Solid-State Circuits **25**(3), 778 (1990).

[8.24] Y. Nakagome et al., "Sub-1 V swing bus architecture for future low-power ULSIs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 82–83, June 1992.

[8.25] Yibi n Ye et al., "A new technique for standby leakage reduction in high-performance circuits", Symp. VLSI Circuits Dig. Tech. Papers, pp. 40–41, 1998.

[8.26] M. Hasegawa et al., "A 256 Mb SDRAM with subthreshold leakage current suppression", ISSCC Dig. Tech. Papers, pp. 80–81, Feb. 1998.

[8.27] T. Sakata et al., IEEE Solid-State Circuits **29**(8), 887 (1994).
[8.28] H. Akamatsu et al., "A low power data holding circuit with an intermittent power supply scheme for sub-1 V MT-CMOS LSIs", Symp. VLSI Crircuits Dig. Tech. Papers, pp. 14–15, June 1996.
[8.29] T. Kuroda et al., "A 0.9 V 150 MHz 10 mW 4 mm^2 2-D discrete cosine transform core processor with variable-threshold-voltage scheme", ISSCC Dig. Tech. Papers, pp. 166–167, Feb. 1996.
[8.30] M. Mizuno et al., "Elastic-V_t CMOS circuits for multiple on-chip power control", ISSCC Dig. Tech. Papers, pp. 300–301, Feb. 1996.
[8.31] K. Kumagai et al., "A novel power-down scheme for low V_T CMOS circuits", Symp. VLSI Circuits Dig. Tech. Papers, pp. 44–45, 1998.
[8.32] H. Makino et al., "An auto-backgate-controlled MT-CMOS circuit", Symp. VLSI Circuits Dig. Tech. Papers, pp. 42–43, 1998.
[8.33] T. Ooishi et al., "A well-synchronized sensing/equalizing method for sub-1.0 V operating advanced DRAMs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 81–82, May 1993.
[8.34] L.S.Y. Wong, G.A. Rigby, "A 1 V CMOS digital circuit with double-gate-driven MOSFET", ISSCC Dig. Tech. Papers, pp. 292–293, Feb. 1997.
[8.35] T. Kawahara et al., IEEE J. Solid-State Circuits **26**(11), 1530 (1991).
[8.36] K. Ishibashi et al., "A 1 V TET-load SRAM using a two-step word-voltage method", ISSCC Dig. Tech. Papers, pp. 206–207, Feb. 1992.
[8.37] K. Ishibashi et al., IEEE J. Solid-State Circuits **30**, 480 (1995).
[8.38] H. Morimura et al., IEEE J. Solid-State Circuits **33**(8), 1220 (1998).
[8.39] H. Mizuno et al., "Driving source-line (DSL) cell architecture for sub-1 V high-speed low-power applications", Symp. VLSI Circuits. Dig. Tech. Papers, pp. 25–26, June 1995.
[8.40] H. Yamaguchi et al., "A 0.8 V/100 MHz/Sub-5 mW-operated mega-bit SRAM cell architecture with charge-recycle offset-source driving (OSD) scheme", Symp. VLSI Circuits. Dig. Tech. Papers, pp. 126–127, June 1996.
[8.41] K. Itoh et al., "A deep sub-V, single power-supply SRAM cell with multi-V_T, boosted storage node and dynamic load", Symp. VLSI Circuits Dig. Tech. Papers, pp. 132–133, June 1996.
[8.42] H. Kawaguchi et al., "Dynamic leakage cut-off scheme for low-voltage SRAMs", Symp. VLSI Circuits Dig. Tech. Papers, pp. 140–141, June 1998.
[8.43] Y. Yamaguchi et al., IEICE Trans. Electron. **E79-C**, 772 (1996).
[8.44] C. Chuang et al., Proc.IEEE **86**(4), 689 (1998).
[8.45] Y. Yamaguchi et al., IEICE Trans. Electron. **E78-C**, 812 (1995).
[8.46] K. Ueda et al., "A CAD-compatible SOI/CMOS gate array having body-fixed partially-depleted transistors", ISSCC Dig. Tech. Papers, pp. 288–289, Feb. 1997.
[8.47] K. Shimomura et al., "A 1 V 46 ns 16 Mb SOI-DRAM with body control technique", ISSCC Dig. Tech. Papers, pp. 68–69, Feb. 1997.
[8.48] S. Kuge et al., "SOI-DRAM circuit technologies for low power high speed multi-giga scale memories", Symp. VLSI Circuits Dig. Tech. Papers, pp. 103–104, 1995.
[8.49] F. Assaderaghi et al., "A novel silicon-on-insulator (SOI) MOSFET for ultralow voltage operation", Symp. Low Power Electronics Dig. Tech. Papers, pp. 58–59, 1994.
[8.50] T. Fuse et al., "A 0.5 V 200 MHz 1-stage 32 b ALU using a body bias controlled SOI pass-gate logic", ISSCC Dig. Tech. Papers, pp. 286–287, Feb. 1997.
[8.51] T. Douseki et al., "A 0.5 V SIMOX-MTCMOS circuit with 200 ps logic gate", ISSCC Dig. Tech. Papers, pp. 84–85, Feb. 1996.

[8.52] F. Morishita et al., "Leakage mechanism due to floating body and countermeasure on dynamic retention mode of SOI-DRAM", Symp. VLSI Tech. Dig. Tech. papers, pp. 141–142, 1995.
[8.53] T. Tanigawa et al., IEICE Trans. Electron. **E79-C**, 781 (1996).
[8.54] S. Tomishima et al., "A long data retention SOI-DRAM with the body refresh function", Symp. VLSI Circuits Dig. Tech. Papers, pp. 198–199, June 1996.
[8.55] Y. Tosaka et al., IEICE Trans. Electron. **E79-C**, 767 (1996).
[8.56] B.S. Amrutur et al., "Techniques to reduce power in fast wide memories", Symp. Low Power Electronics Dig. Tech. Papers, pp. 92–93, Oct. 1994.
[8.57] B.S. Amrutur, M.A. Horowitz, "A replica technique for wordline and sense control in low-power SRAMs", IEEE J. Solid-State Circuits, **33**(8), 1208 (1998).
[8.58] T. Hamamoto et al., "Well conception: a novel scaling limitation factor derived from DRAM retention time and its modeling", IEDM Dig. Tech. Papers, pp. 915–918, Dec. 1995.
[8.59] Y. Nakagome et al., IEEE J. Solid-State Circuits **26**(7), 1003 (1991).
[8.60] R. Khanna et al., "A 0.25 µm X86 microprocessor with a 100 MHz Socket 7 Interface", ISSCC Dig. Tech. Papers, pp. 242–243, Feb. 1998.
[8.61] G. Singh, "A high speed 3.3 V IO Buffer with 1.9 V tolerant CMOS process", European Solid-State Circuits Conference Dig. Tech. Paper, pp. 128–131, Sept. 1998.
[8.62] H. Mizuno et al., "A 18 µA-stanby-current 1.8 V 200 MHz microprocessor with self substrate-biased data retention mode", ISSCC Dig. Tech. Papers, pp. 280–281, Feb. 1999.
[8.63] M. Miyazaki et al., "A 1000 MIPS/W microprocessor using speed-adaptive threshold-voltage CMOS with forward bias", ISSCC Dig. Tech. Papers, pp. 420–421, Feb. 2000.
[8.64] O. Takahashi et al., "1 GHz fully pipelined 3.7 ns address access time $k \times 1024$ embedded DRAM macro", ISSCC Dig. Tech. Papers, pp. 396–397, Feb. 2000.
[8.65] A.H. Montree et al., "Limitations to adaptive back bias approach for standby power reduction in deep sub-micron CMOS", ESSDERC Proc., pp. 580–583, Sept. 1990.
[8.66] K. Itoh et al., "VLSI memory technology: current status and future trends", ESSCIRC Dig. Tech. Papers, pp. 3–10, Sept. 1999.

Index

Access time, 6
Access-time penalty, 188
Activation energy, 206
Active chip power, 389
Active power reduction, 406, 413
Active power source, 402
Active restoring circuit, 162
Address buffer, 111, 141
Address comparators, 178
Address counter, 365
Address decoder, 144
Address multiplex, 20
Address Transition Detector (ATD), 112, 143, 170, 365, 413, 415
α-particle induced soft errors, 208
α-particle irradiation, 196
Alternate transposition, 239
Aluminum-strapped word-line, 141
Ambient temperature, 104
Amplifier activation speed, 220
Amplifier offset voltage, 223
AND cell, 45
ArF, 94
Aspect ratio, 200
Asynchronous operation, 358

Band-gap V_{REF} (BGR) generator, 316, 320
Bank, 357
Battery back-up mode, 105
Battery operation, 153, 290, 391
Battery-based portable system, 389
BGA (ball grid array), 364
Block, 341
Body effect, 53, 144
Boost ratio, 277
Boosted dc voltage, 249
Boosted sense-ground (BSG) scheme, 472
Boosted storage-node scheme, 465
Bootstrap capacitor, 255
Bootstrap inverter, 75
Built-in potential, 258, 469
Built-in self-test, 194
Burn-in test, 152, 250, 297, 323
Burn-in voltage generation, 324
Burst mode, 363
Bus utility, 348
Bypass capacitor, 269

Capacitive imbalance, 220
Capacitor-over-bit-line structure (COB cell), 200
Capacitor-under-bit-line structure (CUB cell), 200
$\overline{\text{CAS}}$ (the Column Address Strobe), 105, 111
$\overline{\text{CAS}}$ before $\overline{\text{RAS}}$ refresh (CBR), 105, 175
Cell-signal charge, 113
Channel length, 52
Channel length modulation, 55, 85
Channel-length modulation constant, 300
Channel width, 52
Charge recycle refresh, 412
Charge-transfer amplifier, 229
Charge-pumping circuit, 249
Chemical mechanical polishing, 200
Chip coating, 209
Chip shrink approach, 98
Chip-size package, 364
CMOS inverter, 79
CMOS latch-up, 65, 173, 258
CMOS NAND decoder, 145, 403
Collector–emitter current, 259
Column access time, 353
Column-address hold time, 361
Column-address set-up time, 361
Column cycle time, 353
Column latency, 363
Column mode, 398
Combined scaling, 92
Command instruction, 360

Command operation, 354, 363, 372
Common I/O, 168
Common-mode noise, 211
Common-mode voltage, 157
Comparator, 249
Concentrated spare lines, 189
Conductance imbalance, 223
Constant-current V_T, 426
Constant electric-field scaling, 90
Constant operation-voltage scaling, 92
Contact printing, 93
Cosmic-ray neutron-induced soft errors, 208
Counter, 89
Coupling noise, 250
Critical path, 126
Cross-coupled amplifier, 86
Cross-coupled differential amplifier, 157
Cross-shaped layout area, 129
CSP (Chip-Size Package), 396
CTT (Center Tapped Terminated), 401
Current mirror, 83
Current-mirror amplifier, 84, 170, 293, 417
Current sense amplifier, 170, 417, 419
Cutoff region, 51
Cycle time, 6

Data line, 8
Data-line and amplifier imbalance, 213, 217
Data-line arrangement, 157, 216
– Folded, 157
– Open, 157
Data-line dissipation charge, 404
Data-line interference noise, 213, 237
Data-line noise, 196, 210
Data-line pitch, 131
Data-output buffer, 172
Data-retention current, 390
Data-retention mode, 176
Data-retention power, 390, 405, 412
Data-retention time, 355
Decoded trimmer, 327
Decoupling scheme, 220
Deep trench capacitor, 203
Defect modes, 188
Defective memory elements, 178
Defective word lines, 178
Depletion NMOSFET (normally on), 50
Design parameter variation, 425, 433
Destructive readout, 16, 352, 403
Device technology, 100
Differential amplifier, 83
Differential signal current, 167
Diode decoupling, 221
Direct sensing, 165
Display device, 340
Distributed spare lines, 189
Divided bit-line NOR cell (DINOR cell), 44
Dominant-pole compensation, 315
Dopant atoms, 224
Double boosting, 279
Double-data-rate (DDR) DRAM, 358
Double-well CMOS, 260
Double-well structure, 451
Drain, 49
Drain–source voltage-reduction circuit, 271
DRAM-embedded system LSI, 203
DRAM main memory, 339
Dry etching, 100
Dual boosting, 151
Dual port, 340
Dual-power supply, 436
Dummy cell, 158, 214
Dummy word line, 214
Dynamic V_T, 466
Dynamic decoder, 144
Dynamic inverter, 76
Dynamic random access memory (DRAM), 1, 389, 403

EDO (Extended Data-Out) DRAM, 359, 365
EEPROM, 7, 33
Electrical programming, 179
Electromigration, 64
Electron–hole pair, 270
Embedded DRAM, 340
Embedded memory, 383
Enhancement NMOSFET (normally off), 50
Epitaxial layer, 252
EPROM, 7, 33
Error-checking and -correcting techniques (ECC), 182
Etching technology, 98
Exclusive OR, 415
Extrapolated V_T, 426
Extrinsic offset voltage, 210
Extrinsic variation, 223

Feedback charge-pump circuit, 283
Fine-pattern technology, 98

First-level cache, 341
Five-to-one projection aligner, 98
Fixed-V_T circuit, 438
Flash memory, 2, 7, 34
Floating body effect, 466
Floating-gate Avalanche-injection Metal Oxide Semiconductor (FAMOS), 33
Folded data-line arrangement, 20, 132, 223, 233
Forward-biased, 253
Four-transistor cell, 13
Fowler–Nordheim tunneling, 37
Full CMOS cell, 25

g line, 93
Gate, 49
Gate (G) boosting circuit, 438
Gate boosting, 87, 439
Gate–source back-biasing circuit, 438, 442
Gate–source offset driving, 442
Gate–source self-back-biasing, 442
Global metal wiring, 113
Graphics system, 340
Guard ring, 255

h line, 93
Half-V_{DD} (or V_{DL}) cell-capacitor plate, 249
Half-V_{DD} data-line precharge, 162, 406, 409
Half-V_{DD} electrode, 246
Half-V_{DD} generator, 332
Half-V_{DD} precharge, 233
Hemispherical-grain (HSG) poly-silicon electrode, 200
Hi-C capacitor structure, 267
Hierarchical I/O configuration, 169
Hierarchical memory system, 339
Hierarchical row decoder structure, 413
Hierarchical word-line structure, 141
High boost-ratio converter, 283
High-density packaging, 364
High-density technology, 98
High-performance circuits, 100
High-permittivity and thin dielectric materials, 248
High S/N ratio circuits, 116
High-speed clocking scheme, 354, 363
High-speed column modes, 108
Hit rate, 341
Hot-carrier breakdown voltage, 56
Hot electron, 270
Hot-electron injection, 36
HSTL (High-Speed Transceiver Logic), 363
Hybrid arrangement, 136
Hybrid converter, 295

IC, 3
Impact ionization, 270
Incomplete equalization of a pair of data lines, 247
Individual replacement, 185
Input/output interface circuit, 3
Input-predetermined logics, 431
Inter-subarray replacement redundancy, 189
Interleaving, 347
Internal chip configuration, 113
Internal reference voltage, 249
Intra-subarray replacement redundancy, 185
Intrinsic offset voltage, 210
Intrinsic random V_T variation, standard deviation, 224
Intrinsic variation, 223
Intrinsic V_T variation, 434
Inversion layer, 255
Iterative circuit block, 429

Junction temperature, 104

KrF, 94

Laser programming, 179
Laser beam, 179
Latch function, 111, 354
Latch-type CMOS sense amplifier, 404, 417
Latch-up immune structure, 252
Latch-up susceptibility, 262
LDD (Lightly Doped Drain–Source), 56, 100, 270
Leakage charge, 196, 204
Leakage currents, 195
Level monitor, 282
Level shifter, 88
Line delay
– Reduction, 124
Line-mode testing, 193
Lithography, 93
LOC (Lead-On-Chip), 364, 396
Local wiring, 114
Logical array, 119, 355
Loop stability, 250, 410

Loss of stored data, 233
Low dropout voltage, 291
Low-level data-line disturbances, 204
Low-power circuits, 117
Low-power DRAM circuit, 406
Low-power RAM circuit, 389
Low-power SRAM circuit, 413
Low-voltage data-bus interface, 396
LSI, 3
LVTTL (Low-Voltage TTL), 363, 400

Main word line, 129, 155
Maintenance of Q_S, 432
Maximum refresh time, 196, 408
MCM (multichip module), 364
Memory capacity quadrupling approach, 98
Memory-cell array, 3
Memory-cell miaturization, 195
Memory-cell signal charge, 425
Memory hierarchy, 341
Memory–processor performance gap, 343
Memory subsystem, 393
Miller compensation, 311
Minority carrier, 252
Minority-carrier reduction, 254
Miss penalty, 341
Mode register set, 372
MOS capacitor, 198
MOS-gate capacitors, 57
MOS threshold voltage, 50
MOSFET-load cell, 24
Multibank interleaving, 354, 355
Multibit testing, 193
Multidivided data-line, 20, 132, 137, 354
– Partial activation, 119, 406, 408
– Scheme, 134
Multidivided memory array, 128, 354
Multidivided power-line
– Partial activation, 456
Multidivided word-line, 30, 139, 354
– Partial activation, 407, 413
Multilayered dielectric film, 200
Multilevel cell, 45
Multipurpose register, 194
Multistage small-signal transmission, 172
Multi-V_T circuit, 438, 440

N-well, 261
NAND cell, 42
NAND decoder, 144
Narrow-channel effect, 55, 253
Negative reference voltage, 335
Negative word-line, 335, 465
Negative-feedback converter, 293
Nibble mode, 108, 347, 399
NMOS NOR decoders, 145
NMOS-output converter, 291
NMOS Static Circuit, 67
Noise cancellation, 211
Noise-generation mechanism, 217
Noise source, 210
Non-destructive readout, 11, 28, 354, 404
Non-saturated region, 52
Non-selected word-line noise, 213
NOR cell, 38
NOR decoder, 144
Normal mode, 398
Numerical aperture number, 94

Off-chip L2 cache, 342
Offset structure, 56
On-chip L1 cache, 342
On-chip mode register, 354, 363
On-chip monitor, 177
On-chip spare elements, 178
On-chip testing circuits, 192
On-chip V_{BB} generator, 456
On-chip voltage down-converter (VDC), 20, 389, 406, 409
On-chip voltage generator, 250, 435
One-to-one projection printing, 93
One-transistor, one-capacitor memory cell, 2, 15
Open data-line arrangement, 132, 231
Over-voltage protection, 282

p^+ barrier, 209
p–n-junction capacitors, 57
p–n junction leakage current, 204
Packaging, 94
Packet protocol, 354, 363
Page, 348
Page mode, 108, 399
Paired MOSFET, 432
Parallel testing, 193
Parasitic npn bipolar transistor, 259
Peripheral circuit, 3
Permittivity of the capacitor insulator, 197
Phase compensation, 293, 301, 410
Photoaligner, 98
Physical memory array, 119, 355

Pin function, 370
Pipeline operation, 359
Pipeline/prefetch operation, 354, 362
Planar-type cells, 199
PMOS-output converter, 291
PMOS output transistor, 151
Pole-zero compensation, 306, 315
Poly-Si fuses, 179
Poly-silicon load cell, 25
Positive reference voltage, 335
Power-on characteristics, 250, 258
Power-on reset-signal generator, 263
Power-supply rejection ratio, 334
Power-supply standardization, 290, 425, 435
Power-supply voltage bounce, 213, 240
Precharging methods, 216
Predecoding scheme, 146
Programming elements, 227
PROM, 7
Pull-down circuit, 213
Purification of materials, 209
Push–pull inverter, 75

Quadrupling of memory capacity, 97
Quiet array, 247, 248

Raised dc supply-voltage driver, 154
Rambus, 363
Rambus channel, 368, 380
Rambus DRAM, 339, 358, 380
Random Access Memory (RAM), 6
Random logic circuit, 429
Random microscopic fluctuation, 224, 432
$\overline{\text{RAS}}$ (Row Address Strobe), 105, 111
$\overline{\text{RAS}}$-clock buffer, 142
$\overline{\text{RAS}}$-only refresh, 105, 175
Ratio current, 404
Read Only Memory (ROM), 2, 6
Read operation, 105
Reduction of dc current, 411
Reduction of word-line delay, 155
Reduction projection printing, 93
Redundancy, 102, 178, 225
Reference voltage, 157
Reference-voltage generation, 158, 316
Reference-voltage stability, 250
Reference-voltage variation, 213
Refresh-busy rate, 119, 207, 408
Refresh charge reduction, 412
Refresh operation, 8, 11, 105, 196
Refresh-relevant circuits, 175
Refresh time extension, 412
Refresh time increase, 408
Refresh timer, 177
Reliability of wiring, 64
Resistivity, 60
Resistor, 60
Ring oscillator, 88, 256
Ripple, 334
Row access time, 352
Row address strobe ($\overline{\text{RAS}}$), 398
Row cycle time, 352
Row-address hold time, 360
Row-address set-up time, 360
Rush current, 258

Saturated E-NMOS load inverter, 68
Saturated region, 52
Scaling law, 90
Second-level cache, 342
Self-aligned contact, 203
Self-refresh, 175
Self-refreshing, 105
Sense amplifier, 15
– Distributed driving, 127
Sense-current distributed array, 164
Sensing circuit, 157
Separated I/O, 168
Series-pass regulator, 334
Set-up/hold timing, 359
Shared amplifier, 20
Shared decoders, 129
Shared I/O, 406
Sheet resistance, 60
Shielded data-line structure, 239
Short-circuit, 191
Short-channel effect, 55, 56
Signal charge, 196
Signal-to-noise ratio (S/N), 18, 104, 195
Silicon cycle, 2
Silicon substrate, 49
SIMOX (Separation by Implanted OXygen), 467
Simultaneous replacement, 185
Single port, 340
Single-power supply, 435
SiO_2 equivalent insulator thickness, 197
Small outline J-leaded package, 103
Small package, 394
Small-voltage signal transmission, 127
Soft error, 20
Soft-error critical charge, 196, 208
Soft-error immunity, 472
SOI CMOS technology, 425

Source, 49
Source control circuit, 461
Source–drain resistance, 55
Source–follower mode, 53
Sources of power dissipation, 402
Spare decoder, 179
SRAM cache, 339
SRAM cell, 24
SSTL (Stub Series Terminated Logic), 363, 367
Stacked-capacitor cell, 199
Stand-by current faults, 183
Stand-alone commodity SRAM, 342
Static column mode, 108, 347, 399
Static current, 391, 407
Static decoder, 144
Static random access memory (SRAM), 2, 391, 404
Stress migration, 64
Stress-released drain structure, 270
Subarray replacement redundancy technique, 192
Substrate-bias effect coefficient, 53, 257, 266, 277
Substrate bias voltage, 249
Substrate-bias voltage (V_{BB}) generator, 251
Substrate bounce, 251
Substrate-current generation, 250
Substrate doping concentration, 251, 457
Substrate (or well) structure, 250
Subthreshold current, 52, 204, 233, 425, 463
Subthreshold current reduction, 407
Subthreshold swing, 426
Subword-line, 129, 155
Switched substrate-impedance, 459
Synchronized decoupling, 221
Synchronous DRAM (SDRAM), 339, 358, 368
Synchronous operation, 354, 358
Synchronous-link DRAM, 365
System clock, 359

Terminated I/O interface, 354, 363
TFT load cell, 25
Thermal resistance of the package, 104
Thin Small Outline Package, 103
Three-transistor cells (3-T), 13
Threshold voltage, 53
Threshold voltage (V_T) scaling, 425
Throughput, 339
Total data-line charging capacitance, 139
Total data-line dissipating charge, 139, 408
Transconductance, 85
Transfer function, 302
Transposed data lines, 20
Transposition (twist) of data-line pair, 239
Trench-capacitor cells, 199
Triple-well structure, 272, 438, 453, 468
TTL (Transistor Transistor Logic), 104, 363, 399
Two-capacitor cell, 15
Two-step amplification, 229
Two-step driving, 220
Two-dimensional (2-D) selection, 456

ULSI, 3
Ultra-low-voltage DRAM circuit, 437
Ultra-low-voltage operations, 425
Ultra-low-voltage SOI circuit, 466
Ultra-low-voltage SRAM circuit, 463
Undershoot, 254
Usage efficiency of spare lines, 185

Variable-V_T circuit, 438
Variation in the reference voltage, 241
Varied boost-ratio driver, 152
V_{BB} bump, 266
V_{BB} clamping, 263
V_{BB} generator, 176
V_{DD} bump, 198
V_{DD} data-line precharge, 158
V_{DD} precharge, 231
Vertical capacitor, 59
Video processor, 340
Video RAM (VRAM), 340
VLSI, 3
Voltage booster, 87
Voltage bump, 244
Voltage clamper, 286
Voltage-division converter, 295
Voltage doubler, 281
Voltage down-converter, 250, 290
Voltage-stress relaxed word driver, 151
Voltage trimming, 250, 327
Voltage up-converter, 208, 250, 276, 464
V_T difference (ΔV_T) V_{REF} generator, 316, 318
V_T mismatch, 432, 463
V_T referenced V_{REF} generator, 316

Well and source control circuit, 456, 462
Wet etching, 100
Wide-bit I/O chip configuration, 354, 393
Wiring parasitic capacitance, 62
Wiring parasitic resistance, 62
Word driver, 147
Word line, 8
Word-line to data-line coupling capacitance, 230
Word-line drive noise, 212
Word-line pitch, 131
Word pulse deformation, 235
Word voltage bootstrapping, 20
Write and relevant circuits, 174
Write operation, 105