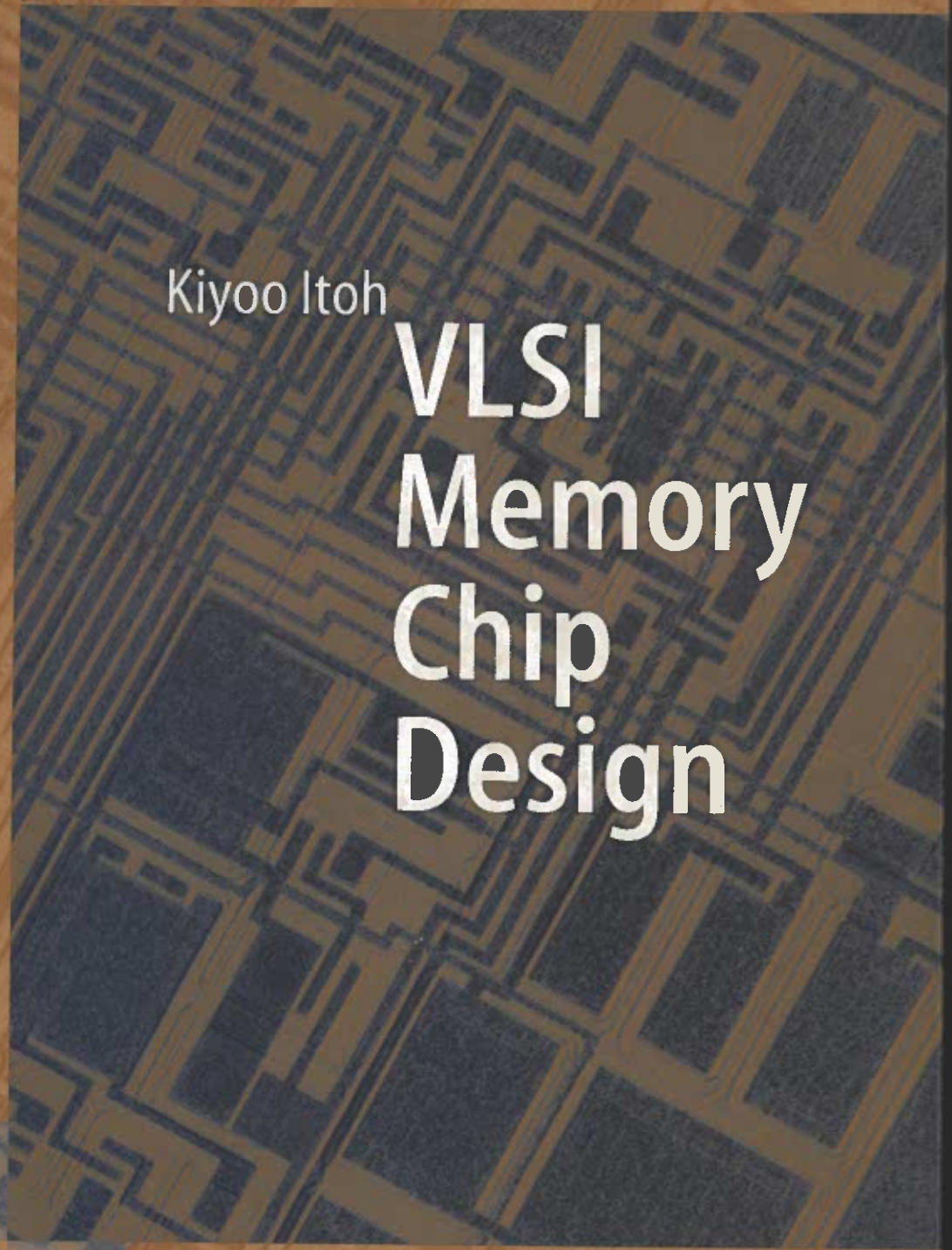
ADVANCED MICROELECTRONICS



Kiyoo Itoh

VLSI Memory Chip Design

Springer

Springer Series in

ADVANCED MICROELECTRONICS 5

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

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.
ISBN 3540678204 (alk. paper)
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

ISSN 1437-0387
ISBN 3-540-67820-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

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: 10771899 57/3144/mf - 5 4 3 2 1 0

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

SRAM Speed. Due to the possession of a gain, the SRAM cell allows the 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 multistage 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,

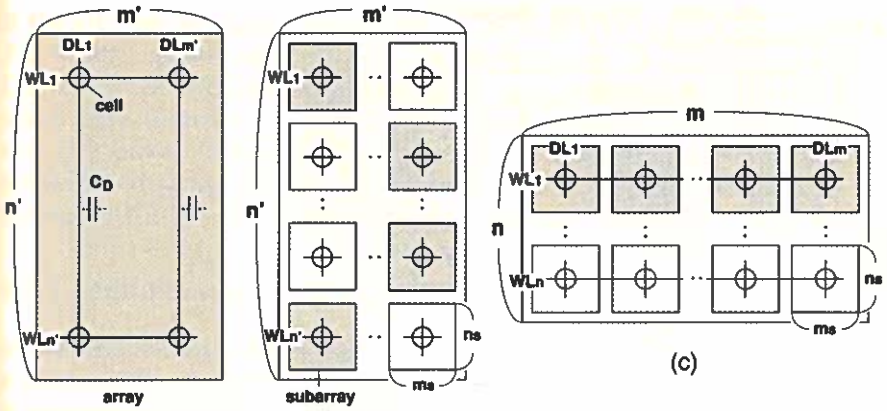


Fig. 6.11. The concept of a multidivided RAM array [6.14, 6.17]: the shading denotes the activated area. $M = n'm' = nm$. (**a**) A non-divided array; (**b**) a multidivided array; (**c**) a logical array

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 1 k 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

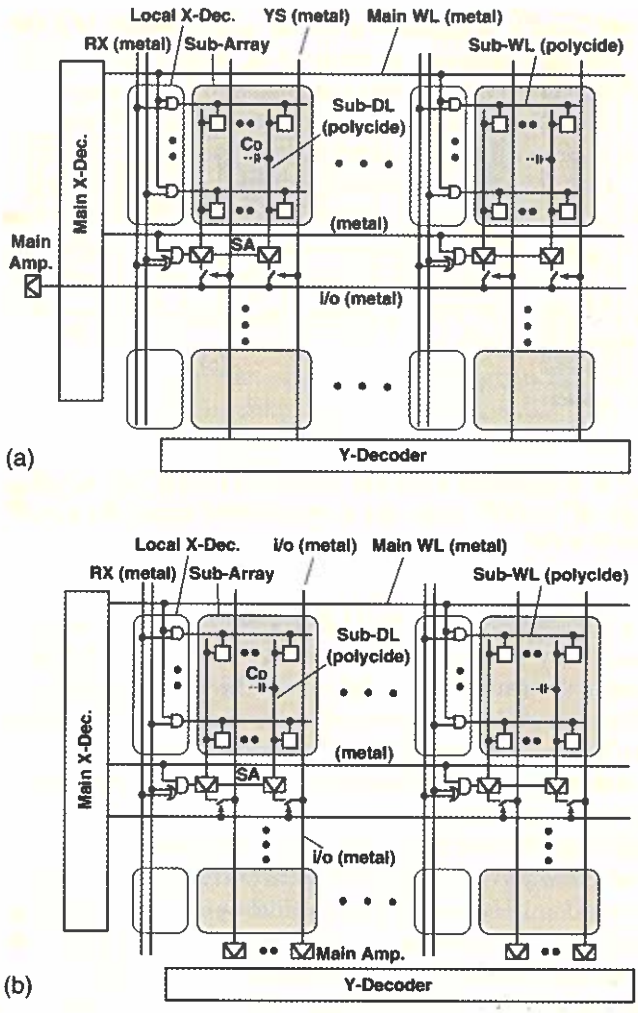


Fig. 6.12. An actual multidivided DRAM array [6.4, 6.17]: the paired-line arrangement is actually applied for each DL and each i/o line. (**a**) DL orthogonal to i/o line; (**b**) DL parallel to i/o line

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

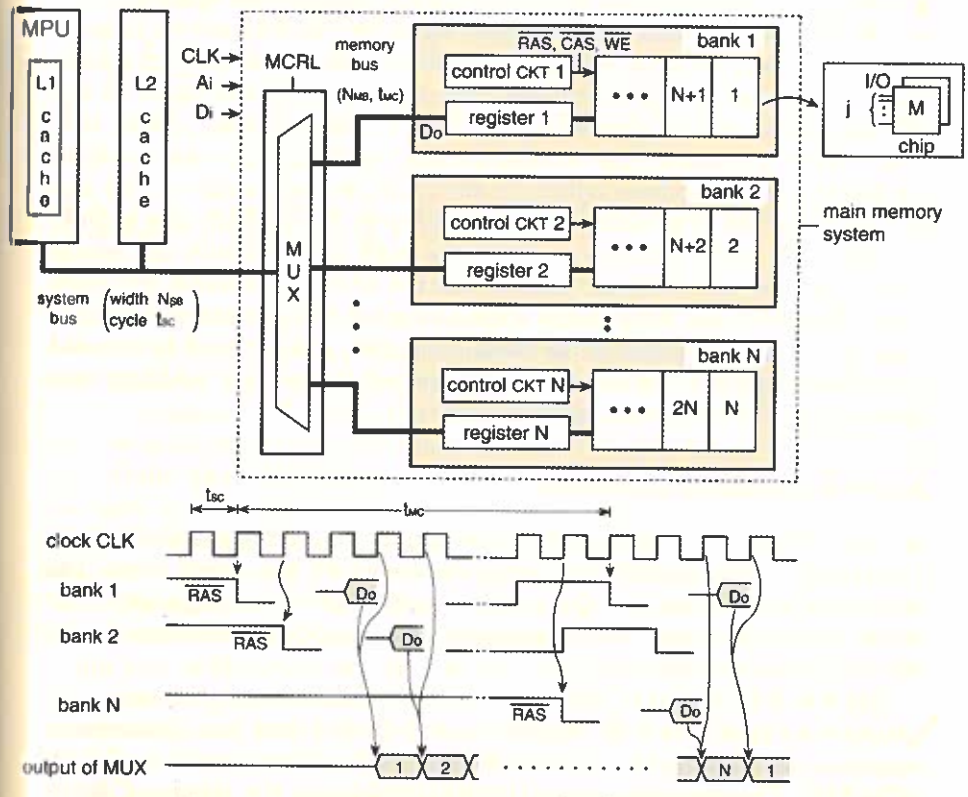


Fig. 6.13. Conventional multibank interleaving [6.8]. MUX, multiplexer; MCRL, memory controller

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

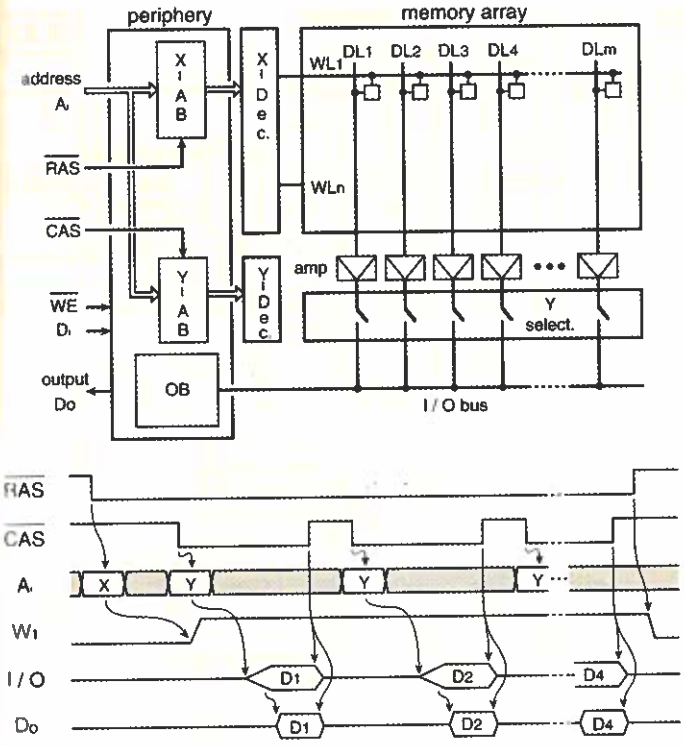


Fig. 6.14. The asynchronous operation of a DRAM chip [6.8]. AB, address buffers; OB, data-output buffer; X, row; Y, column

(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

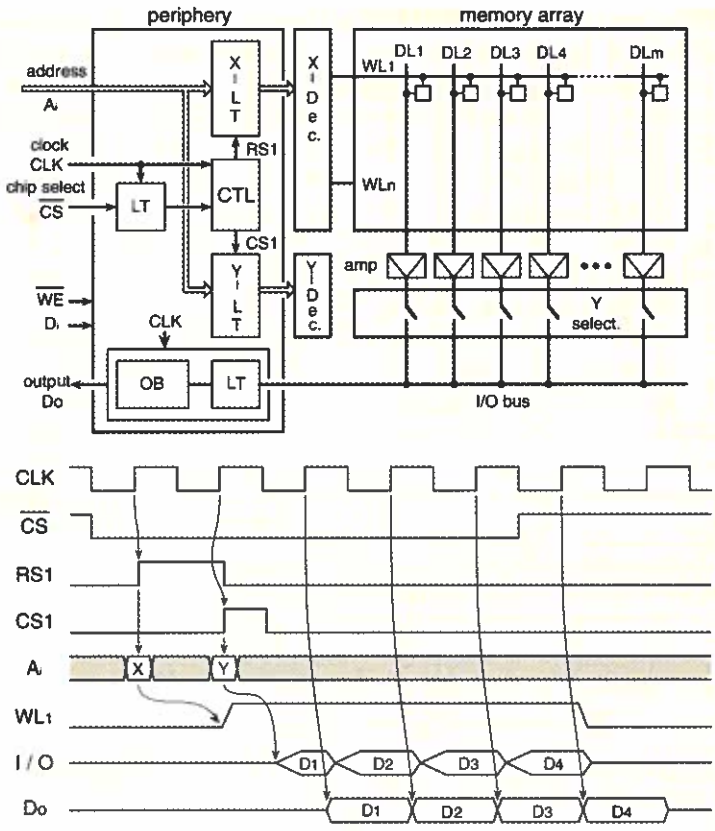


Fig. 6.15. The synchronous operation of a DRAM chip [6.8]. LT, latch; CTL, control circuit; OB, data-output buffer

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-

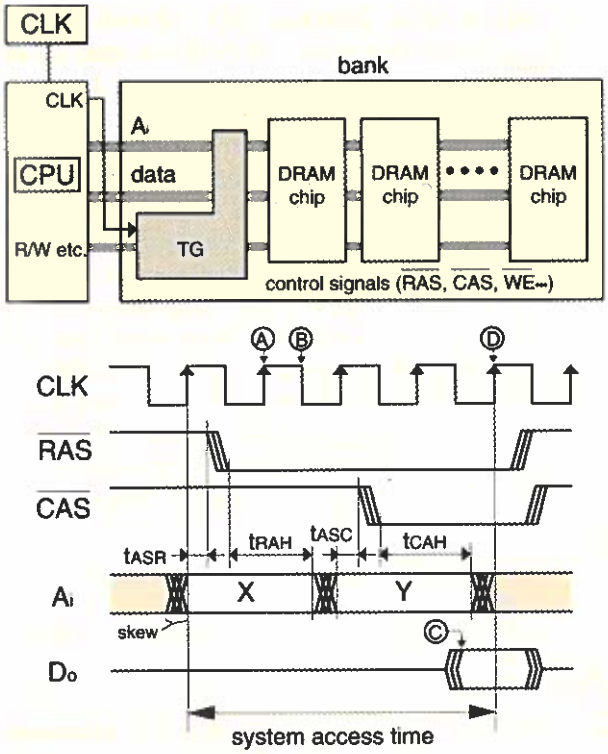


Fig. 6.16. The asynchronous operation of the memory system [6.8]. TG, timing generator; X, row; Y, column

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

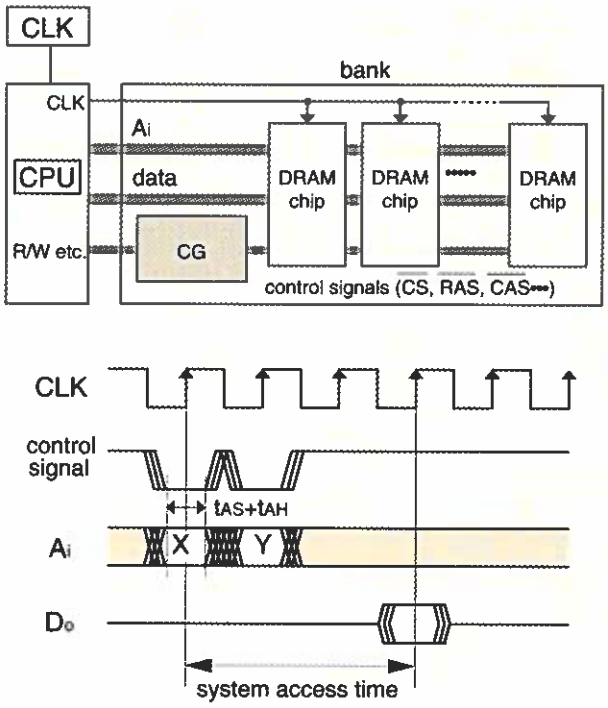


Fig. 6.17. The synchronous operation of the memory system [6.8]. CG, command generator

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.

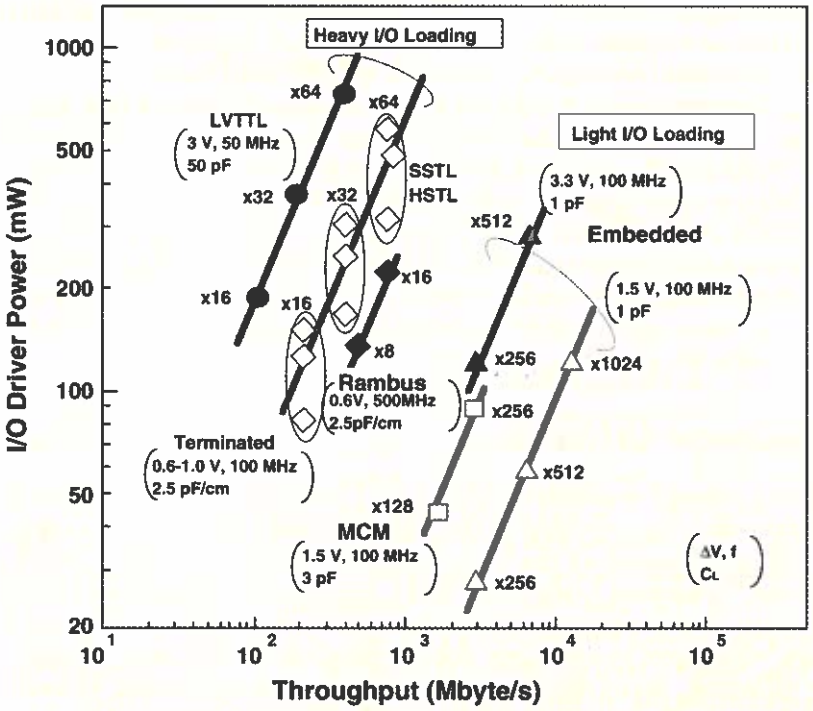


Fig. 6.18. The data transfer rate of a DRAM chip [6.4, 6.17]. x_j, The I/O pin count of a DRAM chip; C_L, the loading capacitance of each I/O line, without termination power