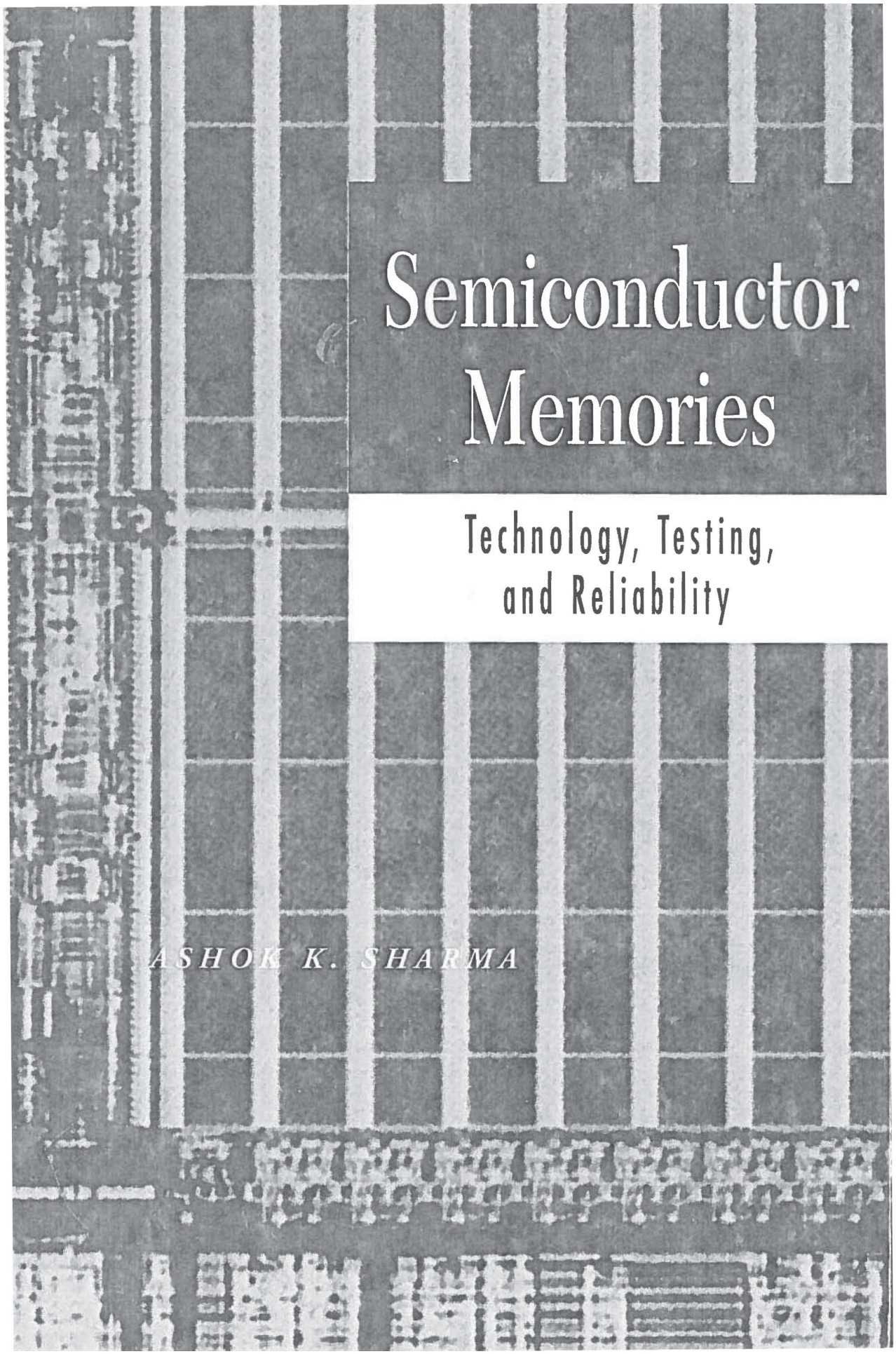


Semiconductor Memories

Technology, Testing, and Reliability

Ashok K. Sharma

IEEE Solid-State Circuits Council, *Sponsor*

The Institute of Electrical and Electronics Engineers, Inc., New York

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Printed in the United States of America
10 9 8 7 6 5 4

Library of Congress Cataloging-in-Publication Data

Sharma, Ashok K.
Semiconductor memories : technology, testing, and reliability /
Ashok K. Sharma.
p. cm.
Includes index.
ISBN 0-7803-1000-4
1. Semiconductor storage devices. I. Title
TK7895.M4S49 1996
621.39'732—dc20 96-6824
CIP

2

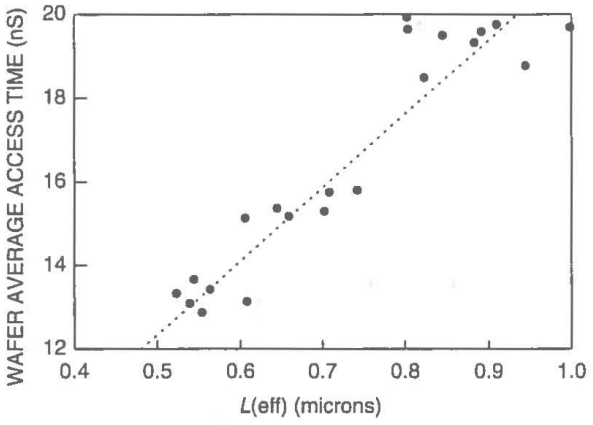
Random Access Memory Technologies

2.1 INTRODUCTION

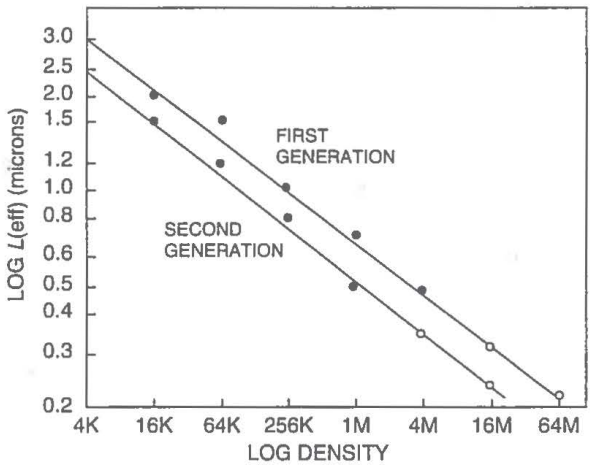
Semiconductor memory devices are generally categorized as volatile or nonvolatile random access memories (RAMs). In RAMs, the information is stored either by setting the state of a bistable flip-flop circuit or through the charging of a capacitor. In either of these methods, the information stored is destroyed if the power is interrupted. Such memories are therefore referred to as volatile memories. If the data are stored (i.e., written into the memory) by setting the state of a flip-flop, they will be retained as long as the power is applied and no other write signals are received. The RAMs fabricated with such cells are known as static RAMs, or SRAMs. When a capacitor is used to store data in a semiconductor RAM, the charge needs to be periodically refreshed to prevent it from being drained through the leakage currents. Hence, the volatile memories based on this capacitor-based storage mechanism are known as the dynamic RAMs, or DRAMs.

SRAM densities have generally lagged behind those for the DRAMs (1:4), mainly because of the greater number of transistors in a static RAM cell. For example, a 256 kb SRAM has about the same number of transistors as a 1 Mb DRAM. However, static RAMs are being widely used in systems today because of their low-power dissipation and fast data access time. Early static RAMs were developed in three separate technologies: bipolar, NMOS, and CMOS. By the middle of the 1980s, the vast majority of SRAMs were made in CMOS technology. SRAMs are currently available in many varieties, ranging from the conventional MOS (NMOS or CMOS) to high-speed bipolar and GaAs SRAM designs. Full bipolar SRAMs, although less than 1% of the total SRAM market, are still available in lower densities for very high-speed applications. The lower end commodity SRAMs are represented by the "mixed-MOS" technology which is a combination of CMOS and NMOS for high-density applications, and by the full CMOS technology for a combination of high-density and low-power requirements. High-speed and high-density SRAMs are fabricated in both CMOS and mixed-MOS technologies, as well as combinations of bipolars and CMOS, called BiCMOS.

SRAM speed has been usually enhanced by scaling of the MOS process since shorter gate channel length, L(eff) translates quite linearly into faster access time. Figure 2-1(a) shows the plot of SRAM wafer average access time versus L(eff). This scaling of the process from first-generation to second-generation SRAMs has resulted in support of higher density products, as shown in the L(eff) versus density logarithmic plot of Figure 2-1(b). Section 2.2 of this chapter discusses in detail various



(a)



(b)

Figure 2-1. (a) SRAM wafer average access time versus L(eff). (b) L(eff) versus SRAM density logarithmic plot. (From [1], with permission of IEEE.)

SRAM (NMOS and CMOS) cell structures; MOS SRAM architectures, cell, and peripheral circuits operation; bipolar SRAM (ECL and BiCMOS) technologies; silicon-on-insulator (SOI) technologies; advanced SRAM architectures and technologies; and some application-specific SRAMs.

In the last decade of semiconductor memories growth, dynamic random access memories (DRAMs) have been produced in the largest quantities because of their high density and low cost per bit advantage. Main computer memories are mostly implemented using the DRAMs, even though they lag behind the SRAMs in speed. Low-power dissipation and faster access times have been obtained, even with an increase in chip size for every successive generation of DRAMs, as shown in Figure 2-2 [2]. The DRAM memory array and peripheral circuitry such as decoders, selectors, sense amplifiers, and output drivers are fabricated using combinations of n-channel and p-channel MOS transistors. As a result, 1 Mb DRAMs have reached maturity in production, and nowadays, 4 and 16 Mb DRAMs are being offered by several manufacturers. While the density of DRAMs has been approximately quadrupling every three years, neither their access time nor cycle time has improved as rapidly. To increase the DRAM throughput, special techniques such as page mode, static column mode, or nibble mode have been developed. The faster DRAMs are fabricated with specially optimized processes and innovative circuit design techniques. In some new DRAMs being offered with wide on-chip buses, 1024–4096 memory locations can be accessed in parallel, e.g., cache DRAMs, enhanced DRAMs, synchronous DRAMs, and Rambus DRAMs. Section 2.3 of this chapter discusses the DRAM technology development; CMOS DRAMs; DRAM cell theory and advanced cell structures; BiCMOS DRAMs; soft-error failures

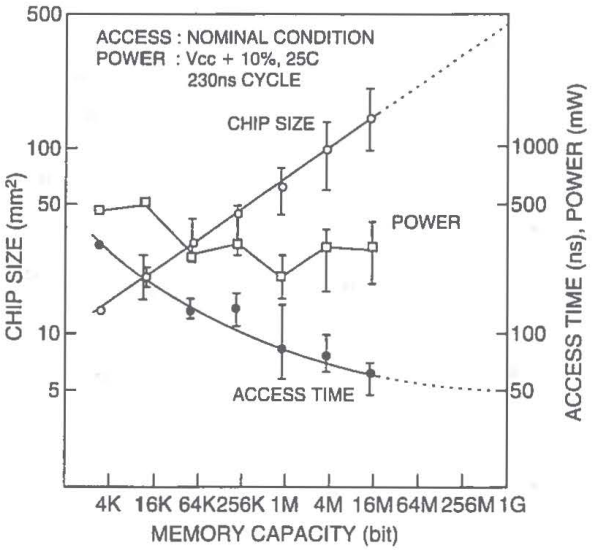


Figure 2-2. Trends in standard DRAM development. (From [2], with permission of IEEE.)

in DRAMs; advanced DRAM designs and architectures; and a few application-specific DRAMs.

2.2 STATIC RANDOM ACCESS MEMORIES (SRAMs)

SRAM is classified as a volatile memory because it depends upon the application of continuous power to maintain the stored data. If the power is interrupted, the memory contents are destroyed unless a back-up battery storage system is maintained. SRAM output width ranges from 1 to 32 b wide. Standard inputs and outputs include interfacing with CMOS, TTL, and ECL circuits. Power supply range includes standard 5 V and new 3.3 V standard for battery-powered applications. A SRAM is a matrix of static, volatile memory cells, and address decoding functions integrated on-chip to allow access to each cell for read/write functions. The semiconductor memory cells use active element feedback in the form of cross-coupled inverters to store a bit of information as a logic "one" or a "zero" state. The active elements in a memory cell need a constant source of dc (or static) power to remain latched in the desired state. The memory cells are arranged in parallel so that all the data can be received or retrieved simultaneously. An address multiplexing scheme is used to reduce the number of input and output pins. As SRAMs have evolved, they have undergone a dramatic increase in their density. Most of this has been due to the scaling down to smaller geometries. For example, the 4 kb MOS SRAM used 5 μm minimum feature size, while the 16 kb, 64 kb, 256 kb, and 1 Mb SRAMs were built with 3.0, 2.0, 1.2, and 0.8 μm feature size, respectively. Now, the 4 and 16 Mb SRAMs have been developed with 0.6–0.4 μm process technology. Also, scaling of the feature size reduces the chip area, allowing higher density SRAMs to be made more cost effectively.

2.2.1 SRAM (NMOS and CMOS) Cell Structures

The basic SRAM cell made up of cross-coupled inverters has several variations. The early NMOS static RAM cells consisted of six-transistor designs with four enhancement mode and two depletion mode pull-up transistors. A significant improvement in cost and power dissipation was achieved by substituting ion-implanted polysilicon load resistors for the two pull-up transistors. These were called *R*-load NMOS, and a successor to these is called the "mixed-MOS" or "resistor-load CMOS." These SRAMs consisted of an NMOS transistor matrix with high ohmic resistor loads and CMOS peripheral circuits which allowed the benefit of lower standby power consumption while retaining the smaller chip area of NMOS SRAMs. The transition from the *R*-load NMOS to the *R*-load CMOS SRAMs occurred at the 4 kb density level. At the same time, the concept of power-down mode controlled by the chip enable pin ($\overline{CE}$) also appeared. In this low-voltage standby mode, the standby current for the mixed-MOS parts is typically in the microamp (μA) range, and for the full CMOS, in the nanoamp (nA) range. This low-power dissipation in the standby mode opened the potential for high-density battery back-up applications.

The mixed-MOS and the full NMOS designs have higher standby currents than the CMOS SRAMs. However, the mixed-MOS technique provides better scaling advantages and relatively lower power dissipation. The early CMOS RAMs with metal gate technology were mostly used in the aerospace and other high-reliability applications for their wider noise margins, wider supply voltage tolerance, higher operating temperature range, and lower power consumption. MOS SRAMs became more popular with the development of silicon gate technology in which the two levels of interconnections through polysilicon and aluminum allowed reduction in cell area, enabling the fabrication of larger and denser memory arrays. In the mid-1980s, a vast majority of all SRAMs were being made in several variations, such as the CMOS or mixed-MOS formation in n-well, p-well, or the twin-tub technologies.

Figure 2-3 shows an MOS SRAM cell with load devices which may either be the enhancement or depletion mode transistors as in an NMOS cell, PMOS transistors in a CMOS cell, or load resistors in a mixed-MOS or an *R*-load cell. The access and storage transistors are

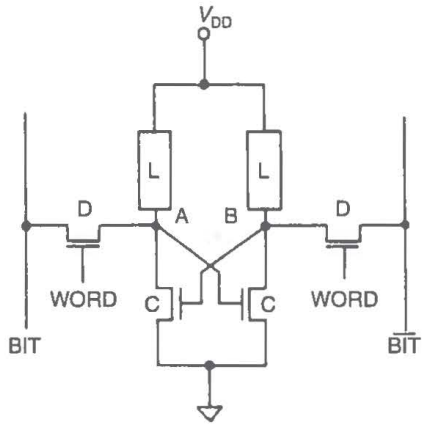


Figure 2-3. A general schematic of a SRAM memory cell.

enhancement mode NMOS. The purpose of load devices (*L*) is to offset the charge leakage at the drains of the storage and select transistors. When the load transistor is PMOS, the resulting CMOS cell has essentially no current flow through the cell, except during switching. The depletion load and resistive load have a low level of current flowing through them, and hence the standby power dissipation is always higher than that of the CMOS cell.

Figure 2-4(a) shows the basic CMOS SRAM cell consisting of two transistors and two load elements in a cross-coupled inverter configuration, with two select transistors added to make up a six-transistor cell. Figure 2-4(b) shows the use of polysilicon load resistors instead of PMOS transistors in the CMOS cell, which allows up to a 30% reduction in the cell size in double-polysilicon technology, using buried contacts.

There can be several application-specific variations in the basic SRAM cell. Figure 2-4(c) shows an eight-transistor, double-ended, dual-port static cell. This is useful in cache architectures, particularly as an embedded memory in a microprocessor chip. This chip can be simultaneously accessed through both of the

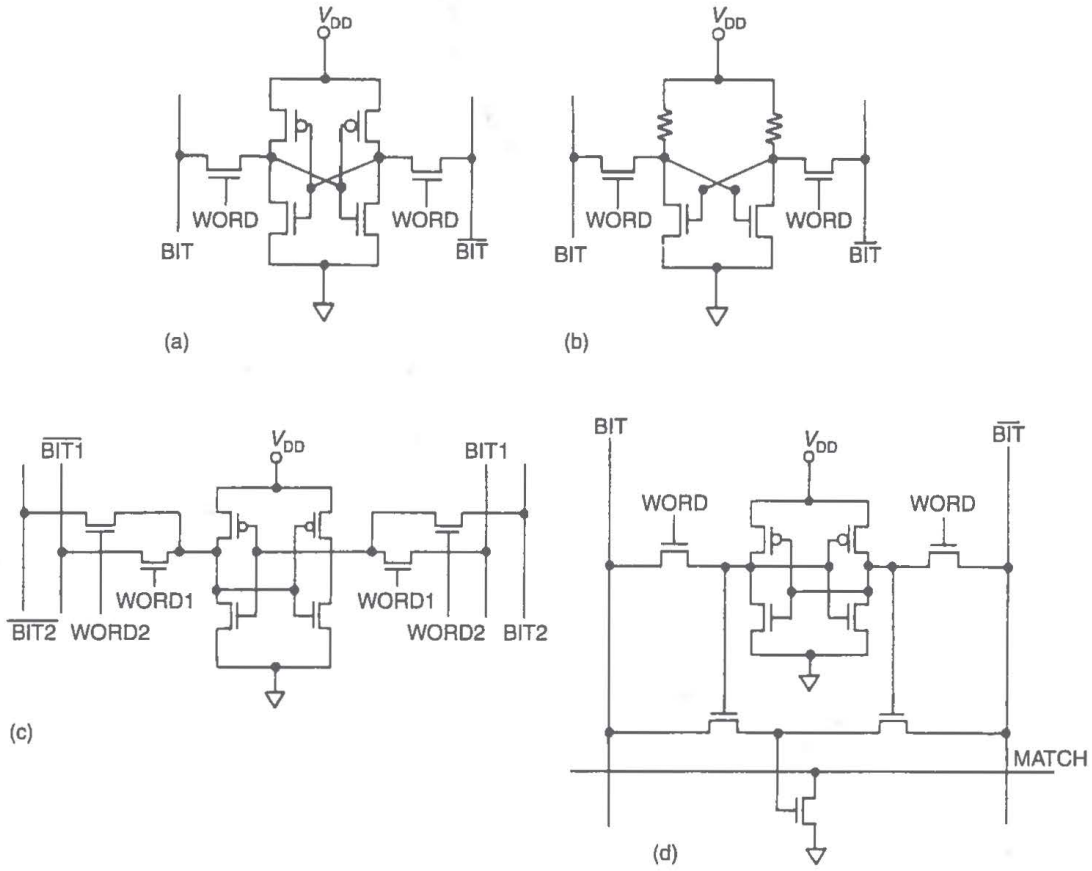


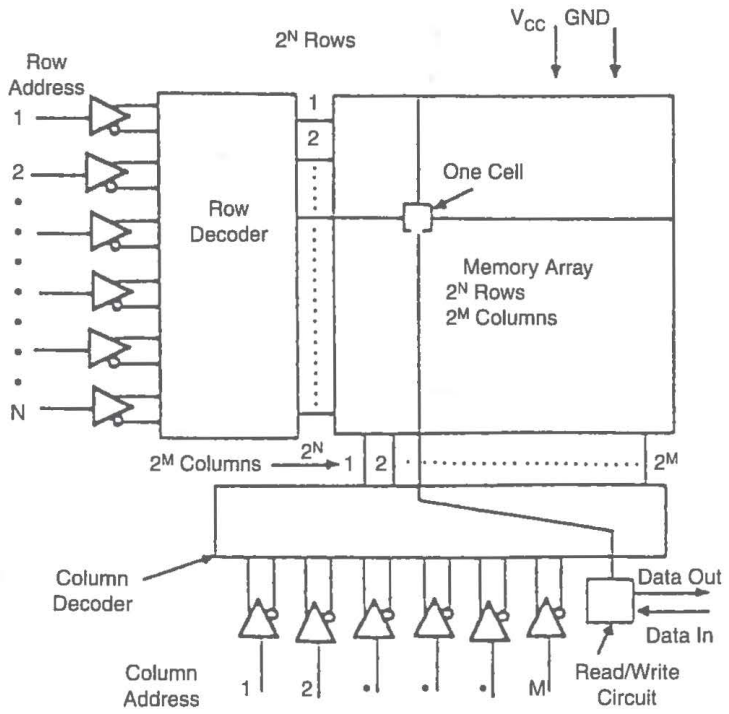
Figure 2-4. Various configurations of CMOS SRAM cells. (a) Six-transistor full CMOS. (b) Four transistors with *R*-load NMOS. (c) Dual-port with double-ended access. (d) Content-addressable memory (CAM).

ports. Figure 2-4(d) shows a nine-transistor content-addressable memory (CAM) cell. This is used in applications where knowledge of the contents of the cell, as well as location of the cell, are required.

2.2.2 MOS SRAM Architectures

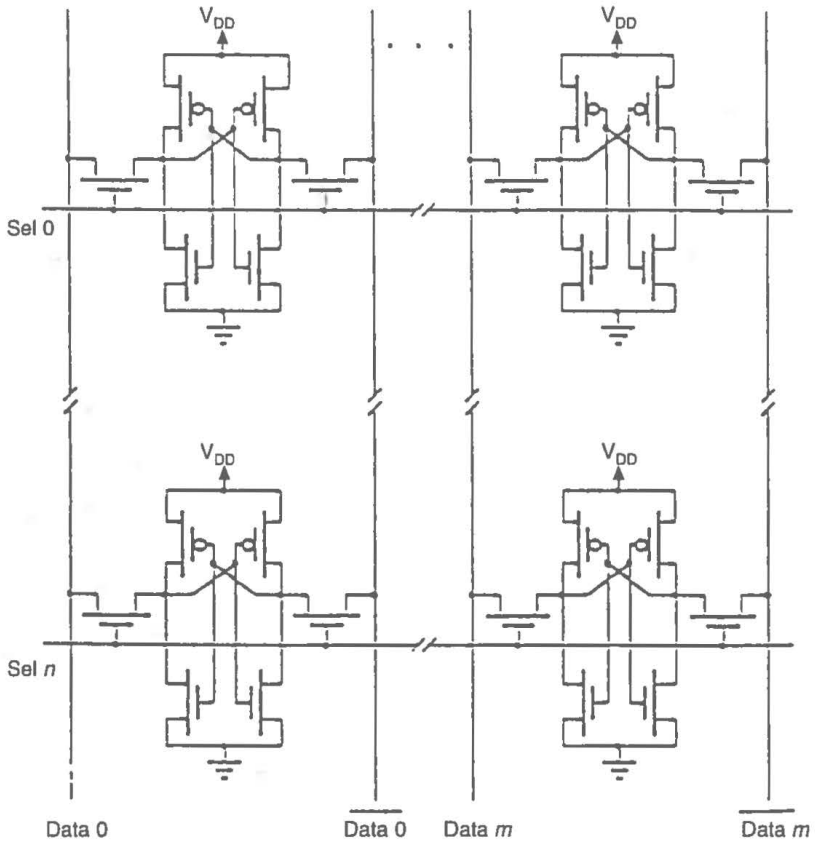
Figure 2-5 shows (a) a typical SRAM basic organization schematic, and (b) the storage cell array details. Each memory cell shares electrical connections with all the other cells in its row and column. The horizontal lines connected to all the cells in a row are called the "word lines," and the vertical lines along which the data flow in and out of the cells are called the "bit lines." Each cell can be uniquely addressed, as required, through the selection of an appropriate word and a bit line. Some memories are designed so that a group of four or eight cells can be addressed; the data bus for such memories is called nibble or one byte wide, respectively.

In a RAM, the matrix of parallel memory cells is encircled by the address decoding logic and interface circuitry to external signals. The memory array nominally uses a square or a rectangular organization to minimize the overall chip area and for ease in implementation. The rationale for the square design can be seen by considering a memory device that contains 16K 1-bit storage cells. A memory array with 16K locations requires 14 address lines to allow selection of each bit ($2^{14} = 16{,}384$). If the array were organized as a single row of 16 Kb, a 14-to-16K line decoder would be required to allow individual selection of the bits. However, if the memory is organized as a 128-row × 128-column square, one 7-to-128 line decoder to select a row and another 7-to-128 line decoder to select a column are required. Each of these decoders can be placed on the sides of the square array. This 128-row × 128-column matrix contains 16,384 cross-points which allow access to all individual memory bits. Thus, the square memory array organization results in significantly less area for the entire chip. However, the 16K memory may also be organized as a 64-row × 256-column (or 256 × 64) array.



(a)

Figure 2-5. (a) A typical SRAM basic organization schematic.



(b)

Figure 2-5 (cont.). (b) The storage cell array details.

Most RAMs operate such that the row address enables all cells along the selected row. The contents of these cells become available along the column lines. The column address is used to select the particular column containing the desired data bit which is read by the sense amplifier and routed to the data-output pin of the memory chip. When a RAM is organized to access n bits simultaneously, the data from n columns are selected and gated to n data-output pins simultaneously. Additional circuitry, including sense amplifiers, control logic, and tri-state input/output buffers, are normally placed along the sides of a cell array.

The two important time-dependent performance parameters of a memory are the "read-access time" and the "cycle time." The first timing parameter (access time) represents the propagation delay from the time when the address is presented at the memory chip until the data are available at the memory output. The cycle time is the minimum time that must be allowed after the initiation of the read operation (or a write operation in a RAM) before another read operation can be initiated.

2.2.3 MOS SRAM Cell and Peripheral Circuit Operation

A basic six-transistor CMOS SRAM cell and its layout are shown in Figure 2-6. The bit information is stored in the form of voltage levels in the cross-coupled inverters. This circuit has two stable states, designated as "1" and "0." If, in the logic state "1," the point C_5 is high and point C_6 is low, then T_1 is off and T_2 is on; also,

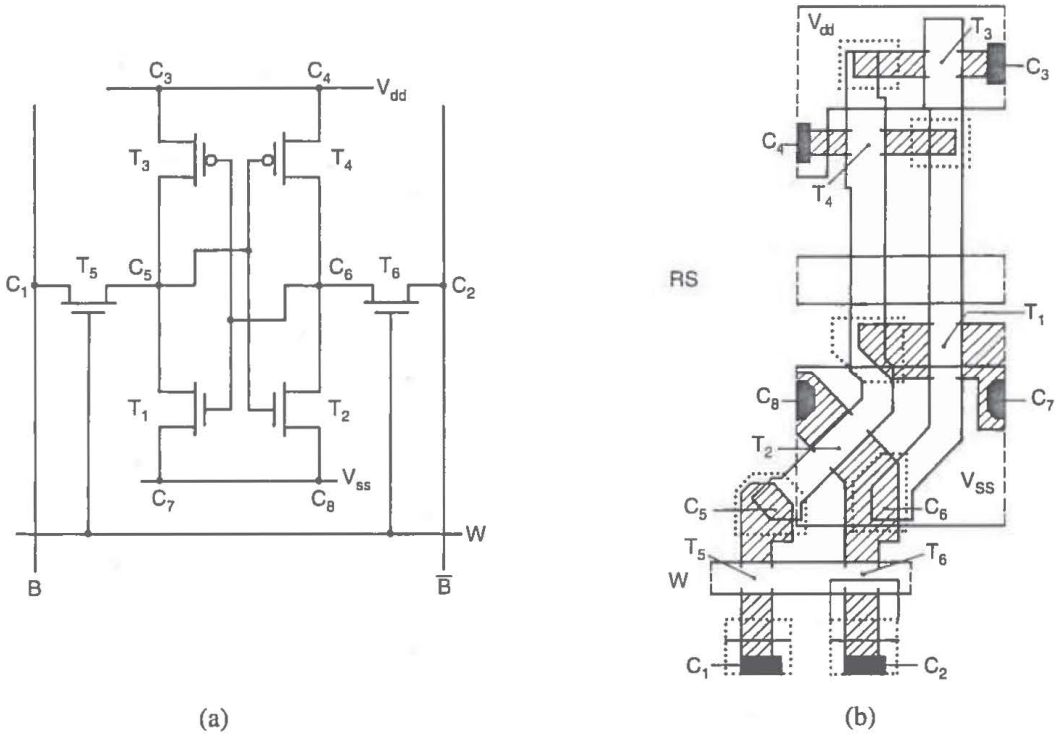


Figure 2-6. Six-transistor CMOS SRAM cell. (a) Schematic diagram. (b) Layout. (From Bastiaens and Gubbels [3], 1988, with permission of Philips Electronics, Eindhoven, The Netherlands.)

T_3 is on and T_4 is off. The logic "0" state would be the opposite, with point C_5 low and C_6 high. During the read/write operations, the row address of the desired cell is routed to the row address decoder which translates it and makes the correct word line of the addressed row high. This makes transistors T_5 and T_6 in all cells of the row switch "on." The column address decoder translates the column address, and makes connection to the bit line B and the inverse bit line $\overline{B}$ of all cells in the column addressed.

A READ operation is performed by starting with both the bit and $\overline{\text{bit}}$ lines high and selecting the desired word line. At this time, data in the cell will pull one of the bit lines low. The differential signal is detected on the bit and $\overline{\text{bit}}$ lines, amplified, and read out through the output buffer. In reference to Figure 2-6, reading from the cell would occur if B and $\overline{B}$ of the appropriate bit lines are high. If the cell is in state "1," then T_1 is off and T_2 is on. When the word line of the addressed column becomes high, a current starts to flow from $\overline{B}$ through T_6 and T_2 to ground. As a result, the level of $\overline{B}$ becomes lower than B. This differential signal is detected by a differential amplifier connected to the bit and $\overline{\text{bit}}$ lines, amplified, and fed to the output buffer. The process for reading a "0" stored in the cell is opposite, so that the current flows through T_5 and T_1 to ground, and the bit line B has a lower potential than $\overline{B}$. The READ operation is nondestructive, and after reading, the logic state of the cell remains unchanged.

For a WRITE operation into the cell, data are placed on the bit line and $\overline{\text{data}}$ are placed on the $\overline{\text{bit}}$ line. Then the word line is activated. This forces the cell into the state represented by the bit lines, so that the new data are stored in the cross-coupled inverters. In Figure 2-6, if the information is to be written into a cell, then B becomes high and $\overline{B}$ becomes low for the logic state "1." For the logic state "0," B becomes low and $\overline{B}$ high. The word line is then raised, causing the cell to flip into the configuration of the desired state.

SRAM memory cell array's periphery contains the circuitry for address decoding and the READ/WRITE sense operations. Typical write

circuitry consists of inverters on the input buffers and a pass transistor with a write control input signal to the bit and $\overline{\text{bit}}$ lines. Read circuitry generally involves the use of single-ended differential sense amplifiers to read the low-level signals from the cells. The data path is an important consideration with the SRAMs, since the power delay product is largely determined by the load impedance along the data path. The read data path circuitry can be static or dynamic. SRAMs have been designed which turn on or turn off the various sections of data path as needed to reduce the operating power of the device. The internal signals in an outwardly "static" RAM are often generated by a technique called "Address Transition Detection" (ATD) in which the transition of an address line is detected to generate the various clock signals. The input circuitry for SRAMs consists of the address decoders, word line drivers, and decoder controls. Figure 2-7 shows the various SRAM circuit elements.

RAMs are considered clocked or not clocked, based on the external circuitry. Asynchronous (nonclocked) SRAMs do not require external clocks, and therefore have a very simple system interface, although they have some internal timing delays. Synchronous (clocked) SRAMs do require system clocks, but they are faster since all the inputs are clocked into the memory on the edge of the system clock.

2.2.4 Bipolar SRAM Technologies

The earliest semiconductor memories were built in bipolar technology. Nowadays, bipolar memories are primarily used in high-speed applications. Bipolar RAMs are often "word-oriented" and require two-step decoding. For example, in a 1-kb × 1-b memory organized as a 32-row × 32-column array, the row decoder selects one of the 32 rows, and all of the 32 b (the "word") are read out and placed in a register. A second 5 b code is used to access the

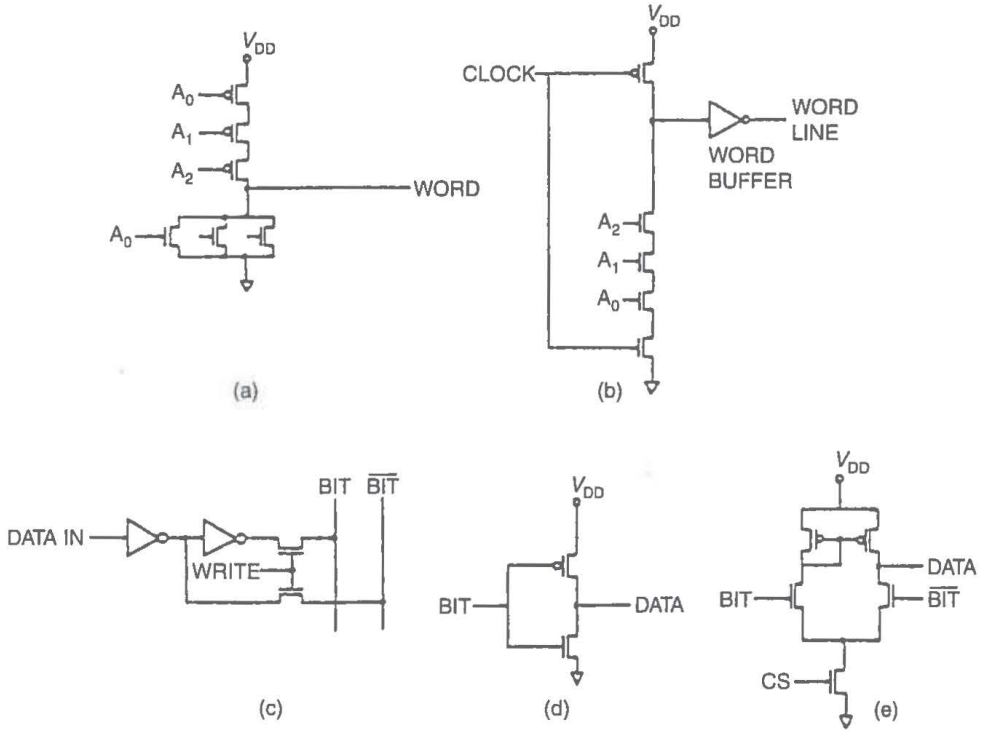


Figure 2-7. Various SRAM circuit elements. (a) Static row decoder. (b) Dynamic row decoder. (c) Simple write circuitry. (d) Inverter amplifier. (e) Differential sense amplifier. (From [4], with permission of Wiley, New York.)

register and select the desired bit. Similarly, the data are stored by writing an entire word simultaneously. The broad category of bipolar memories include the SRAMs fabricated in direct-coupled transistor logic (DCTL), emitter-coupled logic (ECL), and mixed BiCMOS technologies.

2.2.4.1 Direct-Coupled Transistor Logic (DCTL) Technology. The architectures for bipolar SRAMs are basically similar to those of MOS SRAMs. Historically, transistor–transistor logic (TTL) has been the most commonly used bipolar technology. A simple bipolar DCTL RAM memory cell consists of two bipolar transistors, two resistors, and a power source. In this configuration, one of the transistors is always conducting, holding the other transistor OFF. When an external voltage forces the OFF transistor into conduction, the initially ON transistor turns off and remains in this condition until another external voltage resets it. Since only one of the cross-coupled transistors conducts at any given time, the circuit has only two stable states which can be latched to store information in the form of logic 1s and 0s. The cell state is stable until forced to change by an applied voltage.

The circuit in Figure 2-8 is an expanded version of the DCTL memory cell [5], [37]. In this figure, the data lines are connected to Q_1

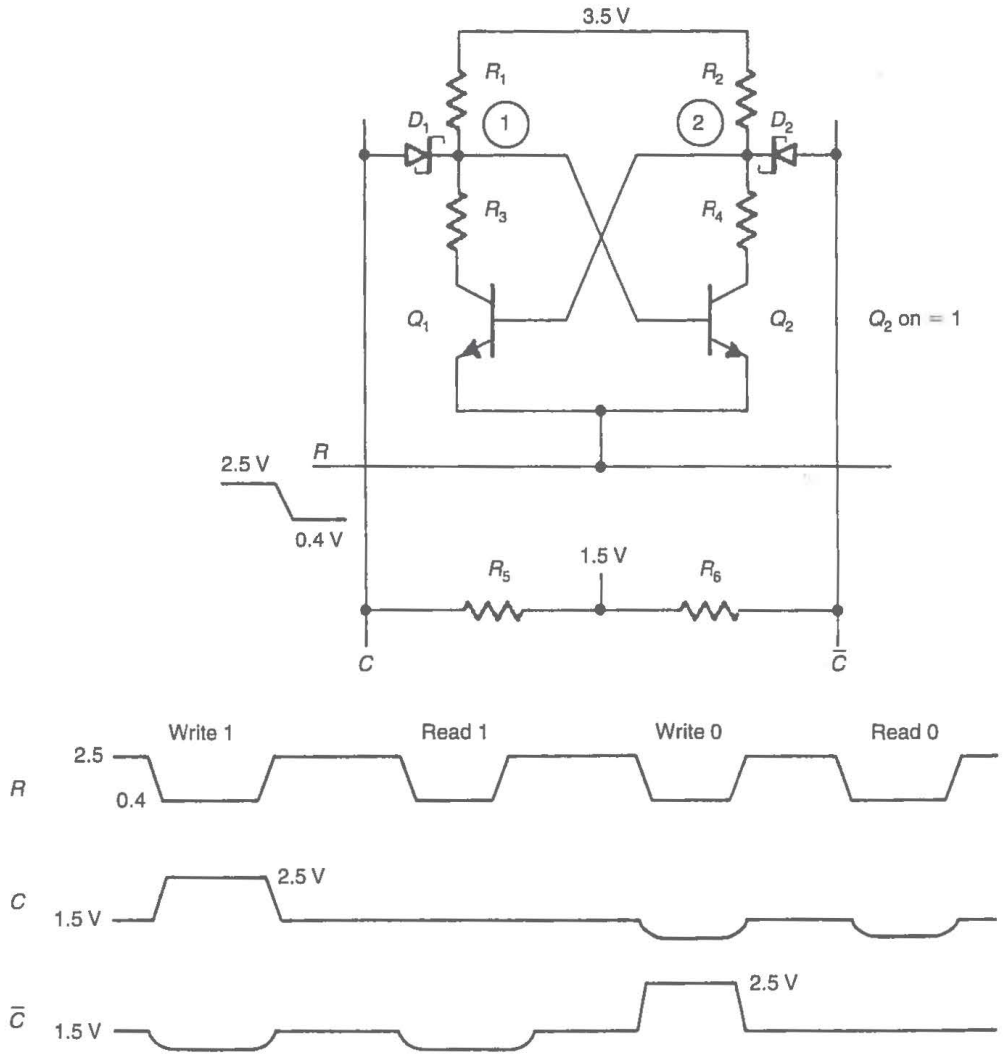


Figure 2-8. Direct-coupled memory cell (DCTL) with Schottky diodes. (From D. A. Hodges and H. Jackson [37], with permission of McGraw-Hill Inc., New York.)

and Q_2 through the Schottky diodes D_1 and D_2. To explain the operation of this cell, assume that a stored 1 corresponds to the state with Q_2 on. The row selection requires that the row voltage should be pulled low. To write a "1," the voltage on line C is raised, forward-biasing diode D_1. This forces sufficient current through R_1 so that the voltage at node 1 increases to turn on Q_2. The current gain of Q_1 is sufficiently high, and it remains in saturation so that most of the voltage drop appears across R_3. When Q_2 turns on, its collector voltage drops rapidly, turning off Q_1. The currents in R_1 and R_2 are always much smaller than the current used for writing, so that the voltage drops across R_3 and R_4 are much smaller than $V_{be(on)}$. In the standby condition, D_1 and D_2 are reverse-biased. To read a stored "1," the row is pulled low, and current flows through C through D_2, R_4, and Q_2 to R. The resulting drop in voltage on C indicates the presence of a stored "1."

2.2.4.2 Emmiter-Coupled Logic (ECL) Technology. Another popular bipolar technology is the emitter-coupled logic (ECL). The ECL memories provide very small access times with typical propagation delays of less than 1 ns and clock rates approaching 1 GHz. This high performance is achieved by preventing the cross-coupled transistors from entering into the saturation region. Figure 2-9 shows an ECL memory cell [37]. In this configuration, the data lines are connected to the emitters of the two transistors. Although both transistors have two

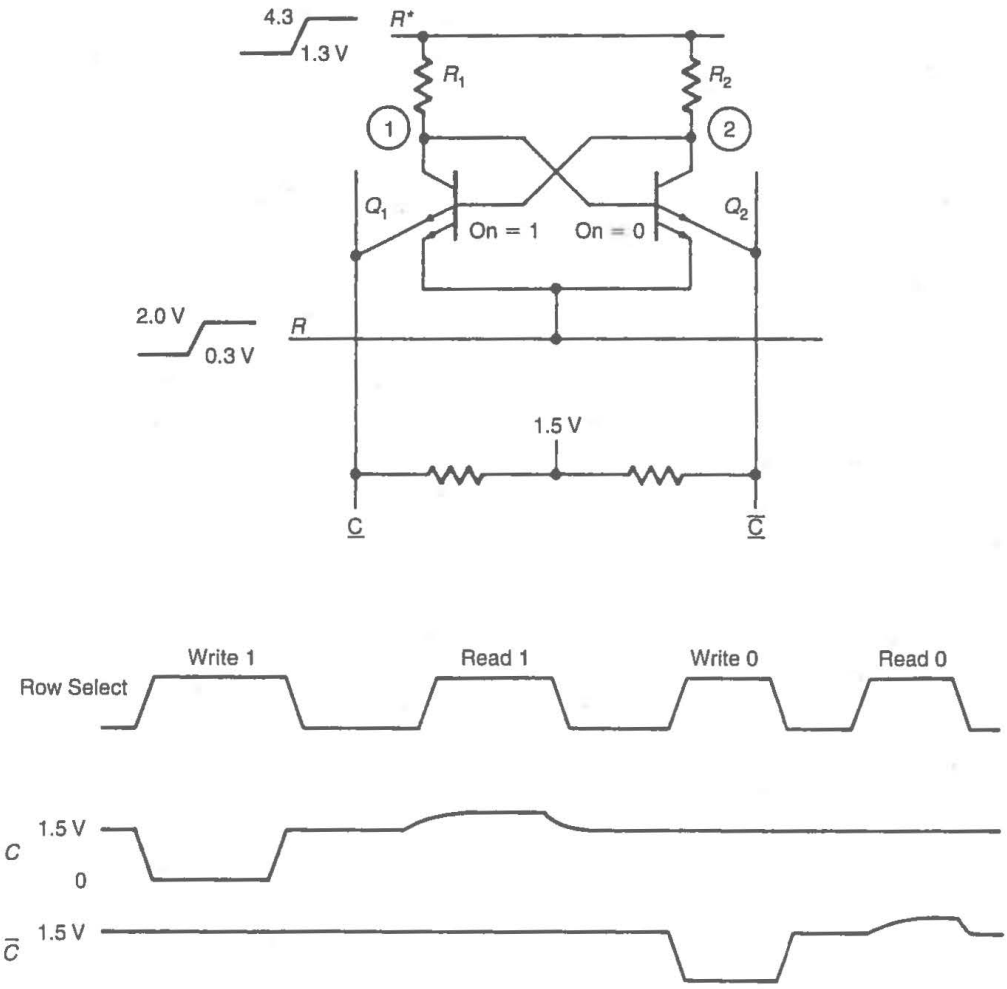


Figure 2-9. Emitter-coupled logic (ECL) memory cell. (From D. A. Hodges and H. Jackson [37], with permission of McGraw-Hill Inc., New York.)

emitters each, this cell is not a TTL circuit because they operate in their normal mode (as opposed to the inverted) modes. The operation of the cell is based on using the multiple-emitter transistors as the current switches. The voltage levels are selected such that these transistors never conduct simultaneously. The read and write operation is controlled by switching the current in the conducting transistor from the row line to the appropriate data line.

The basic operation of an ECL cell is explained with Figure 2-9. Assume that a logic "1" is stored with Q_1 on. The row selection requires that both R and R^* go to the positive levels as shown. To write a "1," the column line C must be held low, which forward-biases the emitter of Q_1, regardless of the previous state of the cell. As a result, the collector–emitter voltage of Q_1 drops quickly, removing the base drive from Q_2. When the row voltage returns to the standby levels, Q_1 remains "on," with its base current coming from R_2. Cell current flows through Q_1 and returns to ground through the line R. The emitters connected to C and $\overline{C}$ are reverse-biased in the standby condition. To read a stored "1," the cell is selected in the same way as for writing. The emitters connected to R become reverse-biased, and the current flowing in Q_1 transfers to the emitter connected to C. The resulting rise in voltage on C indicates the presence of a stored "1." The writing and reading operation for a "0" are complementary to those described for a "1."

A major drawback in the ECL technology is that a very low value of resistors is required, which results in a constant current in the storage cell. This constant current drain causes relatively higher power dissipation compared to the TTL or MOS memories.

Another bipolar logic developed in the 1970s was the integrated injection logic (I^2L). This evolved from the bipolar memory cell design in which the old direct-coupled transistor logic (DCTL) was shrunk to the single complementary transistor equivalent. In memory cells, lateral p-n-p transistors were used as the current source and multicollector n-p-n transistors as the inverters. The major attractive feature of this technology was the high packaging density since the cells were quite compact, in contrast to the conventional bipolar memory cells using resistors for the load impedance of the flip-flop transistors. However, I^2L technology for memories failed to be commercially successful because of the process- and structure-related dependencies of its speed–power performance characteristics. The BiCMOS technology was preferred for high-speed and high-performance applications.

2.2.4.3 BiCMOS Technology. In high density and high-speed applications, various combinations of bipolar and MOS technologies have been investigated. The BiCMOS process is more complex because of the additional steps required. Bipolar ECL input and output buffers have been used with the CMOS memories, both to interface to a bipolar circuit as well as to increase the performance. The CMOS memory cells have lower power consumption, better stability, and a smaller area compared to an equivalent bipolar cell. Therefore, the BiCMOS designs offer optimization of these parameters, since the various circuit elements are selectively chosen to maximize the performance [6]. For example, since the bipolar n-p-n output transistors provide large output drive, BiCMOS gates are effective with high capacitive nodes, such as those in the decoders, word line drivers, and output buffers. The control logic with small fan-out can still use CMOS gates. The sense amplifiers which require high gain and high input sensitivity for fast sensing of small differential bit line swings are built in bipolar logic. Figure 2-10 shows the BiCMOS circuit elements with a typical mix of various technologies [7]. There can be some variations on this technology mix, depending upon the manufacturer's process.

In a typical double-polysilicon technology CMOS SRAM process, the first poly level forms the MOS transistor gate; the second poly level forms highly resistive load resistors in the four-transistor CMOS memory cell, and makes contact to the silicon substrate. Since a poly-to-substrate contact already exists in such a process, a bipolar transistor with a polysilicon

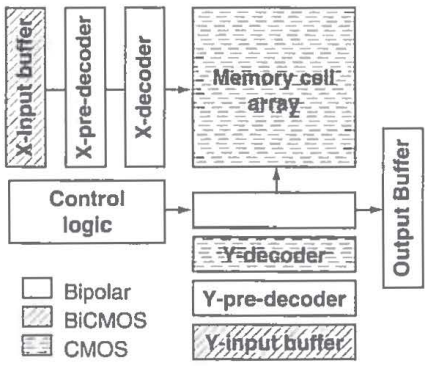


Figure 2-10. BiCMOS circuit elements.

emitter can be produced with minimum complexity. The emitter is created by a dopant diffusion from a doped poly into the silicon substrate where an implanted base is present. Figure 2-11(a) shows the schematic cross-section of a double-polysilicon SRAM with second poly as a bipolar emitter [8].

A high-performance CMOS process can be obtained from a core CMOS process by the addition of a few steps such as: (1) use of a buried layer and an epitaxial layer, (2) a base region for bipolar transistors usually by ion implantation, and (3) an additional collector region (collector sink or collector plug) to reduce collector resistance. Typically, the added steps may increase process complexity by 10–20%. Figure 2-11(b) shows the typical enhancements to a core double-polysilicon process for creating high-performance BiCMOS memories.

The higher density commodity SRAMs tend to exhibit a balance of speed and power characteristics. The various density levels of very high-speed CMOS and BiCMOS SRAMs are specifically designed to optimize performance, often at the expense of chip area and power consumption. The new generation BiCMOS includes many improvements, such as the use of fast data bus in the sense circuit; parallel test circuits using the ECL sense amplifiers; optimized division of memory matrix to reduce parasitic capacitance on the signal lines (bit, word, and decoder); and column redundant circuits suitable for divided word line architectures. The use of low-resistance metal interconnects al-

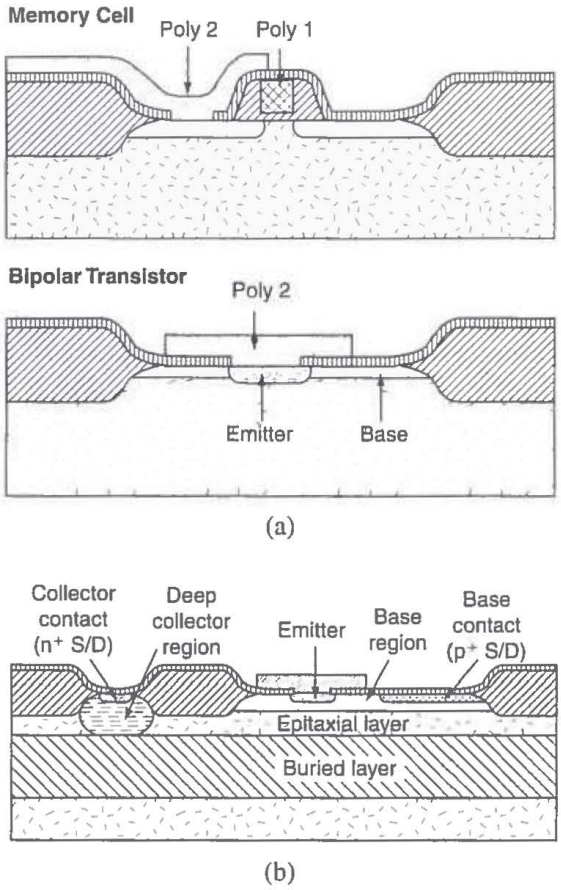


Figure 2-11. Schematic cross-sections. (a) A double-polysilicon SRAM with second poly as bipolar emitter. (b) Typical process enhancements for creating BiCMOS [8].

lows more cells per bit-line pair without significantly increasing the delay.

A 4 Mb BiCMOS SRAM was developed by Fujitsu in 1991. It uses a wired-OR data bus in the sense circuit, a column redundant circuit, and a 16 b parallel test circuit [9]. Figure 2-12(a) shows a block diagram of this SRAM. It has a typical address access time of 7 ns, a write pulse width of 4 ns, and active current consumption of about 120 mA for a 4M × 1-b configuration. A symmetrical chip layout divides the bit lines, word lines, and decoder lines to reduce the ground noise from high-speed operation. The memory matrix is divided into four planes, each consisting of 16 blocks. The column redundant circuit is shown in Figure 2-12(b). The

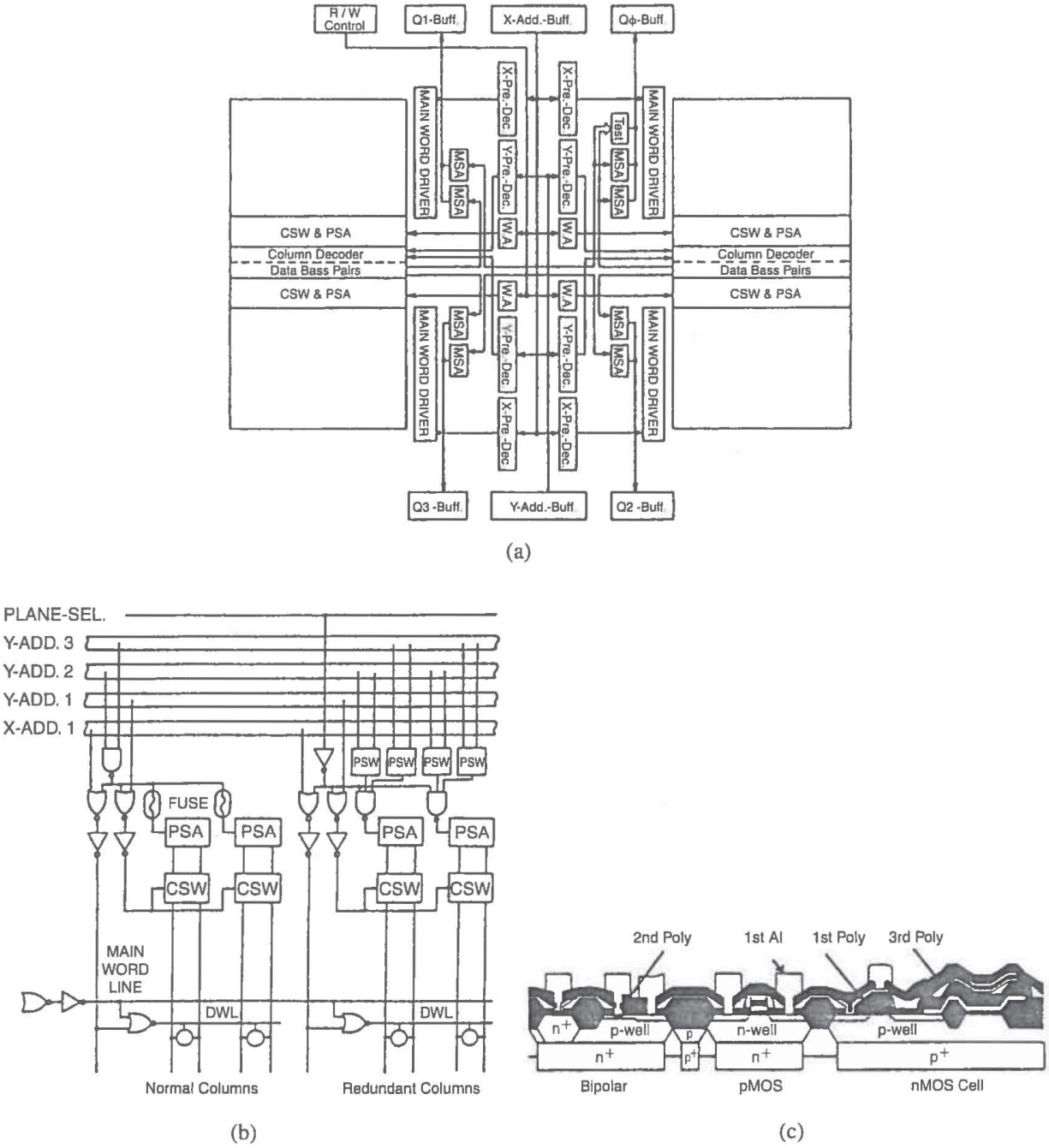


Figure 2-12. Fujitsu 4 Mb SRAM. (a) Block diagram. (b) Column redundant circuit. (c) Cross-section. (From [9], with permission of IEEE.)

redundancy is provided by two rows and eight columns, which provide more flexibility than the conventional architecture. The column arrays are distributed throughout each block and can replace cells within the block. In this 4 Mb SRAM, only four redundant column arrays are placed in the 1 Mb plane to form a redundant section, and they are used for any column array in the plane. The process technology is 0.6 μm triple-polysilicon and double-metal BiCMOS.

The first polysilicon layer forms the gate electrodes of the MOS transistors, and the second layer forms the emitter electrodes of the bipolar transistors. For memory cells, the V_{EE} lines are formed with the second polysilicon layer, and the 50 GΩ loads are formed with the third polysilicon layer. The minimum gate lengths are 0.6 and 0.8 μm for the NMOS and PMOS transistors, respectively. The thickness of the gate oxide is 15 nm. Figure 2-12(c) shows the cross-section of this BiCMOS device.

A 0.5 μm BiCMOS technology has been developed for fast 4 Mb SRAMs in which bipolar transistors are added to an existing 0.5 μm CMOS process [10]. This process requires the growth of a thin epitaxial layer, as well as the addition of three masking steps as follows: self-aligned buried layer, deep collector, and active base. The original CMOS process featured self-aligned twin-well formation, framed mask poly-buffered LOCOS (FMPBL) isolation, a 150 Å thick gate oxide, surface-channel NMOS and buried-channel PMOS transistors, disposable polysilicon spacer module, three levels of polysilicon, and two layers of metallization. Figure 2-13 shows the schematic cross-section of this technology [10]. Three levels of polysilicon are used in this process. The first layer is the gate electrode for the CMOS transistors. The second layer is the tungsten–polycide/polysilicon stack that performs the following three functions: creating self-aligned contact landing pads in the SRAM bit cell, forming an emitter of the bipolar n-p-n transistor, and providing a global interconnect. The third polysilicon layer forms the teraohm resistor load for the bit cell. This process provides a peak cutoff frequency (f_T) of 14 GHz with a collector–emitter breakdown voltage (BV_{CEO}) of 6.5 V, and ECL minimum gate delays of 105·ps at a gate current of 350 μA/μm^2 have been achieved.

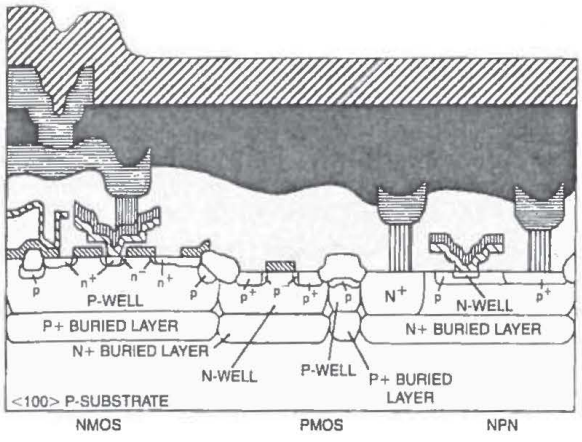


Figure 2-13. Schematic cross-section of the 0.5 μm BiCMOS technology. (From [10], with permission of IEEE.)

The mainframe computers and supercomputers require high-speed ECL I/O SRAMs, whereas the high-performance workstations usually require TTL I/O SRAMs operating with microprocessors at low supply voltages. In order to meet both of these requirements, an approach has been developed that uses the metal mask operation technique to fabricate a 4 Mb BiCMOS SRAM that can have either a 6 ns access time at 750 mW and 50 MHz, or an 8 ns access time at 230 mW and 50 MHz using an ECL 100K or 3.3 V TTL interface, respectively [11]. A 3.3 V supply voltage is desirable for MOSFETs with a 0.55 μm gate length to ensure higher reliability. For an ECL 100K interface, ECL I/O buffers are supplied at −4.5 V V_{EE}, while the 4 Mb SRAM core circuit is supplied at −3.3 V $V_{EE'}$, which is generated by an internal voltage converter. The 4 Mb SRAM core consists of BiCMOS address decoders, MOS memory cells, and bipolar sense amplifiers.

In order to achieve high-speed address access times, a combination of the following technologies were used:

- A Bi-NMOS converter that directly converts the ECL level to a CMOS level
- A high-speed Bi-NMOS circuit with low-threshold-voltage NMOSFETs
- A design methodology for optimizing both the number of decoder gate stages and the size of gate transistors
- High-speed bipolar sensing circuits for 3.3 V supply voltage.

This process was used to fabricate a 4 Mb SRAM in 0.55 μm BiCMOS technology with

quadruple-level polysilicon, double-level metal, p-substrate, and triple-well. The first-level polysilicon layer was used for the MOSFET gate electrodes, the second poly for $V_{EE'}$ (−3.3 V), the third poly for high resistive loads in memory cells, and the fourth level poly for bipolar emitter electrodes. In memory cell arrays, double-level metal layers were used for bit lines (first level) and word lines (second level). The memory cell size is $5.8 \times 3.2\ \mu m^2$ and consists of two high resistive loads and four NMOSFETs.

An experimental 16 Mb BiCMOS SRAM has been developed in 0.4 μm process to achieve 220 MHz fully random read–write operations [111]. The memory circuit design techniques include the following: (1) zigzag double word line, (2) centralized bit-line load layout, and (3) phase-locked loop (PLL) with a multistage-tapped (MST) voltage-controlled oscillator (VCO). This VCO consists of a voltage–current converter and a nine-stage current-limited CMOS inverter ring oscillator. The internal SRAM timing pulses are generated from the output signals at each stage and are proportional to the clock cycle. Since all the timing margins can be expanded by making the clock cycle time longer, this technique allows more flexibility for high-frequency timing designs. The PLL is used as an internal clock generator to achieve a clock frequency above 200 MHz.

In this 16 Mb BiCMOS memory design, the internal SRAM macro is made fully static (asynchronous) by using the bipolar sensing circuits which operate at high speed with only small bit-line voltage swing and do not require clocked equalization that is used in the CMOS sensing operations. It uses the wave pipeline technique, which implies that at a certain time period, two signals (waves) exist simultaneously on the data path. The bit-line delay is minimized by optimizing the location of bit-line load (BL load) PMOSFETs. To reduce the time for write to the same as that for the read operation, postwrite equalizer (PW Eq) PMOSFETs are distributed to three positions on a bit-line pair. A write cycle time of 4.5 ns is achieved using this layout.

2.2.5 Silicon-on-Insulator (SOI) Technology

The MOS and bipolar technologies discussed so far are referred to as the bulk silicon technologies. Traditionally, dielectrically isolated (DI) processing has been used to reduce the charge generation in the substrate and to obtain radiation hardness for military and space applications. This fabrication technique is particularly suited to bipolar devices which need a thick active layer. Since the 1960s, a significant amount of work has been done using sapphire for both substrate and insulator; this technology was called silicon-on-sapphire (SOS), and is a restricted version of the SOI technology. The enhanced device performance (increase in speed) of SOI device structures results from the reduced parasitic substrate capacitance, while the improved radiation hardness results from the absence of charge generation in the silicon substrate.

In SOI technology, small islands of silicon are formed on an insulator film. Using SOI fabrication processes coupled with lateral isolation techniques, completely latchup free circuits can be fabricated because there are no parasitic p-n-p-n paths. Also, the SOI films can be made extremely thin, resulting in fully depleted MOSFETs which have been shown to operate at very high speeds (2 GHz) [12]. Thin-film SOI/CMOS devices and circuits have reduced short channel effects, reduced hot electron generation, and sharper subthreshold slopes [13]. However, because of relatively low wafer yield problems, the major market for SOI chips is in military and aerospace applications. The real impact of SOI technology on the commercial market is expected to occur in the manufacturing of submicron devices (0.25 μm) since, at that point, the production costs for bulk/epi wafers will be greater or equivalent to the SOI wafers.

Some of the important SOI technologies used to produce the starting material include: separation by implanted oxygen (SIMOX), zone melt recrystallization (ZMR), full isolation by porous oxidized silicon (FIPOS), epitaxial

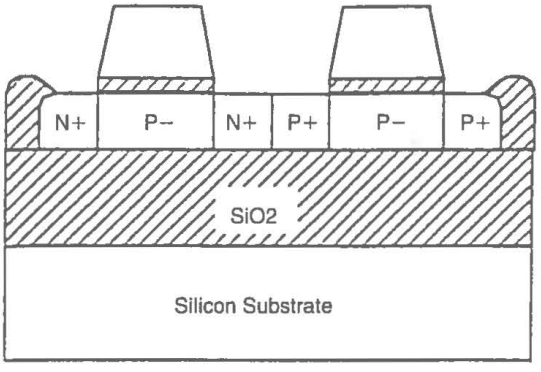
lateral overgrowth (ELO), wafer bonding and etchback (a modified DI process), and silicon-on-sapphire (SOS) process.

SIMOX is the most commonly used technique, and requires a high dose of oxygen (O^-) ions, which provides the minimum concentration necessary to form a continuous layer of SiO_2. The energy of the implant must be high enough so that the peak of the implant is sufficiently deep within the silicon (0.3–0.5 μm). The wafer is normally heated to about 400°C during the implantation process to ensure that the surface maintains its crystallinity during the high implantation dose. A postimplant anneal is performed in neutral ambient (N_2) or He for a sufficient time (3–5 h) at a high temperature (1100–1175°C) to form a buried layer of SiO_2. This anneal step also allows excess oxygen to out-diffuse, thereby increasing the dielectric strength of the buried-oxide layer. After the annealing step, the crystalline silicon surface layer is typically only 100–300 nm thick [14]. Therefore, an additional layer of epitaxial silicon is deposited so that single-crystal regions (0.5 μm or greater) are available for fabrication.

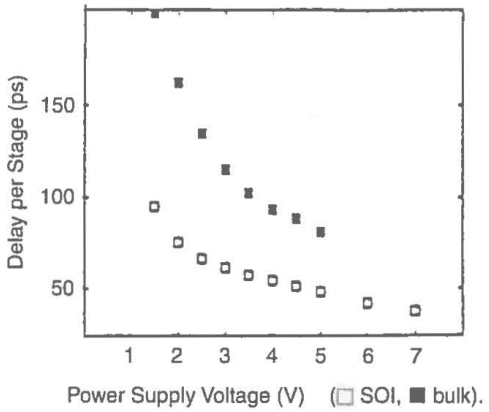
A simplified cross-section of a thin-film SOI CMOS inverter is shown in Figure 2-14(a) [15]. A comparison was made between the device and the circuit performance of 0.5 μm CMOS on an ultra-thin SOI ring oscillator with 0.4 μm bulk CMOS devices. Figure 2-14(b) shows smaller delays (or higher operating speeds) per stage for the SOI devices relative to bulk CMOS [16].

SOI 64 kb SRAM fully functional prototypes have been demonstrated by some companies. The SOI SRAMs with 256 kb and higher densities are under development. SIMOX is attractive for military and space electronic systems requiring high-speed, low-power, radiation-hardened devices that can operate reliably over a wide temperature range. It is a good candidate for power IC and telecommunications applications which must integrate low-voltage control circuits with the high-voltage devices [14]. This technology is well suited for integrating the bipolar and CMOS devices on the same chip. The use of wafers on which no epitaxial layer is deposited following oxygen implantation and annealing (referred to as thin-film SIMOX) has also been studied. Thin-film SIMOX processing is simpler than a bulk process because there is no need to form doped wells, and shallow junctions are automatically formed when the doping impurity reaches the underlying buried oxide. However, the SIMOX process does have a number of disadvantages, including the following:

- The process requires the availability of special oxygen implanters.
- The thickness of the buried oxide layer is limited to about 0.5 μm, whereas for



(a)



(b)

Figure 2-14. (a) Cross-section of a thin-film SOI CMOS inverter [15]. (b) Delay for bulk CMOS and SIMOX for a ring oscillator [16]. (From [15] and [16], with permission of IEEE.)

some high-voltage applications, a greater thickness may be required.

- The SOI surface silicon film and the deposited epitaxial layer may contain some residual damage from the implant process. The implantation parameters and annealing schedules must also be appropriately chosen to provide optimum IC performance because the microstructure of the surface silicon film is very sensitive to the oxygen dose and the postoxygen-implant annealing temperature.

Due to the floating substrate, SOI MOSFETs show a dip in their *I–V* characteristics, generally known as "kink effect." However, in the SOI RAMs, this kink effect has been eliminated by electrically shorting the source and substrate. In the zone-melting recrystallization (ZMR) process, which is currently in evolution, the basic steps consist of oxidation and poly-Si deposition, followed by a zone-melting technique to recrystallize the polysilicon. The melting of the polysilicon layer can be accomplished by lasers, electron beams, arc lamps, or strip heaters. The ZMR fabrication equipment is relatively inexpensive compared to that for the SIMOX process. The ZMR defect densities are being improved through seeding, the use of encapsulation layers, and other process-enhancement techniques [13].

The silicon-on-sapphire (SOS) technology has been extensively used for military and space applications. In SOS, the active device is built upon a silicon island on a sapphire substrate. Thus, all n-channel transistors are completely isolated from the p-channel transistors by the sapphire substrate. There are no parasitic transistors to latch up in the SOS, and the capacitance associated with the reverse-biased substrate junctions is eliminated. This gives the SOS process an important advantage over the bulk CMOS in terms of switching speed [17]. The sapphire substrate provides inherent radiation hardness against the alpha-particle-induced soft errors; ionizing total-dose, transient, and neutron radiation; and cosmic ray upsets. These radiation effects and circuit-hardening techniques are discussed in more detail in Chapter 7.

The early CMOS/SOS logic devices such as the gates, flip-flops, buffers, and small counters were fabricated with an aluminum-gate process on n-type silicon islands, where the NMOS and PMOS transistors were formed and interconnected. The gate aligned over the channel after formation of the channel region overlapped both the source and drain regions to assure a total channel coverage. However, this overlap created gate-to-source and gate-to-drain capacitance, which increased the switching delays. This Al-gate process was subsequently replaced by the silicon-gate technology, which offered improved performance. In the silicon-gate process, the channel oxide and the polysilicon gate structure are defined prior to the source and drain. The polysilicon gate then serves as a mask to define both the source and the drain; thus, the gate is perfectly aligned with no overlap in this "self-aligned silicon-gate" technique. In the following steps, an oxide is deposited (or grown) over the gate, contact windows are defined in the oxide, and aluminum metallization is applied for the interconnections. Figure 2-15 shows the various processing steps for RCA (now Harris Corporation) CMOS/SOS technology [17].

In summary, the key features of a SIMOX-based process are the extended temperature range compared to bulk silicon, lower parasitic capacitance from reduced junction areas, and dielectric isolation which provides latchup immunity. SIMOX is the most commonly used SOI technique suitable for VLSI fabrication, although it requires expensive ion implanters and long implant times. Thomson-CSF Semiconductors uses an SOI-HD process with side and bottom oxides which provide total isolation of the device. Figure 2-16 shows the cross-section of this thin CMOS SOI-HD structure which uses a special gate oxide process to achieve high resistance against various types of radiation, including the neutrons and heavy ions [112]. This technology, available in two versions as 1.2 μm (HS013-HD) and 0.8 μm (HS014-HD) design rules with two and three metal levels, is used in fabricating 16, 64, and 256K SRAMs.

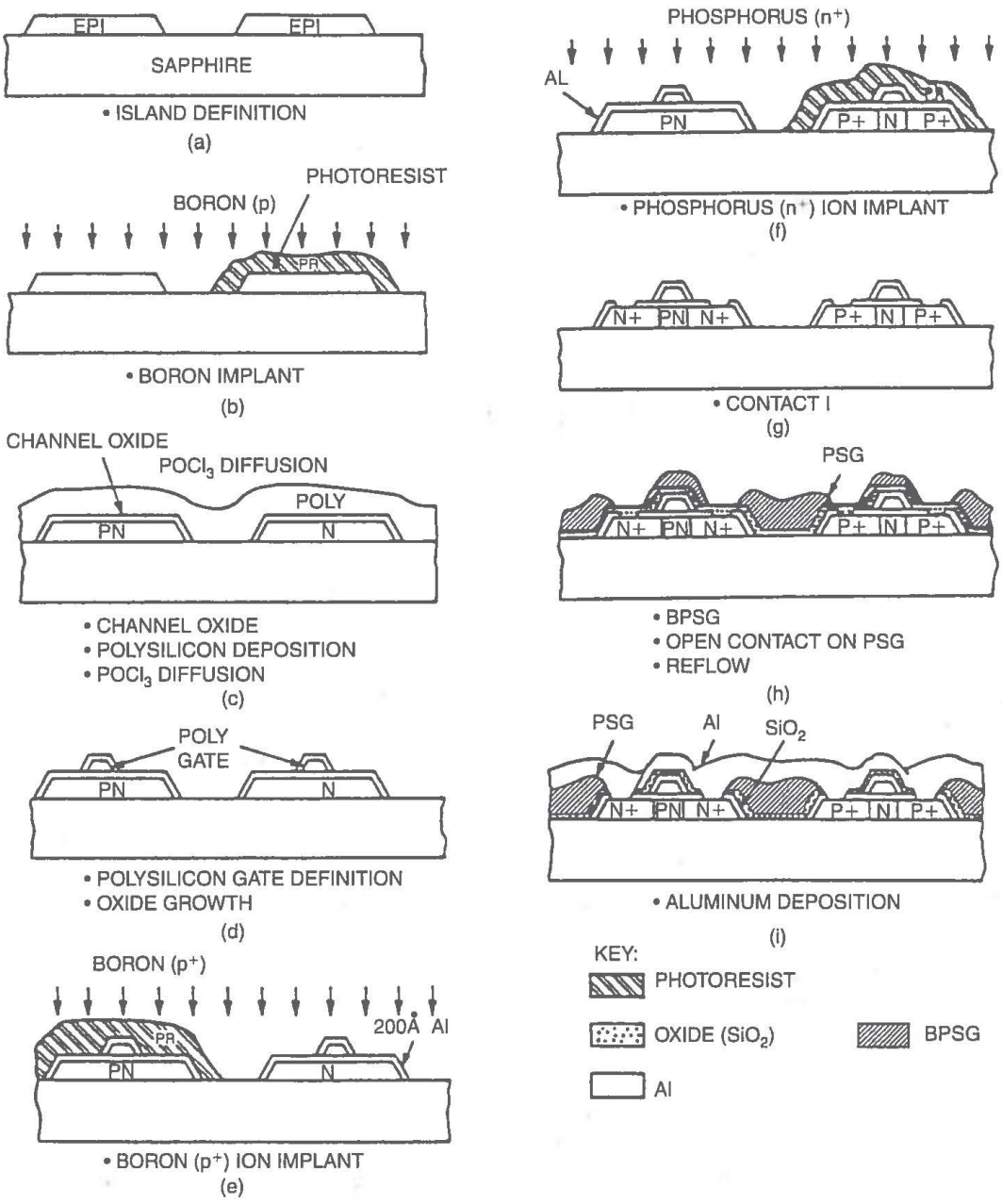


Figure 2-15. Typical RCA CMOS/SOS processing steps. (From *RCA Databook* [17], with permission of Harris Corporation.)

As bulk CMOS technology advances toward sub-0.25 μm channel length, the SOI technologies appear promising because of their reduced parasitic capacitance, lower threshold voltage and body-charging effects. In general, substantial advantages are being projected for the SOI memory technology, especially in the military and space environment, because of a substantial reduction in the soft-error rate and latchup elimination. The SOI MOSFETs can be made fully depleted, allowing only unidirectional current flow in the circuit. This virtually eliminates the kink effect normally associated with the SOS/SOI MOSFETs because of the floating substrate. The hot electron effect which causes threshold instabilities in the MOS devices through charge injection in the gate oxide is also reduced in the fully depleted SOI.

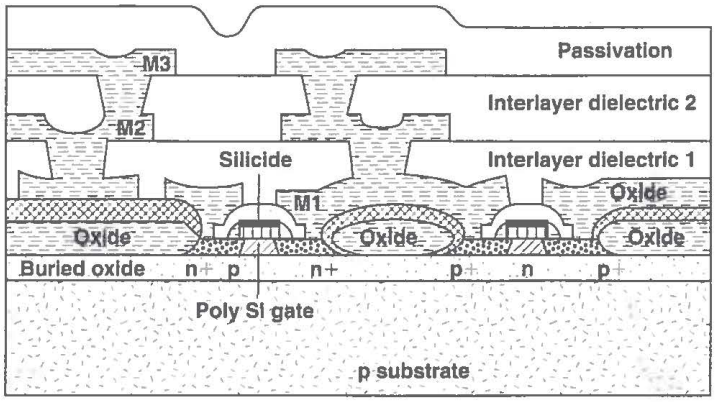


Figure 2-16. Thomson-CSF Semiconductors CMOS SOI-HD process cross-section [112].

A new low-power and high-stability SOI SRAM technology has been developed by using laser recrystallization techniques in which self-aligned, p-channel SOI MOSFETs for load devices in memory cells are stacked over bottom n-channel bulk MOSFETs for both driver and transfer gates in a 3-D configuration [18]. For process simplicity, no specific planarization process is utilized for an intermediate insulating layer between the top SOI and bottom bulk MOSFETs. This technique increases the on/off ratio of SOI MOSFETs by a factor of 10^4, and reduces the source–drain leakage current by a factor of 10–100 compared with that for polysilicon thin-film transistors (TFTs) fabricated by using low-temperature regrowth of amorphous silicon. A test 256 kb SRAM was fabricated using this process with only memory cell regions in a stacked 3-D structure and peripheral circuits in conventional CMOS technology. The total number of process steps is roughly the same as that for polysilicon TFT load SRAMs. The process steps can be reduced to 83% of those for the TFT load SRAMs if both the peripheral circuit and memory cells are made with p-channel SOI and n-channel bulk MOSFETs.

In recent years, several circuits with operating frequencies above 400 MHz have been designed in the industry standard CMOS processes. Typically, these designs are highly pipelined, i.e., they use a clocking strategy with a single clock net. Also, these designs use careful device sizing and their logical blocks have low fan-in, low fan-out. However, it has been difficult to design RAM blocks for these high-speed CMOS circuits. A technique has been developed for a high-speed CMOS RAM structure based on a six-transistor (6-T) static memory cell, which is capable of several hundred million accesses per second [113]. This technique utilizes the following key design approaches:

- A proper pipelining of the address decoder with an option to include precharged logic in the latches.
- Integration of a high-performance sense amplifier in the clocking scheme that includes two types of latches—one that generates new output during a clock low (p-latches), and the second type, during a clock high (n-latches).

This new RAM technique was demonstrated in the design of a 32-word × 64-b (2048 b), three-ported (two read and one write) register file running at 420 MHz, using a standard 1.0 μm process.

2.2.6 Advanced SRAM Architectures and Technologies

2.2.6.1 1–4 Mb SRAM Designs. SRAM speed has usually been enhanced by scaling down the device geometries. The shorter gate channel length, L(eff) translates quite linearly into faster access times. The scaling of the

process also results in higher density memories. Since 50–60% area of a SRAM is the memory array, the problem of reducing the chip size directly relates to reducing the cell area. Figure 2-17(a) shows the cell size as a function of SRAM density [4]. Figure 2-17(b) shows the evolution of MOS SRAMs: the shrinking line widths and a variety of process enhancements which have been primarily responsible for the density and performance improvements [1].

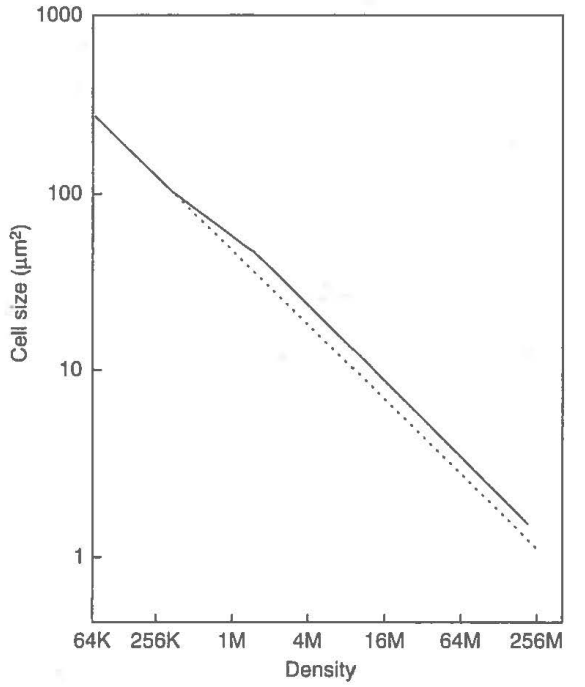


Figure 2-17. (a) SRAM cell size plotted against density [4] (source: various *ISSC Proceedings*, 1976–1990). (b) Evolution of MOS SRAM technology. (From [1], with permission of IEEE.)

The polysilicon load resistors have been commonly used up to 1 Mb SRAM designs; however, their use at 4 Mb and higher densities to achieve both low power and high speed becomes quite challenging. The requirement for a small chip size demands a very small memory cell, which can only be obtained by the scaled polyresistor cell. The standby current requirement sets a minimum value to the load resistance. On the other hand, the resistance must be small enough to provide sufficient current to the storage node so that the charge does not leak away. Another factor in the design of an *R*-load cell is that the power supply of the memory array in the 0.5–0.6 μm technologies has to be less than 5 V, even though the external chip supply voltage is maintained at 5 V.

Several companies continued to optimize the *R*-load cell at the 4 Mb SRAM density level, whereas the others looked for an alternative cell which would reduce the drawbacks of the *R*-load cell but retain its small size. An alternative approach in the 4 Mb generation was the development of a six-transistor cell with four NMOS transistors in the silicon substrate and the two PMOS load transistors formed in the thin-film polysilicon layers above the cell, in the same manner as the resistor loads were stacked in the *R*-load cell. A 1 Mb SRAM with stacked transistors-in-poly cell in 0.8 μm technology was developed by NEC in 1988 [19]. This SRAM employed the p-channel polysilicon resistors as cell load devices with offset gate–drain structure in the second-level poly with the gate below it in the first-level poly. It was stacked on the n-channel driver transistors. This vertical cell offered the size benefit of the *R*-load resistor cell and the noise margin of the six-transistor cell. Subsequently, various process enhancements were made by many companies to reduce the cell size and the standby current.

In 4 Mb and higher density SRAMs with increasing cell density and reduced supply voltage to eliminate hot carrier degradation, it becomes difficult to maintain low standby current without degrading the noise immunity. A

stacked-CMOS SRAM cell with polysilicon thin-film transistor (TFT) load has been developed that leads not only to an increase of the cell static noise margin (SNM), but also to reduction of the cell area [20]. In this cell configuration, the off-current and on-current of the TFT load are the important parameters to reduce standby current and increase noise immunity, respectively. In this approach, an improvement of the electrical characteristics of the TFT load has been achieved by using a special solid phase growth (SPG) technique. The fabrication of this cell is based on three-level polysilicon technology: the first polysilicon layer is used for the gates of the access and driver NMOS transistors and for the *Vss* line, the second polysilicon layer for the gate of the TFT load, and the third polysilicon layer for the active layer of the TFT load and the *Vcc* line. Figure 2-18 shows the stacked-CMOS SRAM cell (a) schematic cross-section, (b) circuit diagram, and (c) layout [20].

Hitachi has developed a 4 Mb CMOS SRAM in 0.5 μm process, four-level poly, two-level metal, with a polysilicon PMOS load memory cell with an area of 17 μm^2 [21]. This had a standby current of 0.2 μA at a supply voltage of 3 V and address access time of 23 ns. The

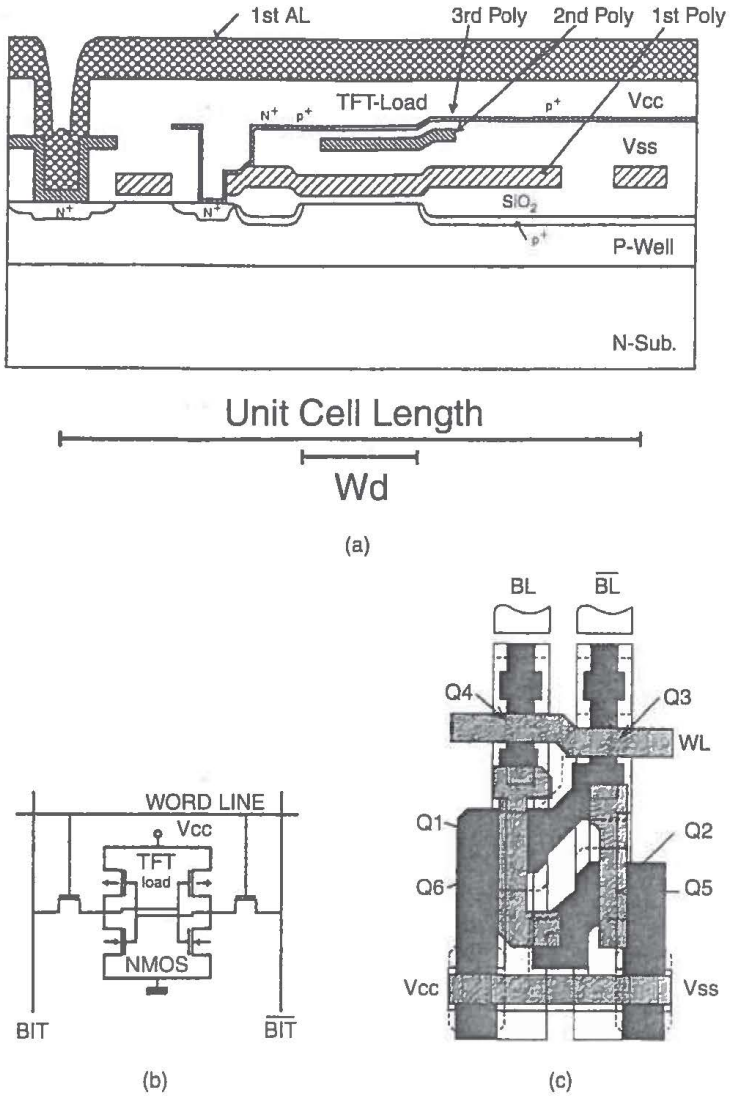
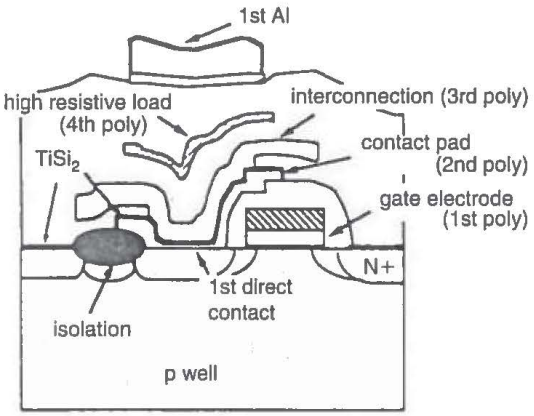


Figure 2-18. A stacked-CMOS SRAM cell. (a) Schematic cross-section. (b) Circuit diagram. (c) Layout. (From [20], with permission of IEEE.)

memory cell has excellent soft-error immunity and low-voltage operating capability. Fast sense amplifier circuits were used to reduce the access time. A noise immune data-latch circuit was used to get a low active current of 5 mA at an operating cycle of 1 μs. A quadruple-array word-decoder architecture was employed in which the word decoders for four memory sub-arrays are laid out in one region. This architecture drastically reduced the word-decoder area, achieving a small chip area of 122 mm^2.

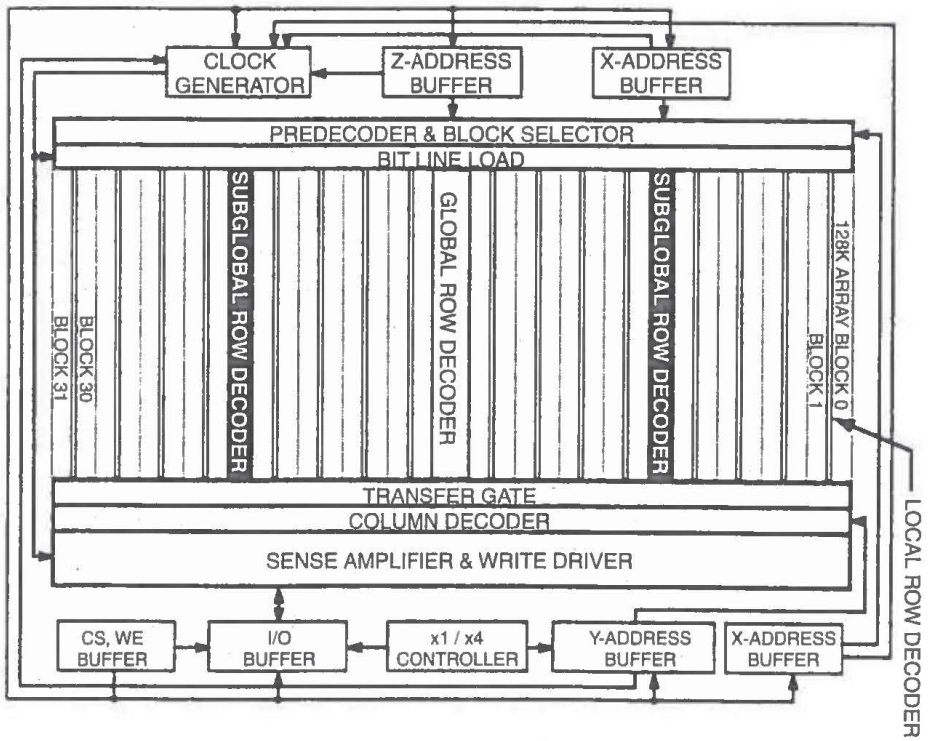
In 1990, Mitsubishi introduced a 4 Mb CMOS SRAM fabricated with 0.6 μm process technology which utilized quadruple-polysilicon (including polycide), double-metal, twin-well CMOS technology [22]. Figure 2-19(a) shows a memory cell cross-section, as well as the process and device parameters. The memory



Process and Device Parameters

TECHNOLOGY	0.6μm CMOS
	N-SUB, TWIN-WELL QUADRUPLE POLY-Si DOUBLE METAL
DESIGN RULE	0.6μm (minimum)
TRANSISTOR	L = 0.6μm (NMOS) 0.8μm (PMOS) Tox = 15nm
MEMORY CELL	HIGH RESISTIVE LOAD R = 10TΩ

(a)



(b)

Figure 2-19. A 4 Mb CMOS SRAM. (a) Memory cell cross-section, process and device parameters. (b) Block diagram. (From [22], with permission of IEEE.)

cell size is 3.5 × 5.3 μm^2. The first polysilicon (polycide) is used for the gate electrode of the transistors. The second polysilicon layer, which is used to form a "contact pad," is electrically isolated from the first polycide and laid over it; this results in cell size compaction. The second polysilicon and the diffusion area are silicided with a self-alignment technique to reduce the sheet resistance. The third polysilicon is used for the interconnection and power supply line of the memory cell. The fourth polysilicon forms the high resistive load of 10 TΩ to achieve a low standby current. The bit lines are formed by the first metal with a barrier metal layer. This SRAM has a 20 ns access time at a supply voltage of 3.3 V, achieved by a new word-decoding scheme and high-speed sense amplifiers. It also has a fast address mode which is suitable for cache memory applications. Figure 2-19(b) shows a block diagram of the SRAM.

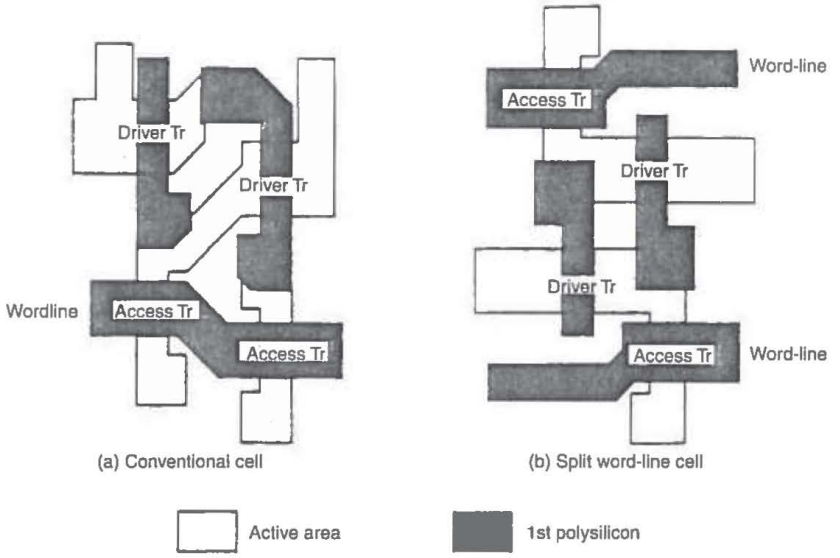
2.2.6.2 16–64 Mb SRAM Development. For 16 Mb and higher density SRAMs, improvements in the MOSFET characteristics alone cannot provide the high performance necessary for advanced processing requirements. As the cell size is reduced through scaling, it becomes more difficult to maintain stable operation at a low supply voltage of 3 V. Therefore, it is necessary to incorporate improved process and circuit techniques and optimization of the chip architecture. Currently, the 16 Mb SRAM is in development stages. In the ISSC Conferences during the last few years, several advanced 16 Mb CMOS SRAM designs and architectural developments have been presented.

In 1992, Fujitsu presented development plans on a 16 Mb CMOS SRAM, organized as 4M words × 4 b [23]. Low power dissipation and access time of 15 ns are obtained by a reduced voltage amplitude data bus connected to a latched cascaded sense amplifier and a current sense amplifier. Double-gate pMOS thin-film-transistor (TFT) load-cell technology reduces the standby current. A "split word-line" memory cell layout pattern, quintuple polysilicon layers, double metal layers, and a 0.4 μm photolithography result in a 2.1 × 4.15 μm^2 memory cell. This SRAM has a 16 b parallel test mode. Figure 2-20(a) shows a memory cell layout, conventional versus split word-line cell. Figure 2-20(b) shows a block diagram of the 16 Mb CMOS SRAM.

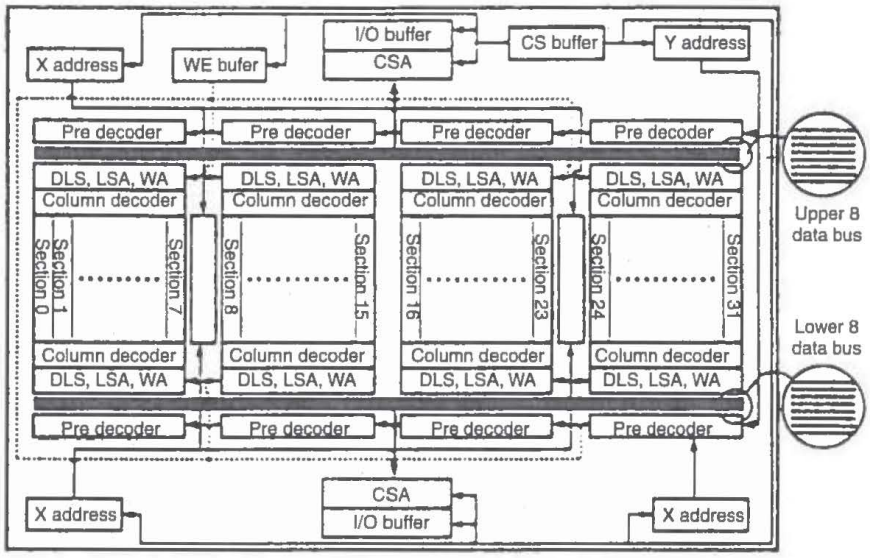
NEC has developed a 3.3 V, 16 Mb CMOS SRAM with an access time of 12 ns and typical standby current of 0.8 μA by using a 0.4 μm quadruple-polysilicon, double-metal technology with thin-film transistor (TFT) load memory cells [24]. The fast access time (12 ns) was achieved through the various design optimization techniques, including:

- An automated transistor size optimization.
- Chip architectural optimization that used center pads to shorten the bit and data-bus delays. Also, the rotated memory cell architecture halves the length of the global word lines. A hierarchical data read-bus structure with a buffer between the subglobal and global read buses helps reduce the data bus delays by 33%. To further reduce the read bus transmission delay, presetting the bus lines to the midlevel called read-bus midlevel preset scheme (RBMIPS) is used.
- This SRAM has three test modes: redundancy row test mode, redundancy column test mode, and 16 b parallel test mode.

A single-bit-line cross-point cell activation (SCPA) architecture has been developed by Mitsubishi to reduce the size of 16 Mb and higher density SRAMs [25]. The supply voltage of this experimental device is 3 V. The column current which constitutes a major portion of active current in the SRAMs refers to the current flowing from the bit line to the storage node of the memory cell through the access transistor when the access transistor is in the on-state. The conventional divided word-line (DWL) structure to reduce the column current requires a large number of block divisions, proportional to the memory density in 16 Mb SRAM chips, whereas the SCPA architecture realizes a smaller column current with a 10% smaller memory core versus the DWL structure. The experimental device with 2.2 × 3.9 μm^2 cell was fabricated using a five-layer polysilicon, sin-



(a)



(b)

Figure 2-20. A 16 Mb CMOS SRAM Fujitsu. (a) Memory cell layout. (b) Block diagram. (From [23], with permission of IEEE.)

gle-aluminum, 0.4 μm CMOS wafer process technology. These are some of the major features of the SCPA architecture:

- A new PMOS precharging boost circuit which is used to boost the word line and activated only during the write cycle.
- A modified data-sensing scheme utilizing a dummy cell since the conventional differential amplifier cannot be used in the SCPA architecture which has a different memory cell configuration.

A 9 ns 16 Mb CMOS SRAM has been developed by Sony Corporation using a 0.35 μm process with quadruple-polysilicon and double-aluminum technology [26]. This design utilizes a current-mode, fully nonequalized data path by using a stabilized feedback current sense amplifier (SFCA) that provides a small input resistance

and an offset compensating effect. The chip is organized as 4M words × 4 b, and the memory cell array is divided into 4 Mb quadrants with 1 b of the 4 b data word residing in each quadrant. Also, each quadrant is further divided into 32 sections of 1024 rows × 128 columns. By placing the sense amplifiers adjacent to the I/O pins and dedicating each quadrant to one data bit, the *RC* delay of the read data path is reduced. It uses a line mode parallel test circuit which is a current mode EXCLUSIVE-OR gate consisting of two SFCAs with reference current equal to one-half of the memory cell current. A parallel access occurs in write and read cycles. As a result, the 16 Mb SRAM is compressed to 256 kb memory in the test mode.

A novel architecture has been developed for 16 Mb CMOS SRAM by Hitachi that enables fast write/read in poly-PMOS load or high-resistance polyload, single-bit-line cell [27]. The write operation involves alternate twin word activation (ATWA) and pulsing of a single bit line. A dummy cell is used to obtain a reference voltage for reading. To obtain the required balance between a normal cell signal line and a dummy cell signal line, a balanced common data-line architecture is used. A novel self-bias control (SBC) amplifier provides excellent stability and fast sensing performance for input voltages close to V_{cc} at a low power supply of 2.5 V. This single-bit-line architecture, used to fabricate a 16 Mb SRAM in 0.25 μm CMOS technology, reduced the cell area to 2.3 μm^2, which is two-thirds of a conventional two-bit-line cell with the same process.

In the last few years, in addition to process enhancements in the planar transistors, several advanced structures have been developed. In 1988, Toshiba proposed the "surrounding gate transistor" (SGT) which had a gate electrode surrounded by a pillar silicon island and reduced the occupied area of the CMOS inverter by 50% [28]. The source, gate, and drain were arranged vertically using the sidewalls of the pillar silicon island as the channel region. The pillar sidewalls were surrounded by the gate electrode, and the gate length was adjusted by the pillar height. This SGT structure potentially reduces serious problems from both the short channel effects (gate length) and the narrow channel effects (gate width) of the planar transistors with submicron geometries.

Some other possibilities for the 64 Mb generations and beyond include combinations of the vertical and stacked transistor and cell technologies. In 1989, Texas Instruments proposed an innovative six-transistor cell with a thin-film polysilicon PMOS transistor and a vertical sidewall trench NMOS transistor in 0.5 μm technology, which resulted in a 23 μm^2 cell area [29]. Since the thin-film transistor may be stacked over the vertical NMOS, this cell has the potential for future scaling. Another interesting combination structure investigated for 0.2 μm technologies was the "delta transistor" proposed by Hitachi in 1989 [30]. This combines the vertical pillar technology with the fully depleted SOI technology using only bulk silicon process techniques. In this process, a pillar of silicon was formed using the standard trench etching technique, and silicon nitride was deposited as the gate dielectric over it. Then beneath the pillars, a selective oxidation of the bulk silicon isolates the silicon of the pillar and forms an SOI vertical transistor above the oxide. This process eliminated the need for isolation, permitting the larger gate length and width for the vertical transistor. These vertical stacking techniques have a potential for future 256 Mb and higher density SRAMs.

2.2.6.3 Gallium Arsenide (GaAs) SRAMs. Some experimental gallium arsenide (GaAs) SRAMs have been developed for very high-speed applications that have 2–3 ns cycle time and densities of 1–16 kb. However, there are a number of technological and processing problems with this technology. The normal processing is more difficult for a GaAs substrate because, unlike SiO_2, As_2O_3 is conductive, and Ga and As tend to separate above 650 °C. The cost of GaAs wafers is also higher than the silicon wafers, the wafer size is smaller, and yield is low. Also, the voltages that could be supported in a sub-5 ns GaAs SRAM are much lower than the standard interface circuitry, and

thus special design considerations are required to interface the RAM to an external ECL system environment.

A schematic cross-section of a GaAs memory cell is shown in Figure 2-21 [31]. This is fabricated with 1.0 μm tungsten self-aligned gate field-effect transistors using double-level metal interconnect technology. The ohmic metal is AuGe–NiAu and the interconnect metals are Ti–Au. The dielectric insulator is phosphorous-doped CVD deposited Si_3N_4. An ion implant was used to form the source and the drain regions.

The process improvements underway in the GaAs technology including 0.3–0.5 μm gates may permit the development of 64 kb SRAMs with access times near 1 ns. These ultra-high-speed devices can be used in cache applications for the next generation supercomputers and supercomputing workstations.

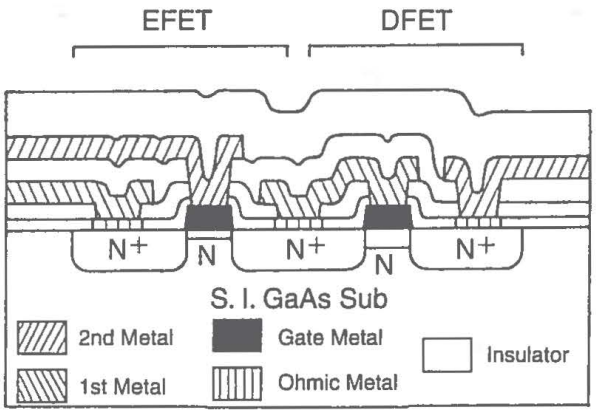


Figure 2-21. Schematic cross-section of a GaAs memory cell. (From [31], with permission of IEEE.)

2.2.7 Application-Specific SRAMs

Application-specific SRAMs include some extra logic circuitry added to make them compatible for a specific task. There is no clear-cut distinction between the application-specific SRAMs and the SRAM arrays embedded in the logic. Usually, the application-specific SRAMs are made in the high-density, optimized processes which include customized features such as the buried contacts and straps to reduce the memory cell size. A few application-specific SRAMs are described in the following subsections.

2.2.7.1 Serially Accessed Memory (Line Buffers). The first-in first-out (FIFO) buffer is an example of the shift register memory architecture through which the data are transferred in and out serially. The serial or sequential architecture requires that the memory cell containing the information be arranged so as to receive and retrieve one bit at a time. The data stored in each cell are retrieved strictly in accordance with their position in the sequence. The data may be retrieved on a first-in first-out (FIFO), first-in last-out (FILO), or last-in first-out (LIFO) basis. The FIFOs are widely used in telecommunication types of applications. These are generally made using the SRAM cells if there is a requirement for the data to be maintained in the FIFO, and from the DRAM cells if the data simply pass through the FIFO. The data are transferred from cell to cell by the sequential operation of the clock.

Many FIFOs currently available are RAM-based designs with self-incrementing internal read and write pointers, which results in very low fall-through times compared to the older shift-register-based designs. The fall-through time of a FIFO is the time elapsing between the end of the first write to the FIFO and the time the first read may begin. Also, the serial registers are not shift registers, but bit wide memory arrays with self-incrementing pointers. These FIFOs can be expanded in any depth to any level by cascading the multiple devices. The FIFOs can range in depth from 64 × 4 b and 64 × 5 b to 16K × 9 b and higher devices.

Figure 2-22 shows the functional block diagram of an IDT 2K × 9 b parallel CMOS FIFO [24]. It is a dual-port memory that loads and empties data on a first-in first-out basis. This device uses Full and Empty flags to prevent the data overflow and underflow, and expansion logic to allow for unlimited expansion capability in both the word size and depth. The reads and writes are internally sequential through the use of ring pointers, with no address

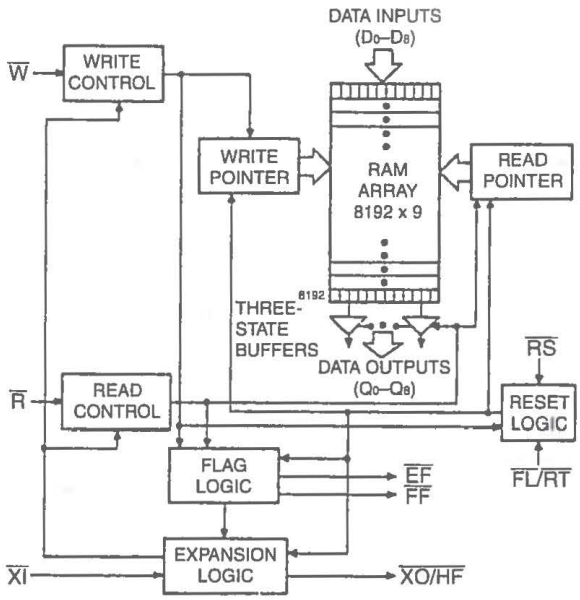


Figure 2-22. Functional block diagram of IDT7205 FIFO (2K × 9 b). (From [32], with permission of Integrated Device Technology, Inc.)

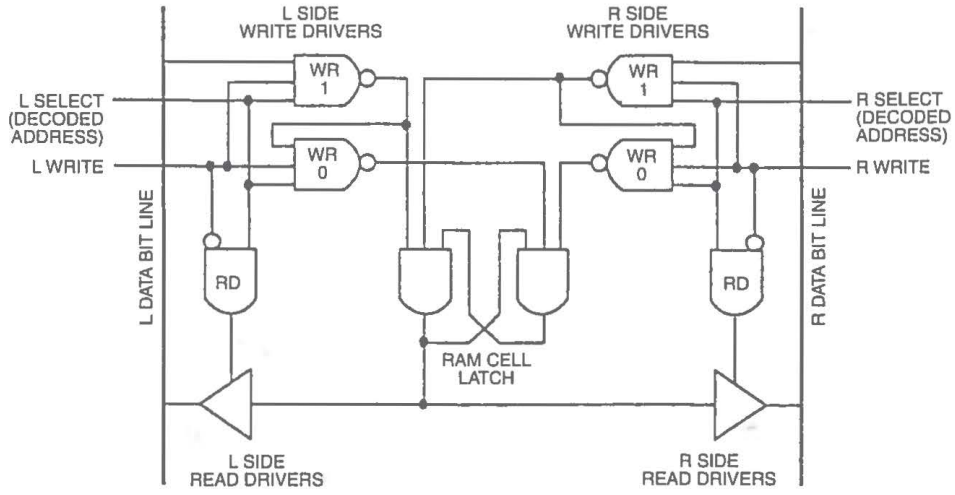
information required to load and unload the data. It is toggled in and out of the device using Write (*W*) and Read (*R*) pins. It utilizes a 9 b wide data array to allow for control and parity bits at the user's option. It also features a retransmit (*RT*) capability that allows for the reset of the read pointer to its initial position. A Half-Full flag is available in the single device and width expansion modes. The device has a read/write cycle time of 30 ns (33 MHz).

2.2.7.2 Dual-Port RAMs. The dual-port RAMS allow two independent devices to have simultaneous read and write access to the same memory. This allows the two devices to communicate with each other by passing data through the common memory. These devices might be a CPU and a disk controller or two CPUs working on different but related tasks. The dual-port memory approach is useful since it allows the same memory to be used for both working storage and communication by both devices without using any special data communication hardware. A dual-port memory has sets of addresses, data, and read/write control signals, each of which accesses the same set of memory cells. Each set of memory controls can independently and simultaneously access any word in the memory, including the case where both sides are accessing the same location at the same time.

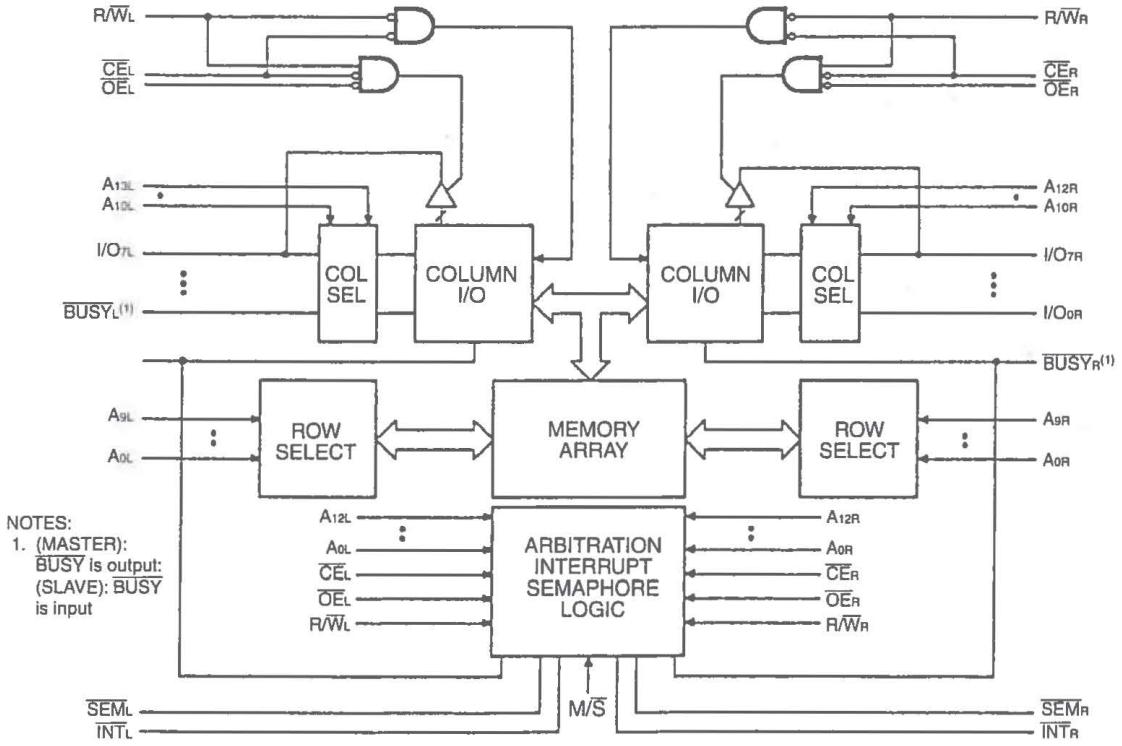
Figure 2-23(a) shows an example of a dual-port memory cell [32]. In this cell, both the left- and right-hand select lines can independently and simultaneously select the cell for a readout. In addition, either side can write data into the cell independently of the other side. The problem occurs when both sides try to write into the same cell at the same time, and the result can be a random combination of both data words rather than the data word from one side. This problem is solved by the hardware arbitration Busy Logic design by detecting when both sides are using the same location at the same time, and then causing one side to wait until the other side is done. It consists of a common address detection logic and a cross-coupled arbitration latch. A common problem in dual-port RAMs is the need to temporarily assign a block of memory to one side. Some devices currently available have "semaphoric logic," which provides a set of flags specially designed for the block assignment function. This ensures that only one side has permission to use that block of memory.

Figure 2-23(b) shows the functional block diagram of an IDT 8K × 8 dual-port SRAM designed for use as a standalone 64 kb dual-port RAM or as a combination MASTER/SLAVE for 16 b or more word systems [32]. This device provides two independent ports with separate control, address, and I/O pins, which permit asynchronous access for reads or writes to any location in the memory. These are rated for a maximum access time of 35 ns.

A family of modular multiport SRAMs with a built-in self-test (BIST) interface has been developed using a synchronous self-timed architecture. The basic port design is self-contained and extensible to any number of ports sharing access to a common-core cell array. The same design was used to implement modular one-port, two-port, and four-port SRAMs and a one-port DRAM based on a four-transistor (4-T)



(a)



(b)

Figure 2-23. (a) A dual-port RAM cell. (b) Functional block diagram of IDT7005 dual-port SRAM. (From [32], with permission of Integrated Device Technology, Inc.)

cell. Figure 2-24 shows the schematic of a four-port SRAM core array as an example [33]. One access transistor pair is added per core cell to provide fully independent read/write capability for each port. This provides symmetry in design, layout, simulation, and BIST circuitry. Each port functions independently with its own control block and model signal path. Also, each core cell is provided with adequate margin for stability to permit simultaneous reads from the

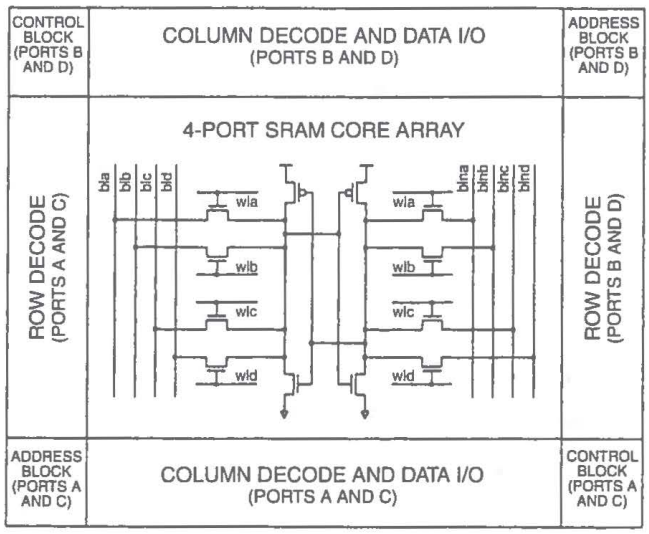


Figure 2-24. Schematic of a four-port SRAM core array. (From [34], with permission of IEEE.)

same address on all ports. No simultaneous writes to the same address location are permitted.

The layout architecture of the generic port periphery places the address block in one corner, with the row and column decoders extending from this block along the adjacent sides of core cell array. The data I/O circuitry is integrated into the same layout block as the column decoder. The timing control circuitry is located at the far end of the column decode block. The ports are mirrored around the cell array to achieve multiport configuration. This design was implemented in a 0.8 μm BiCMOS process. A system level performance of 100 MHz has been verified on a memory management chip from a broad-band switch core containing a 12K × 8 two-port SRAM subdivided into four 3K × 8 blocks.

2.2.7.3 Nonvolatile SRAMs The battery-operated power systems require not only very low operating power and battery-backup data retention current levels from the memory, but also logic circuitry which will protect the stored data if the battery voltage gets low. This logic circuitry also contains protection against accidental writes during power-down. In the early 1980s, Mostek (later SGS-Thomson Microelectronics) introduced byte-wide SRAMs called the "zero power RAMs," which included a lithium battery. They also included a temperature-compensated power fail detection circuitry to monitor the V_{dd}, and switched from the line voltage to back-up battery when the voltage drops below a specified level, and then switched back to the line voltage when the power is restored.

Figure 2-25 shows the functional block diagram of an IDT 71256 (32K × 8 b) available in a low-power (*L*) version which also offers a battery backup data retention capability [32]. The circuit typically consumes only 5 μW when operating from a 2 V battery. When $\overline{CS}$ goes high, the circuit automatically goes into a low-power standby mode and remains there as long as $\overline{CS}$ remains high. In the full standby mode, the device typically consumes less than 15 μW.

2.2.7.4 Content-Addressable Memories (CAMs). The content-addressable memories (CAMs) are designed and used both as embedded modules on the larger VLSI chips and as a standalone memory for specific systems applications. Unlike the standard memories which associate data with an address, the CAM associates an address with data. The data are presented to the inputs of CAM which searches for a match to those data in the CAM, regardless of the address. When a match is found, the CAM

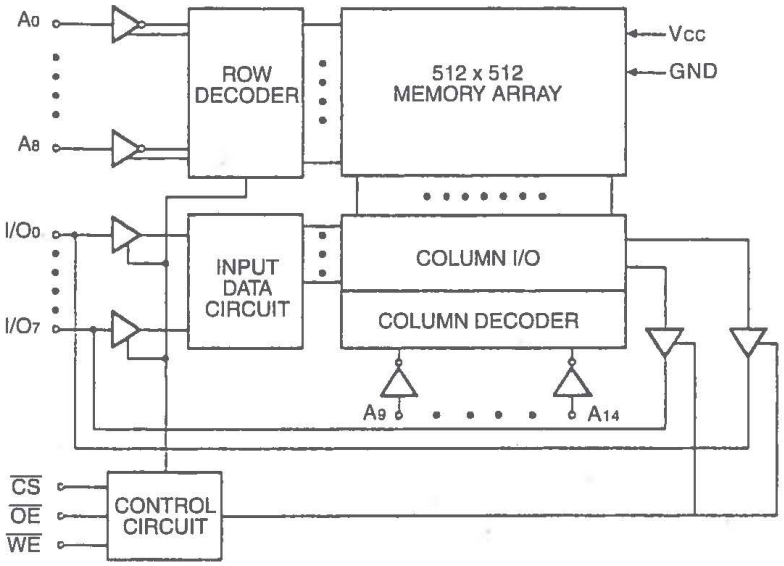


Figure 2-25. Functional block diagram of IDT71256 (32K × 8 b). (From [32], with permission of Integrated Device Technology, Inc.)

identifies the address location of the data. The applications using CAMs include database management, disk caching, pattern and image recognition, and artificial intelligence.

Figure 2-26 shows the block diagram of Advanced MicroDevices AM99C10A, a CMOS CAM with a capacity of 256 words (each consisting of a 48 b comparator and a 48 b register)

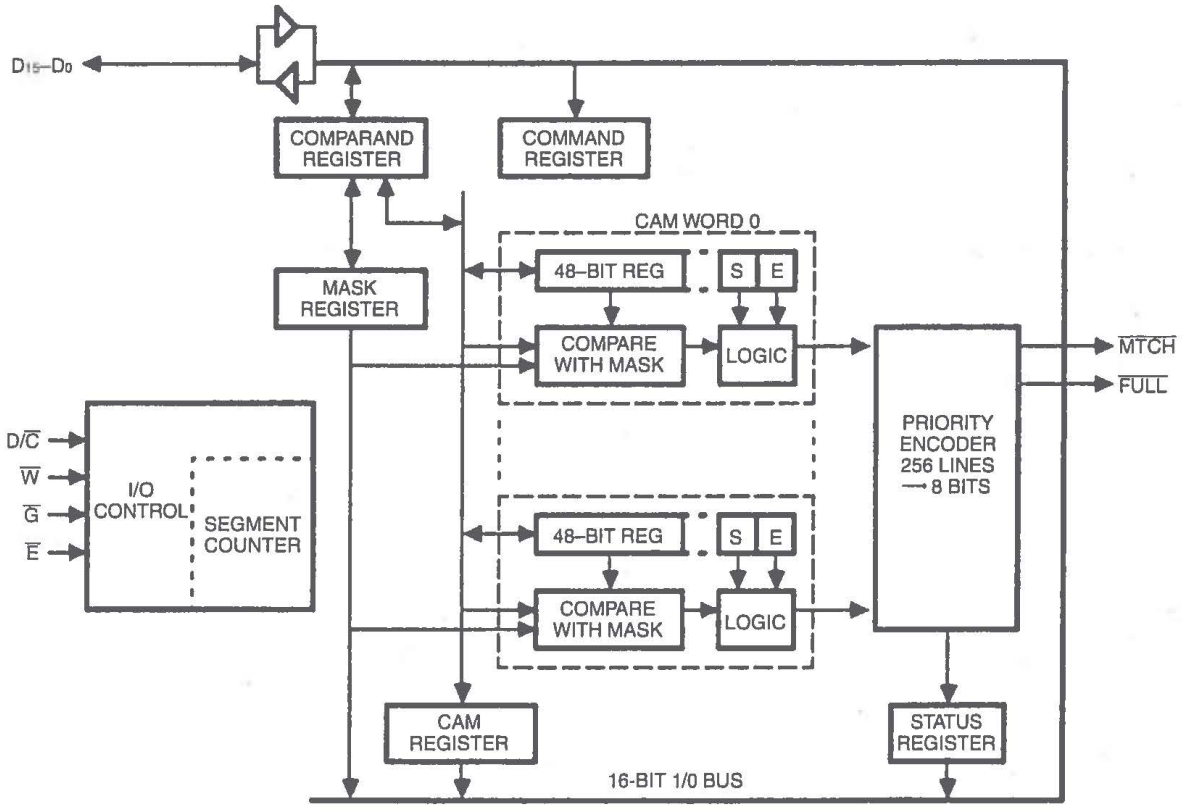


Figure 2-26. A block diagram of Advanced MicroDevices AM99C10A CMOS CAM. (From [35], with permission of Advanced MicroDevices, Inc.)

and a user-programmable word width of 16 or 48 b [35].

2.3 DYNAMIC RANDOM ACCESS MEMORIES (DRAMs)

Dynamic random access memories use charge storage on a capacitor to represent binary data values. These are called dynamic because the stored charge leaks away, even with power continuously applied. Therefore, the cells must be read and refreshed at periodic intervals. Despite this apparently complex operating mode, the advantages of cost per bit and high density have made DRAMs the most widely used semiconductor memories in commercial applications. In the early 1970s, the DRAM cell designs ranging from a six-transistor to a one-transistor configuration were proposed. In 1973, the one-transistor (1-T) cell became standard for the 4K DRAMs [36]. In the following years, the density of DRAMs increased exponentially with rapid improvement in the cell design, its supporting circuit technologies, and photolithographic techniques.

The key circuit technologies which have contributed to the progress in DRAM development are advanced vertical cell structures with trench and stacked capacitors; improvements in differential sensing and folded data-line arrangement for noise reduction; and the use of dynamic amplifiers and drivers, CMOS circuits, half V_{dd} data-line precharge, a shared I/O combined with multidivided data lines for reduced power dissipation. These technology developments are discussed in more detail in the following sections.

2.3.1 DRAM Technology Development

The first commercial DRAM, a 1 kb chip using three transistor cells in p-channel silicon gate technology, was announced by Intel in 1970. This was followed by 4 kb DRAMs in a single-polysilicon, single-aluminum metallization in typically 10 μm feature size. The sense amplifier was a simple structure with two static inverters and two switches with a bootstrapped output stage which resulted in fast charging of the load capacitances. It had nonmultiplexed addressing. A shift register generated the three regulated timing signals needed to operate the sense amplifier from a single external clock. Subsequently, the first address-multiplexed DRAMs were introduced, also at the 4 kb density level. With time-multiplexed addressing, these 4 kb DRAMs used six address pins per device to select one of the 64 rows, followed by one of 64 columns to select a particular bit instead of using 12 address pins to access one specific location. The six multiplexed addresses were clocked into the row and column decoders by the row address strobe ($\overline{RAS}$) and the column address strobe ($\overline{CAS}$). Input and output latches were also used, and the data to be written into a selected storage cell were first stored in an on-chip latch. This simplified the memory system design.

The first generation DRAMs were manufactured using a three-transistor (3-T) cell as shown in Figure 2-27(a) [37]. The transistors M_2, M_3, and M_4 are made small to minimize the cell area. The charge stored on C_1 represents stored binary data. The selection lines for reading and writing must be separated because the stored charge on C_1 would be lost if M_3 were turned on during reading. The cell operates in two-phase cycles. The first half of each read or write cycle is devoted to a precharge phase during which the columns D_{in} and D_{out} are charged to a valid high logic level through M_{y1} and M_{y2}. A "1" is assumed to represent a high level stored on C_1, and is written by turning on M_3 after D_{in} is high. The D_{in} line is highly capacitive since it is connected to many cells. The read–write circuits do not need to hold D_{in} high since sharing the charge on D_{in} with C_1 does not significantly reduce the precharged high level. A "0" is written by turning on M_3 after the precharge phase is over, then simultaneously discharging D_{in} and C_1 via a grounded-source pull-down device (not shown) in the read–write circuit.

The read operation is performed by turning on M_4 after the precharge is over. If a "1" is stored, D_{out} will be discharged through M_2 and M_4. If a "0" is stored, there will be no conducting