## 5.5 Basics of ROM

### 5.5.1 Chip architecture

The fundamental architecture of a simple ROM is illustrated in Fig. 5.15 [14, 2]. The row and column address decoders are incorporated to select one out of $2^n$ words by decoding the $n$-bit address. The conditioning circuit allows precharging of the bit-lines and the information is retrieved or read from the bit-lines by using the sense amplifiers. In most cases, column decoders are utilized to select a set of desired data from the bit-lines prior to the sensing operation [14]. Subsequently, when the data output are ready for retrieval, the output enable signal will activate

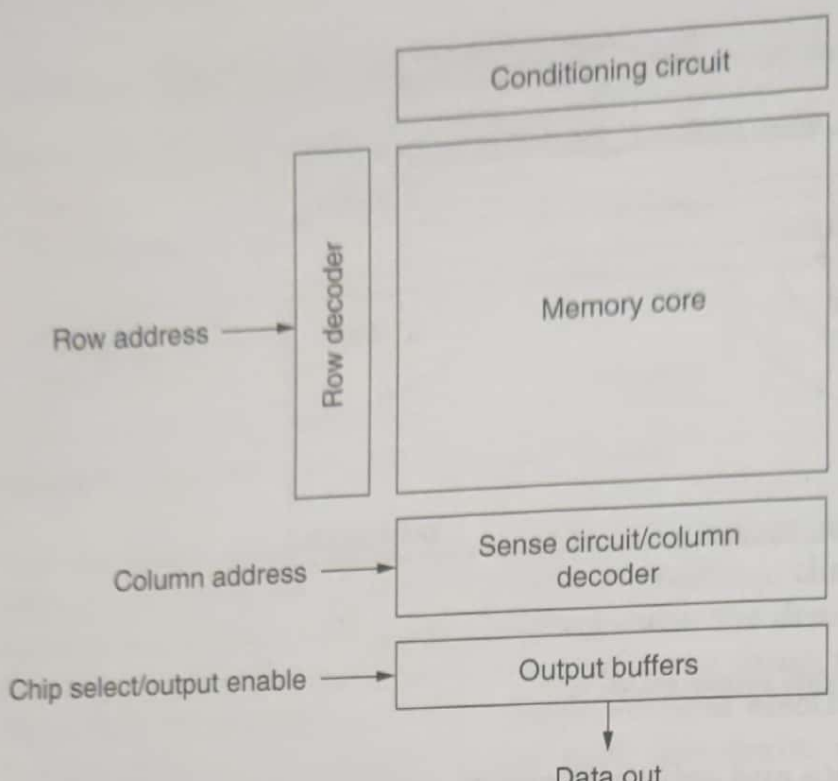



**Figure 5.15** Basic ROM architecture.

the output buffers and transport the data to appear at the outputs. The chip select/output enable signal is often included for memory extension, which augments the number of words and the word size (number of output bits) [11].

A typical example of a 1 kbit ROM organized as 256 words of 4 bits each is shown in Fig. 5.16 [2]. With an 8-bit address, the row address and column address are chosen as 5-bit and 3-bit, respectively. For this reason, there are 32 word-lines and a core of memory cells with a 32×32 matrix is established. The row decoder will select one of the 32 rows, while the group of four 8-to-1 column decoders will serve as data selector to choose four of the 32 columns. Together with the chip select signal, which feeds the output buffers, a 4-bit output, namely D0–D3, is obtained from the selected word-line.

### 5.5.2 ROM cell arrays

A bit is stored in a ROM in accordance to the presence or absence of a data path from a particular row to a corresponding column. The existence of a path is easily realized by connecting a nonlinear element to adjoin the row and column [15]. The nonlinear device, which can be a transistor or a diode, is employed so that the data flows only in the determined direction from the input to the output. The two basic struc-

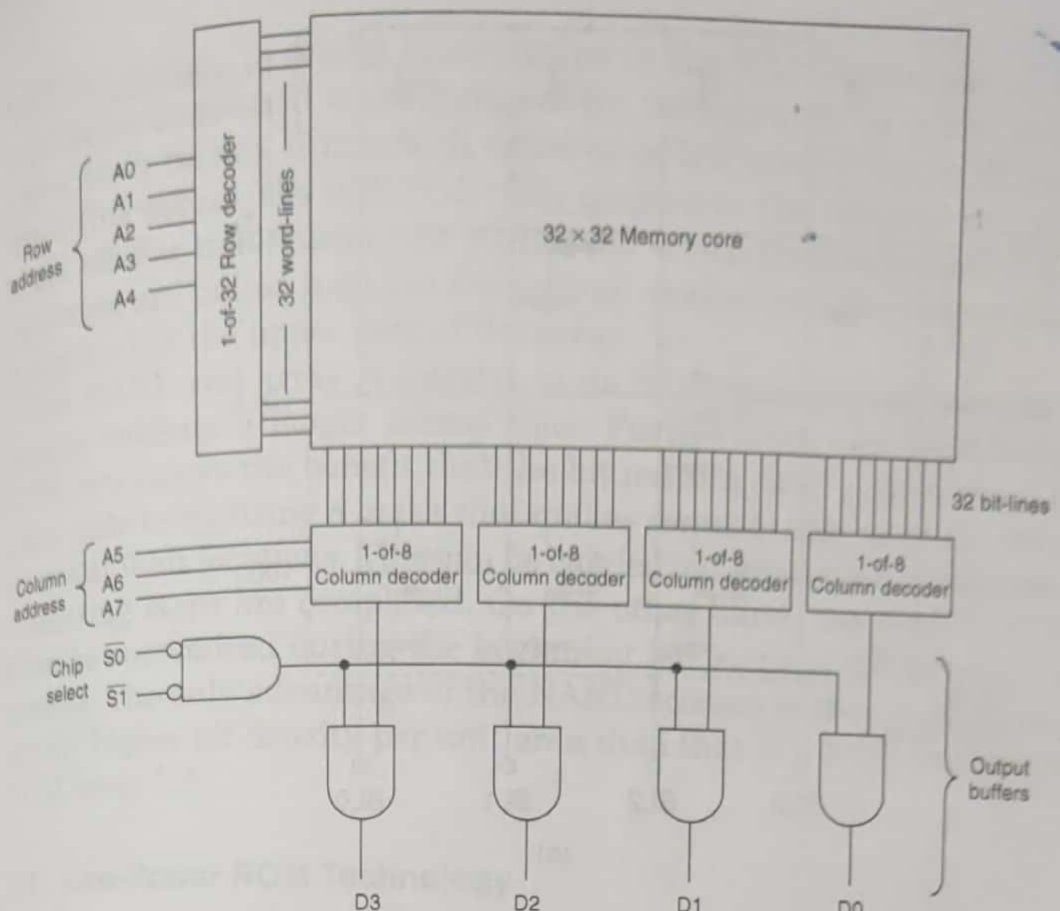
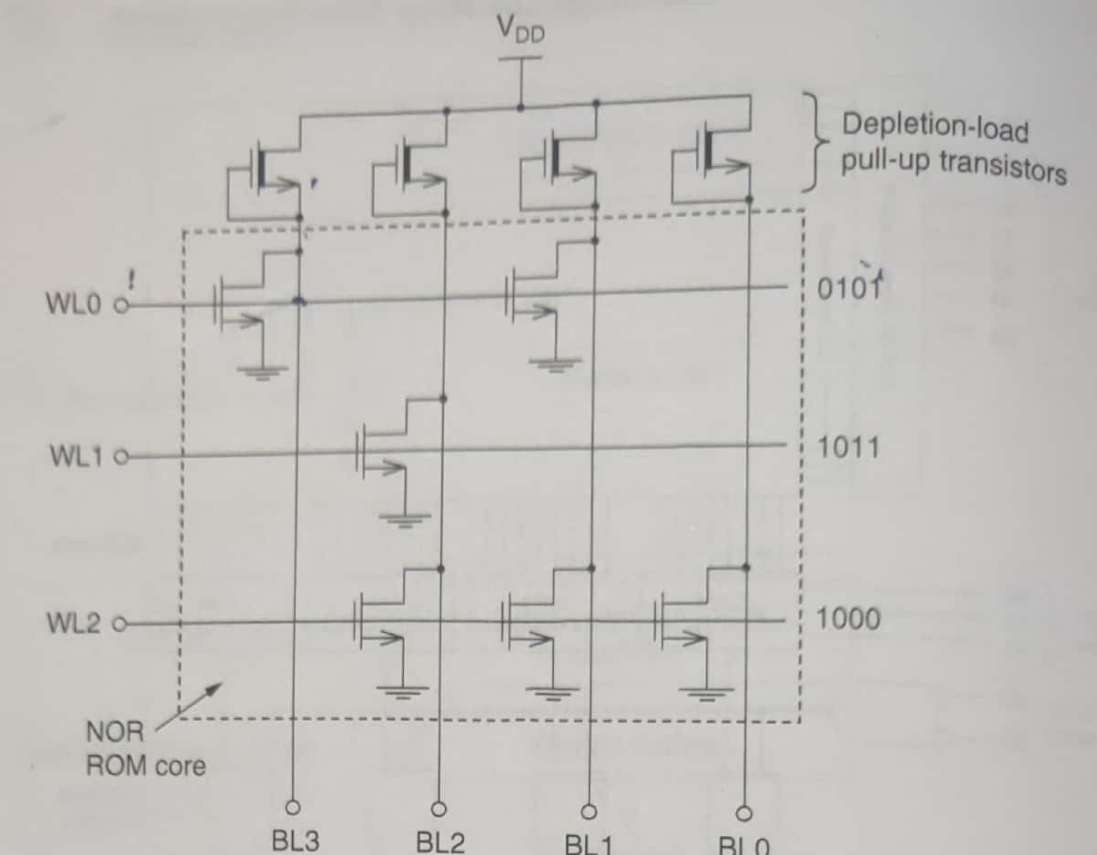



Figure 5.16 A 1024-bit ROM architecture.

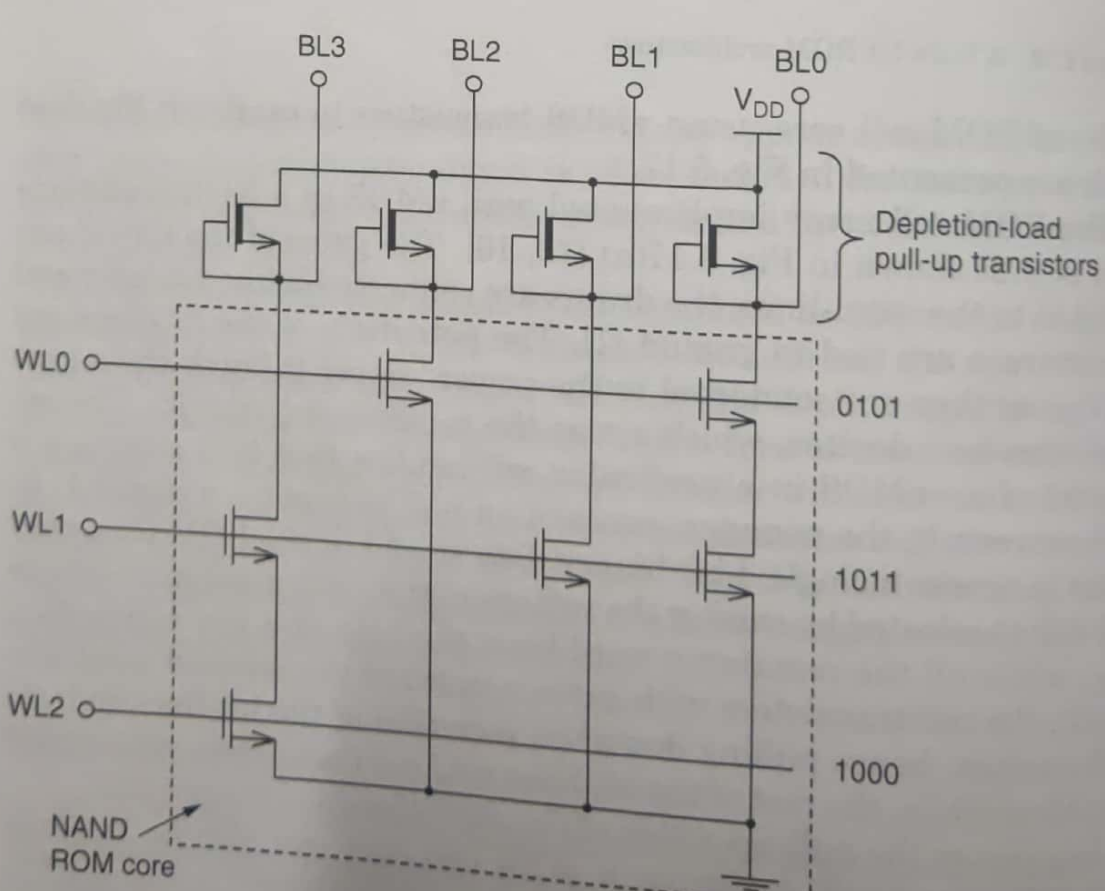
tures of ROM cell core using nMOS transistors to establish the data path are presented in Fig. 5.17.

The ROM cell array implemented with nMOS in a NOR configuration is first shown in Fig. 5.17($a$) [15, 16]. The gates of the nMOS are coupled to the word-lines, the drains are connected to the bit-lines and the sources are tied to ground [1]. The potentials of the bit-lines are at $V_{DD}$ as they are connected to the power supply through the $n$-type depletion-load devices, which act as the conditioning circuits. The existence of an nMOS in a particular cell implies that it is storing a 0 and conversely, the nonappearance of an nMOS denotes a stored 1. In order to access a single 4-bit binary data stored in the ROM core, only one row is selected by raising the voltage of the equivalent word-line to $V_{DD}$, while all the remaining word-lines are held at a low value. As a result, the cell transistors with gates coupled to the selected word-line will conduct, hence pulling down the potential of the bit-lines to logic low. Meanwhile, the rest of the bit-lines without transistors will be held at $V_{DD}$ due to the pull-up action of the depletion-load transistors.

Besides the NOR structure, a ROM cell core can also be realized using an nMOS NAND gate configuration, as illustrated in Fig. 5.17($b$) [15, 16]. It is often referred to as a NAND ROM in the sense that it necessitates all series bit locations to provide a conducting path toward ground so that a specific bit-line output can go low. In contrast to the




(a)




(b)

**Figure 5.17** The ROM cell array based on (*a*) NOR configuration. (*b*) NAND configuration.

NOR cell design, to read a binary value in the NAND core, one of the word-lines is selected by bringing down its potential to a low value rather than raising it to a high value as in the case of the NOR core. When this occurs, the nMOS devices coupled to the selected row will be cut off. For this reason, the bit-lines to which these transistors are connected are pulled high via the pull-up devices and the output data is acquired at the upper part of the array.

The NAND cell array is inferior to its NOR counterpart because it usually exhibits a longer access time. Furthermore, the NOR-based ROM core enjoys the benefit that the bit pattern can be customized at a fast pace by utilizing a mask that makes contacts only with the transistors at 0-bit locations. This can be carried out after most of the manufacturing steps are completed. On the other hand, the NAND ROM must be customized during the beginning few steps of the fabrication process. The only advantage of the NAND scheme is that it offers relatively higher bit density per unit area than that of a NOR configured ROM array [15].

## 5.6 Low-Power ROM Technology

In order to fulfill the present-day escalating market trend encompassing low-voltage low-power applications, many low-power design methodologies have been proposed [14, 17, 18, 19]. This is in accordance and in line with the advancement of submicron technologies and the emergence of highly sophisticated VLSI chips with increased complexity in handheld gadgetry [20]. The high area density ROM is broadly used in numerous digital systems, such as digital filters, digital signal processors, and microprocessors to store fixed rigid data. In view of its wide coverage in today's upsurging portable equipments, the power consumption of ROM should be diminished as much as possible to ensure performance leverage. This section reports on several low-power techniques at the circuit and architecture levels to realize significant power dissipation reduction in the ROM structures [18].

### 5.6.1 Sources of power dissipation

The architecture of ROM in its conventional form is depicted in Fig. 5.18. The ROM core is the memory bank, which consists of an array of transistors arranged to store bits of information. As discussed previously, two major types of ROM are the NAND array and the NOR array, where the pull-down transistors are connected in series and in parallel, respectively. The decoder selects one word-line in the ROM core at a given time based on the corresponding input address. The column multiplexer and driver assert a particular column to be read and drive the data bus. Meanwhile, the control logic block generates the internal signals of the

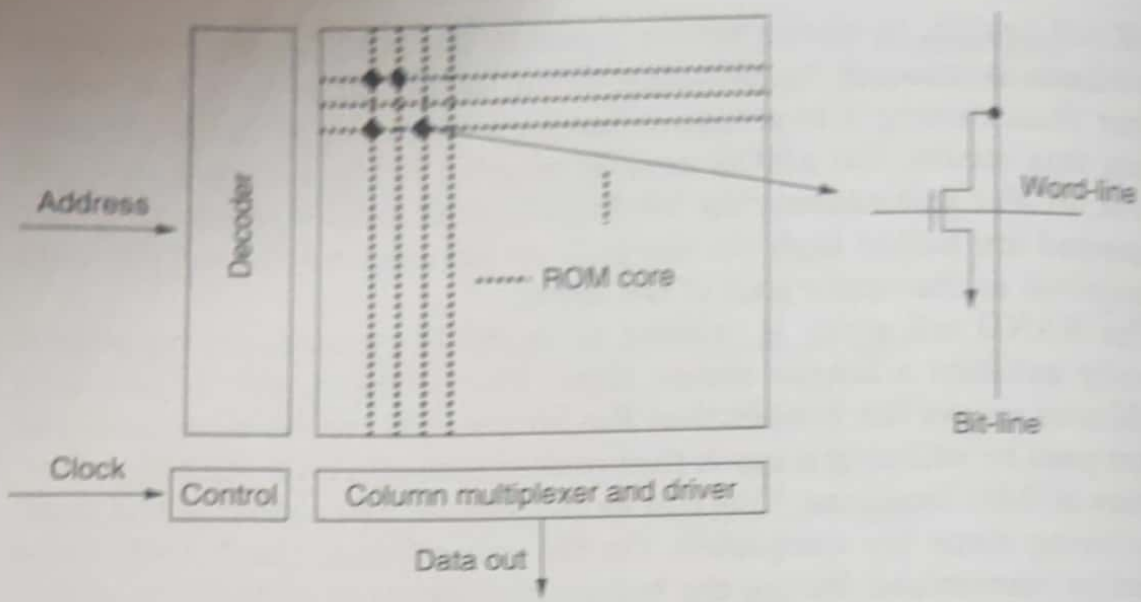



**Figure 5.18** Conventional architecture of ROM.

ROM such as precharge and read. The ROM core using a NOR array is reported herein since this configuration offers higher speed than its NAND counterpart and is also the most extensively used.

In a memory system, most of the power dissipation stems from the highly capacitive elements, for instance the predecoder lines, word-lines and bit-lines. A 2 K × 18 ROM has been designed based on a 0.6 μm technology at a power supply voltage of 3.3 V and a clock frequency of 10 MHz [18]. Its sources of power consumption are tabulated in Table 5.1. Although the technology used is somewhat old, it can still serve the purpose of illustrating the percentage of power dissipated in each major block of ROM. As shown in the table, the precharge of the bit-lines in the ROM core contributes the largest slice of the overall power loss. This is due to the high bit-line capacitance with the drain of the transistors coupled to this line. Also, more than 18 bit-lines are switched per access. This is because the word-line selects more bit-lines than is necessary. In this example, a multiplexer of 12 to 1 ratio is employed. As a result, at least an additional four bit-lines will switch instead of one.

**TABLE 5.1 Power Dissipation of 2 K × 18 ROM**

| Block | Power (mW) | Percentage (%) |
|---|---|---|
| Decoder | 0.06 | 2.1 |
| ROM Core | 2.24 | 89 |
| Control | 0.18 | 7.2 |
| Drivers | 0.05 | 1.7 |

The power dissipated in the control logic is quite substantial because it generates the precharge signal for the ROM core and also enables the output drivers and decoding logic. The power consumption in the decoder is a small amount since only one word-line is asserted per access.

### 5.6.2 Low-power techniques at architecture level

Since the switching activity of the bit-lines occupies the largest proportion of power dissipation in a ROM architecture, a range of different low-power schemes will be dedicated to the ROM core to enhance its performance.

**Divided hierarchical word-line structure.** The divided word-line technique has been introduced for implementation in SRAMs [21]. This method will be briefly explained here and will then be illustrated in detail in Sec. 6.9.2. Principally, the basic idea of this concept is to distribute the memory core into various subblocks and then run the subblock word-line and global word-line in different layers (i.e., metal or poly). Therefore, only the bit cells of the specific selected block are accessed through the address bits. The same design idea can be applied to ROMs. Although this technique can achieve a considerable amount of power savings, it does not solve the problem entirely. This is because due to layout considerations, a ratio of at least 4 to 1 is still required in the multiplexer. This implies that more than one bit-line per bit could still be switching.

**Selective precharge.** During the first part of the cycle, all the bit-lines are being precharged to a high voltage level. As a result, a large amount of capacitance is being switched per cycle because many bit-lines are discharged even when these locations are not accessed. This will lead to excessive power dissipation. By making use of the selective precharge method, only bit-lines, which will be accessed, are precharged [22]. This technique is cost-effective because its hardware overhead is low due to the fact that most of the control logic used here is the same control logic feeding the multiplexers.

**Difference encoding.** For some implementations, such as digital filters, ROM is operated in a sequence. If the values between contiguous data do not change radically between one address and the next, the ROM core can be designed in such a way that only the difference between the data is stored rather than the entire value [23]. This technique is called difference encoding and it may be used to minimize the entire size of the ROM core. Nevertheless, a main drawback is having to introduce an adder to calculate the original value. An adaptation of the same

concept is to hard wire different constants (i.e., offsets) and store only the difference relative to the constant.

### 5.6.3 Low-power techniques at circuit level

Besides reducing power consumption at the architectural level, another effective means to realize power saving in VLSI systems is to exploit design techniques at the circuit level.

**nMOS precharge.** It is well known that voltage swing limitation is a prevailing technique used to reduce power dissipation at the bit-lines. This can be accomplished by using nMOS transistors to precharge the bit-lines within the ROM core to a high value. As a result, the bit-lines are precharged to $V_{DD} - V_{\text{th}}$, where $V_{\text{th}}$ denotes the threshold voltage. The voltage swing will be reduced since the bit-lines switch only between $V_{DD} - V_{\text{th}}$ and ground and thus significant amounts of power savings can be achieved. However, this technique deteriorates the noise margin and body biasing effect (which increases the threshold voltage), and therefore demands careful design of the output drivers.

**Voltage scaling.** Voltage scaling is one of the most eminent and prominent approaches to power dissipation reduction. A quadratic enhancement can be readily attained via voltage scaling. Moreover, the short circuit current dissipation can be immensely reduced through scaling of the power supply voltage. High-precision expressions estimating short circuit currents have been reported [24]. Nonetheless, this improvement in power consumption is achieved with degradation in the speed performance of the circuits. A first-order derivation representing the delay of CMOS gates can be expressed as [20]:

$$T_{\text{delay}} = \frac{C_L V_{DD}}{I} = \frac{2C_L V_{DD}}{\mu C_{\text{ox}} \left(\frac{W}{L}\right) (V_{DD} - V_{\text{th}})^2} \tag{5.1}$$

where $C_L$ denotes the load capacitance and $V_{DD}$ is the power supply voltage. Meanwhile, $\mu$ represents the electron mobility, $C_{\text{ox}}$ is the oxide capacitance, $V_{\text{th}}$ is the threshold voltage, and $W$ and $L$ imply the channel width and length, respectively. When the $V_{DD}$ goes down, approaching the threshold voltage of a particular device, the delay will be drastically increased due to the quadratic effect and this will then lead to a severe speed deterioration problem.

**On-chip high-voltage generator.** The prime disadvantage of the EEPROM cells is the need for a high voltage for the programming step. Nowadays, with the ever-increasing emphasis on single-battery operation, the external high voltage supply is eliminated by replacing it

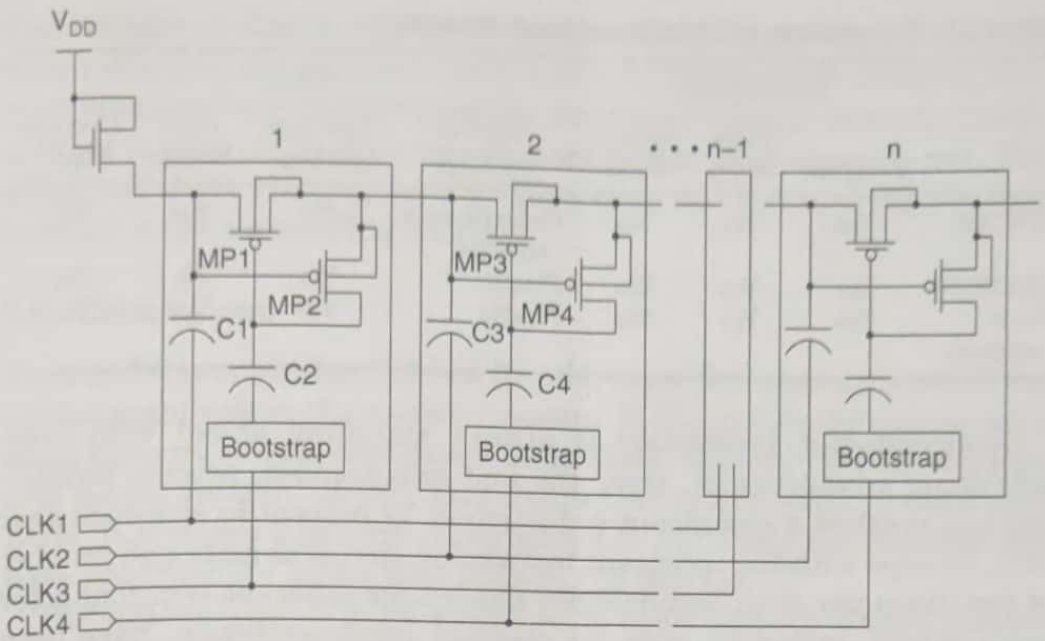



**Figure 5.19** Low-voltage charge pump.

with an on-chip high-voltage generator. The schematic diagram of a pMOS-based high-voltage generator designed for EEPROM, which is suitable for a power supply voltage below 2 V is shown in Fig. 5.19 [25]. The voltage gain at each unit stage is not deteriorated by a drop in the threshold voltage. This charge pump is able to generate sufficient programming voltage, which is applicable to a power supply as low as 1 V. C1 and C3 denote the clock-coupling capacitors, while M1 and M3 with their gates driven by the bootstrapped clock generators through C2 and C4, serve as the charge-transfer transistors. Besides, M2 and M4 are transistor switches incorporated to precharge the gates of M1 and M3. Simulation result shows that for a power supply voltage of 1 V, an output voltage of 1.75 V is acquired.

## 5.7 Future Trend and Development of ROMs

A comparison of the features and uniqueness of EPROM, EEPROM, and flash memory that determine their acceptance of commerciality in the memory market are tabulated in Table 5.2. The high-density field alterable features of the UV-EPROM have allowed it to develop a large share of the nonvolatile data storage market in spite of the high package cost and relative inflexibility in the system. EEPROM's flexibility in the system has given it sizable market niches in the low-density regime below 256 K. There is an increasing market demand for flash memories due to their high-density, low-cost process, and package as well as high system flexibility [26].

TABLE 5.2 Comparison of EPROM and Flash Memory

| | Functionality | | | Cost | | | |
|---|---|---|---|---|---|---|---|
| | Program | Field erase | System erase | Type of package | Low cost package | High density | Low cost test |
| EPROM | Yes | Yes | No | Ceramic with window | No | Yes | No |
| EEPROM | Yes | Yes | Yes | Plastic | Yes | No | Yes |
| Flash memory | Yes | Yes | Yes | Plastic | Yes | Yes | Yes |

Semiconductor technology is always marching ahead with overwhelming advancement. Over the last two decades, process technology has marked a significant reduction of 12 percent in size each year [27]. To cope with the incessant increase in circuit density and number of functions per chip, technologies needed for different semiconductor applications, including ones for memory implementation, have been emerging. For example, the Flash+ concept from STMicroelectronics integrates flash memory and EEPROM on the same chip [28]. Since flash memory allows dynamic software alteration, it is the most appropriate of all existing nonvolatile memory for storing microcomputer codes. On the other hand, for storing parametric data, EEPROM is most favorable because it is byte-alterable. Hence, Flash+ memory enjoys the key characteristic features of flash memory and EEPROM.

In the past, DRAM was the dominant technology in the memory market, but today it is apparent that flash memories have successfully surfaced to become the main focus of high demand. This is because flash memory renders the lowest cost per bit amid all semiconductor memory architectures. Its cell size is much smaller and fewer processing steps are required as compared with its DRAM counterpart [13]. The rapidly rising sales of wireless consumer products such as cellular phone, digital still cameras, internet audio devices, handheld computers and set-top boxes are no doubt the hottest drivers for tremendous development in flash memory technology. A technology roadmap is shown in Table 5.3 [29].

Memory vendors intend to intensify the density of flash memory by storing more than one data bit within a memory cell [30]. Just like the DRAM chips, EPROMs, EEPROMs, and older flash memories store only one bit of information in each cell, which is a small capacitor.

TABLE 5.3 Technology Roadmap

| Year | 1998 | 1999 | 2000 | 2001 | 2002/3 |
|---|---|---|---|---|---|
| Process | 0.35 µm | 0.25 µm | 0.18 µm | 0.15 µm | 0.13 µm |
| Products | 16 Mb | 32 Mb | 64 Mb | 128 Mb | 256 Mb-1 Gb |

The amount of charge (sufficient or vice versa) in the capacitor determines whether the cell contains a 1 or a 0. Newer flash memory, however, allocates four or more possible amounts of charge per cell thereby making it possible to store two or more bits in each memory cell. This effort will drive the cost per bit down even further, making the flash memory a more appealing solution.

## 5.8 Conclusions

First and foremost, Sec. 5.1 briefly addresses the memory system and its dominant role in the memory system. It illustrates the categories of main memory, which encompass Read Only Memory (ROM) and Random Access Memory (RAM). This chapter is devoted solely to explaining ROM. It provides a general introduction to ROM and the features that have prompted its active adoptions in the market.

The various types of ROM available for use with each of their positive and negative points are individually presented in the succeeding section. The myriads of ROM discussed include standard ROM, MROM, PROM, and floating gate memories. This progressive development of ROM architectures is achieved by the unrelenting effort applied in realizing applications that are more easily modifiable and user-friendly. Next, in view of the dominance and supremacy of the types of memories with the underlying principle being the floating gate devices, the basic physics of the floating gate concept is elaborated. The information in this section is mainly extracted from the IEEE Standard 1005–1998. Having understood the fundamental physics, the schemes of the memories based on the floating gate nonvolatile model, namely EPROM, EEPROM, and flash EEPROM are reported in Sec. 5.4.

The subsequent sections describe the ROM chip architecture, the conventional cell array arrangements, and techniques for attaining low-power consumption both at the architectural level and also at the circuit level. One important point to note is that excessive power savings can only be achieved through the use of multiple techniques. Eventually, the chapter concludes with a concise write-up on the future trend and development of the ROM technology.

## References

1. A.S. Sedra and K.C. Smith, *Microelectronic Circuits*, Toronto: Saunders College Publishing, 1998.
2. T.L. Floyd, *Digital Fundamentals*, New York: Maxwell Macmillan, 1994.
3. J. Caywood, E. Dollar, M. Knoll, H. Maes, W. Rau, J. Schreck, D. Sweetman, M.V. Buskirk, R. Wegener, and K. Yoshikawa, "IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays," *IEEE Std. 1005–1998*, Feb. 9, 1999.
4. D. Frohman-Bentchkowsky, "A Fully-Decoded 2048-Bit Electrically-Programmable MOS ROM," *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 1971, pp. 80–81.

Chapter

# 6

# Low-Voltage Low-Power Static Random-Access Memories

## 6.1 Introduction

Over the years, power requirement reduction in Random-Access Memory (RAM) has undergone tremendous advancement. Many circuit techniques have been devised to achieve active and standby power reduction in both static and dynamic RAMs [1–4]. This trend is due to the advent of high-density low-power electronic applications such as handheld terminals, mobile phones, and laptops. RAM is a Read/Write (R/W) memory, meaning that each data bit stored in its memory cells may be altered to a different bit value easily and quickly [5]. As its name implies, the time required for RAM to store (write) or retrieve (read) information is independent of the data's physical memory location. It is organized and controlled in a manner that enables each data bit to be stored and retrieved directly to a specific memory address location. Indeed, in R/W memory, the data being stored and retrieved at different memory address locations can be accessed at comparable speeds, independent of the memory locations' address values [6]. Furthermore, RAM is a volatile memory. It retains its memory patterns for as long as power is being supplied. Conversely, its contents vanish if the power is removed. Basically, RAM can be classified into two categories:

- Static RAM (SRAM)
- Dynamic RAM (DRAM)

SRAM utilizes a flip-flop mechanism, which uses static latches, and these operate in a manner similar to the way in which memory cells work [6]. The electrical feedback ensures that voltage levels are sustained so that the data may remain active indefinitely for as long as power supply continues. On the other hand, the data in a DRAM storage cell gets stored in the form of charge on a capacitor, which inevitably leaks, causing gradual charge depletion. Therefore, the major difference between SRAM and DRAM is that SRAM does not need a power refresh operation, whereas DRAM warrants periodic refreshing of the stored charge [7]. Moreover, SRAM has higher speed because it operates with differential pairs of bit-lines [8]. This chapter presents SRAM, whereas the DRAM presentation shows up later, in Chap. 7.

## 6.2 Basics of SRAM

A block diagram of a typical SRAM architecture is depicted in Fig. 6.1 [8]. SRAM chips consist mainly of address buffers, row/column decoders, the memory array, sense amplifiers and input/output buffers. The conditioning circuit activates the precharging of the bit-lines. The memory chip array contains the cells in which the data bits are already stored (as in a read operation) or in which they are about to be stored (as in a write operation). Each cell is connected to one of the many word-lines and a pair of bit-lines. Therefore, in each R/W operation, a particular cell gets selected by triggering its word-line and bit-line. Cell selection therefore occurs as the address buffers drive the row decoder and the column decoder. The intersecting point of the activated word-line and the activated bit-line marks the position of the addressed

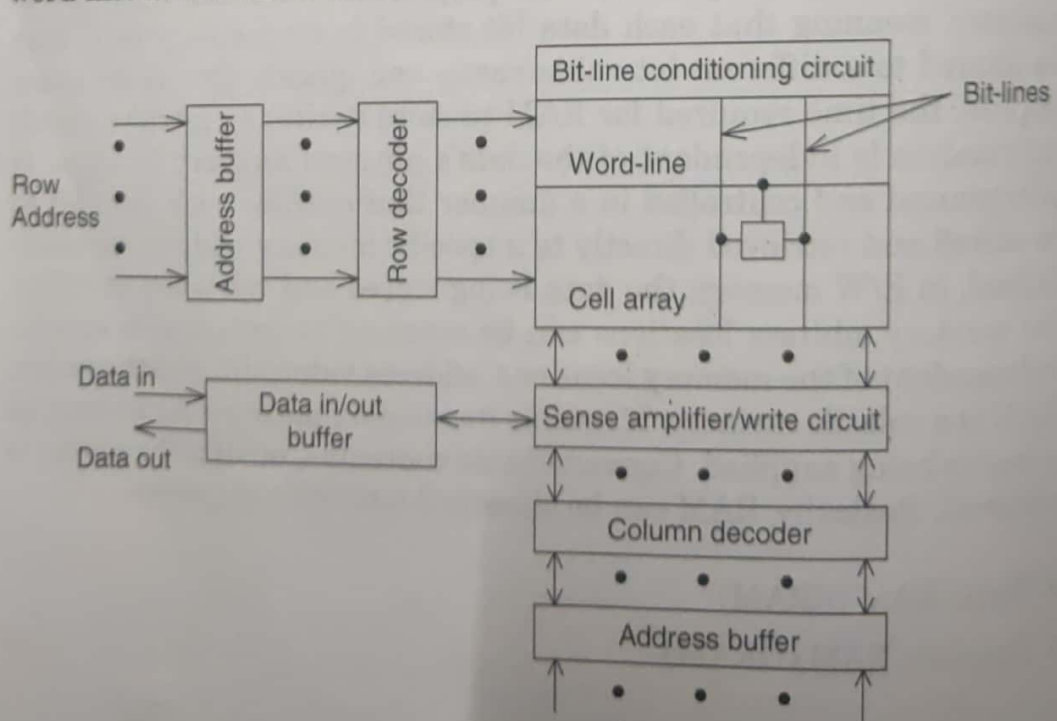



Figure 6.1 Typical SRAM architecture.

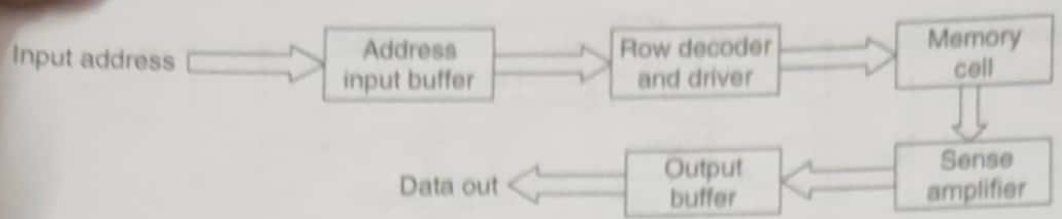



**Figure 6.2** Typical path for read access in SRAM.

cell [7]. Then, a column sense amplifier detects the contents of the selected cell in the form of small voltage variations via the memory complementary bit-lines. To achieve minimal access time, the detection must occur as quickly as possible.

The access time is determined by the critical path from the address input to the data output, as shown in Fig. 6.2. The word-line decoding and bit-line sensing add to access delays. Therefore, to reduce the sensing delay in a read operation, it makes sense to have a small swing at the bit-lines and to use a highly sensitive sense amplifier.

The commonly used control signals in an SRAM include the following: Write-Enable ($\overline{WE}$), Chip-Select ($\overline{CS}$), and Output-Enable ($\overline{OE}$). The $\overline{WE}$ signal determines the read or write mode. For the read operation, the bit-line signals are sent to the output; however, for the write operation, the bit-lines are driven by the input. In a memory system where a myriad of SRAMs are pooled together, the SRAMs are connected to common address and data buses. Only one memory section may be selected at any given time from the memory bank to place information on the data bus. In this case, the $\overline{CS}$ signal aids in the selection of a single SRAM from the memory bank. Meanwhile, $\overline{OE}$ serves as the enabling signal for the output buffer.

The timing diagram of a read operation is shown in Fig. 6.3($a$). During this time the data stored in a specific SRAM location (designated by the address) is read out. The read cycle time, $t_{RC}$, and the address access time, $t_{AA}$ are indicated. Figure 6.3($b$) shows the write cycle, which

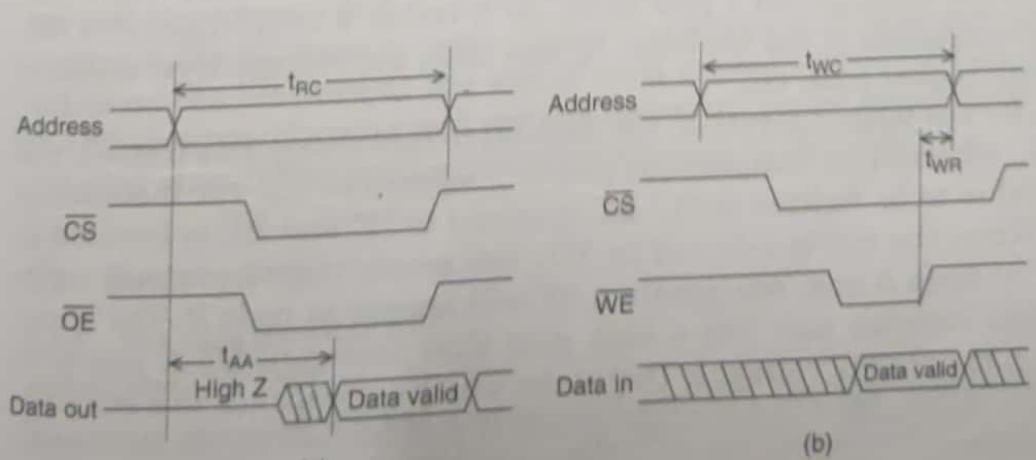



**Figure 6.3** Typical timing of an SRAM: ($a$) read cycle; ($b$) write cycle.

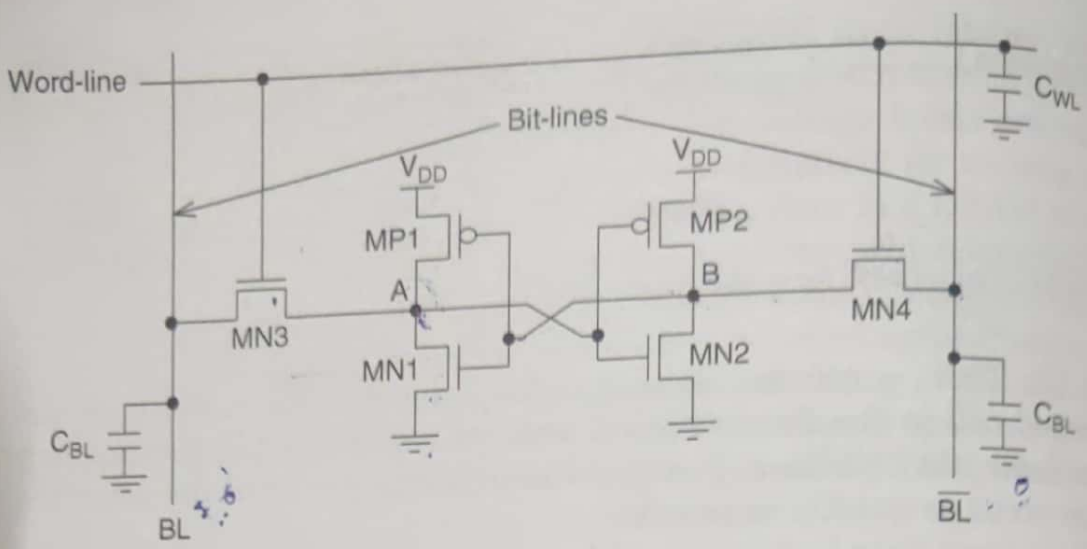



**Figure 6.4** The six-transistor CMOS memory cell.

permits changes to the information within an SRAM. Two times are indicated—the write cycle time, $t_{WC}$, and the write recovery time, $t_{WR}$.

The memory cell (MC) is a fundamental element in the design of low-power high-density SRAMs because a significant portion of the memory's size is taken by the cell area. There are various static memory cells. The memory cell implementation in its simplest form is shown in Fig. 6.4 [8]. The cell is made up of two pass-transistors (MN3 and MN4) and a flip-flop formed by two cross-coupled inverters (MP1, MN1 and MP2, MN2). The pass-transistors, which are connected to the two complementary bit-lines BL and $\overline{BL}$, are controlled by the Word-Line signal WL. They act as transmission gates providing bidirectional access between the flip-flop and bit-lines BL and $\overline{BL}$ [6].

Before the read operation, the voltages at both bit-lines get precharged to an equalized potential. Consider a stored logic "1" at node A and a complementary logic "0" at node B. When this particular memory cell is selected by asserting signal WL, $\overline{BL}$ will be discharged to the ground terminal via MN4 and MN2. Concurrently, current flows from $V_{DD}$ to BL through MP1 and MN3. As a result a small potential difference appears at the bit-lines. In this case, the voltage level at BL is slightly higher than $\overline{BL}$.

As for the write operation, the data bit to be written gets transferred to BL, whereas its complement gets transferred to $\overline{BL}$. For instance, if the value to be written is "1," BL is raised to "1" and $\overline{BL}$ is lowered to "0." When the cell is selected by WL, the access transistors will write a "1" at node A and, conversely, a "0" will appear at node B. The cross-coupled inverter pair has a high gain that causes nodes A and B to switch to opposite voltages. This state, which corresponds to a stored 1, will be maintained indefinitely unless it gets altered by another write operation.

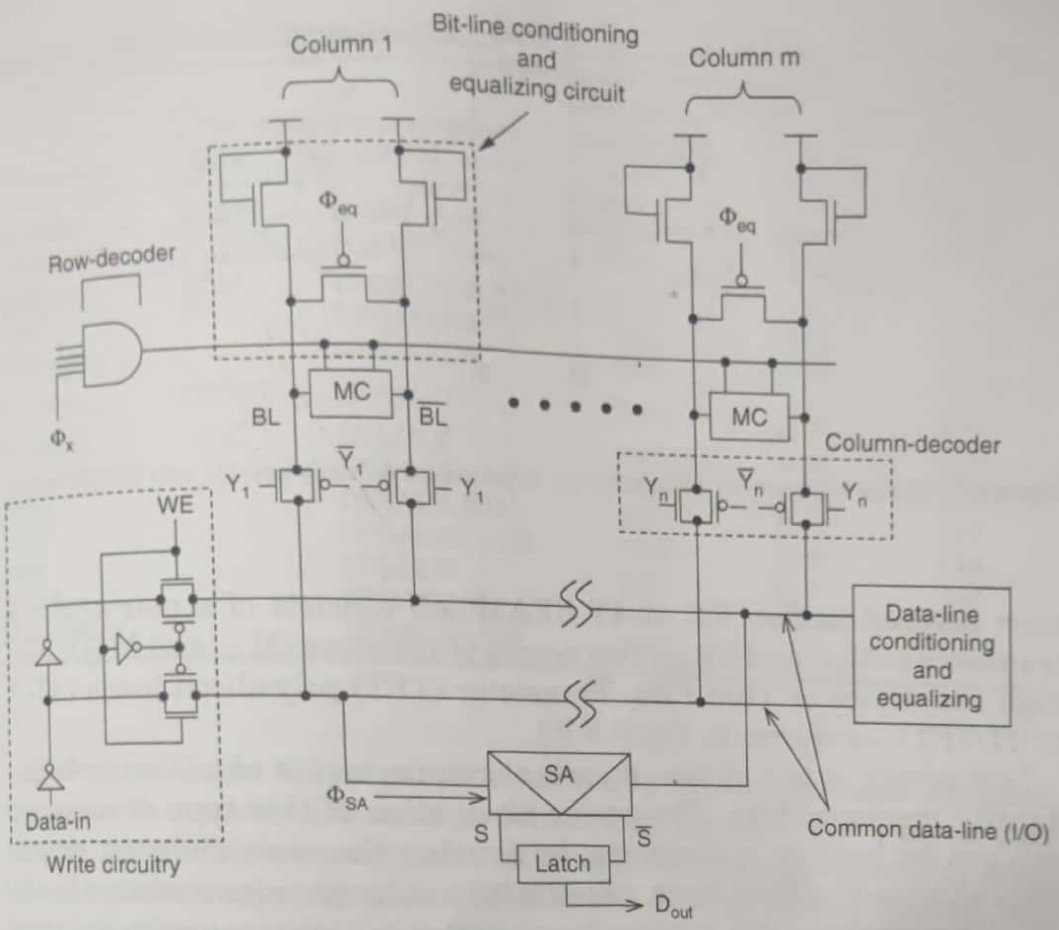



**Figure 6.5** A simplified SRAM schematic.

Figure 6.5 portrays a simplified schematic of an SRAM with read/write circuitry [8]. During a read operation, a small voltage swing variation on the bit-lines is desirable to ensure high-speed sensing. This small swing will be detected and amplified by the Sense Amplifier (SA). For the write operation, when WE is asserted, the input data and its complement get placed on the bit-lines and then get written into a memory cell just as soon as a memory cell is selected. After a write cycle, the swing on the bit-lines is large. Given the above, a differential voltage exists on the bit-lines at the end of each memory cycle. This is where the bit-line conditioning and equalizing circuit comes into play. A more detailed elaboration on the aforesaid appears in Sec. 6.4.

## 6.3 Memory Cell

### 6.3.1 The race between 6T and 4T memory cells

The most common SRAM storage cell designs are the Six-Transistor (6T) and the Four-Transistor (4T) cells. The operation of the 6T cell has

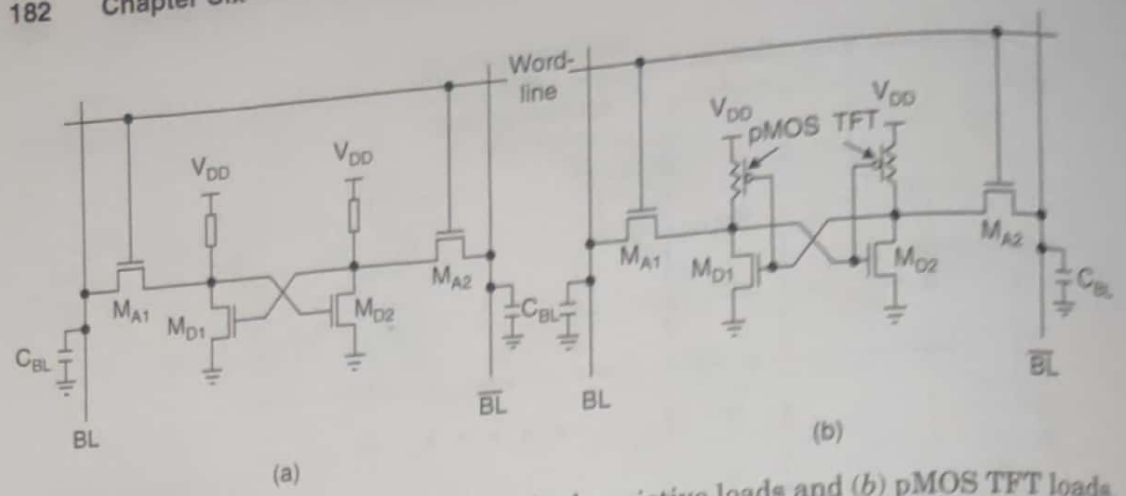



**Figure 6.6** Static 4T memory cells with (*a*) high resistive loads and (*b*) pMOS TFT loads.

been covered in Sec. 6.2. A 4T SRAM cell consists of a pair of drive transistors ($M_{D1}$ and $M_{D2}$), two access transistors ($M_{A1}$ and $M_{A2}$), and high resistance or Thin-Film Transistor (TFT) polysilicon loads (4T-R or 4T-TFT), as shown in Fig. 6.6 [8].

Low-power is one of the important requirements to realizing high-density memory chips. The power dissipation of this type of memory cell can be kept to a minimum by forming the resistor in an extra-high-ohmic polysilicon layer, which results in large-value resistive loads [6]. Nevertheless, this process necessitates a higher complexity that requires one fabrication step more than the full CMOS 6T cell [7]. An advantage of the 4T cell is that it occupies 30 to 40 percent less area than the 6T cell because the two polysilicon loads (R or TFT) can be formed above the two nMOS driver devices ($M_{D1}$ and $M_{D2}$) [8].

Owing to the small-area feature of the 4T SRAM cell as compared to its 6T counterpart, it has dominated the stand-alone SRAM market since it emerged in the 1970s. Meanwhile, the 6T cell has maintained mainstream status for on-chip storage [9]. In the early 1990s, however, despite its larger size, the 6T cell began to overtake the 4T cell and came to dominate the stand-alone SRAM market.

The motivation behind this trend is based on the advantages offered by the 6T cell. The areas of concern include process complexity, process compatibility, memory cell stability at low voltage, and soft-error rate. Table 6.1 shows the characteristics of various 4T and 6T SRAM cells [9]. They are basically classified into three types—the 4T cells, the Simple 6T (S-6T) cells and the Advanced 6T (A-6T) cells. The S-6T cells utilize the basic CMOS logic process, whereas the A-6T cells incorporate enhanced process steps such as Self-Aligned Contacts (SAC) or Local Interconnects (LI) in order to achieve reduced cell area. The 4T cells with either highly resistive or TFT loads also typically include the SAC for area reduction.

TABLE 6.1 SRAM Cells Characteristics

| Company | Type | Process features (P: poly M: metal) | Minimum gate length (μm) | Memory cell area ($\mu m^2$) | Relative process complexity |
|---|---|---|---|---|---|
| Motorola | S-6T | 1P/2M | 0.30 | 15.80 | 100 |
| Intel | | 1P/2M | 0.35 | 20.50 | 100 |
| Motorola | 4T-R | 4P/SAC/1M | 0.25 | 3.57 | 125 |
| NEC | | 3P/SAC/1M | 0.30 | 5.64 | 116 |
| Mitsubishi | | 3P/SAC/1M | 0.40 | 8.36 | 116 |
| Hitachi | 4T-TFT | 5P/SAC/1M | 0.40 | 7.16 | 135 |
| Motorola | A-6T | 1P/LI/1M | 0.28 | 9.70 | 108 |
| Motorola | | 1P/SAC/LI/1M | 0.30 | 8.80 | 109 |
| Sony | | 1P/SAC/LI/2M | 0.28 | 5.01 | 119 |
| IBM | | 1P/SAC/2LI/1M | 0.35 | 15.00 | 114 |
| IBM | | 1P/LI/2M | 0.45 | 33.6 | 109 |
| Toshiba | | 2P/SAC/1M | 0.30 | 7.65 | 109 |
| Matsushita | | 1P/SAC/LI/2M | 0.25 | 6.82 | 117 |

Based on Table 6.1, a graph contrasting relative process complexity against memory cell area is depicted in Fig. 6.7. For a given gate length, a smaller memory cell can be realized by implementing additional process steps. Advanced 6T cells can have processes that are less complex than ones laid out for 4T-R and 4T-TFT. However, even though higher process complexity always translates into increased wafer cost and fewer good dice (in most cases), the reduction in area size correlates with the increase in process complexity. This improvement in

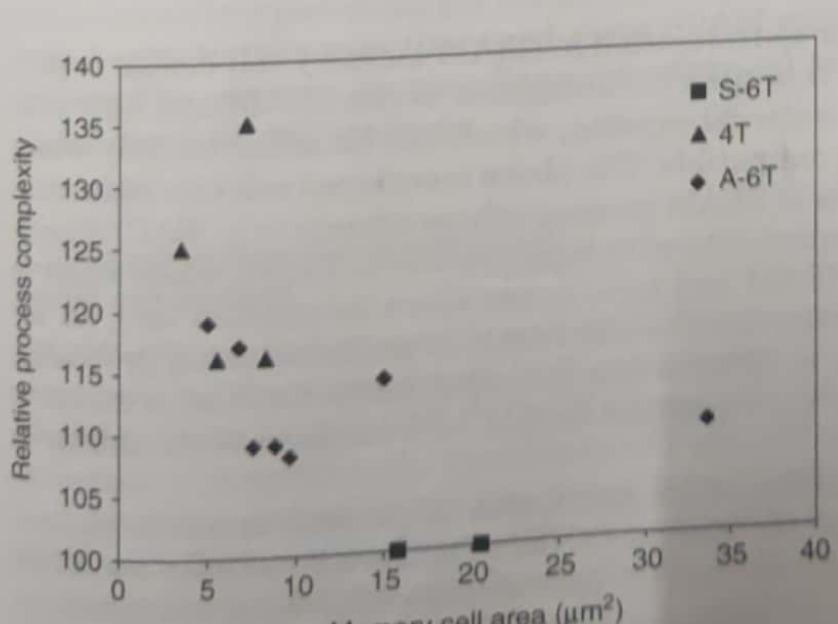



Figure 6.7 Relative process complexity vs. memory cell area.

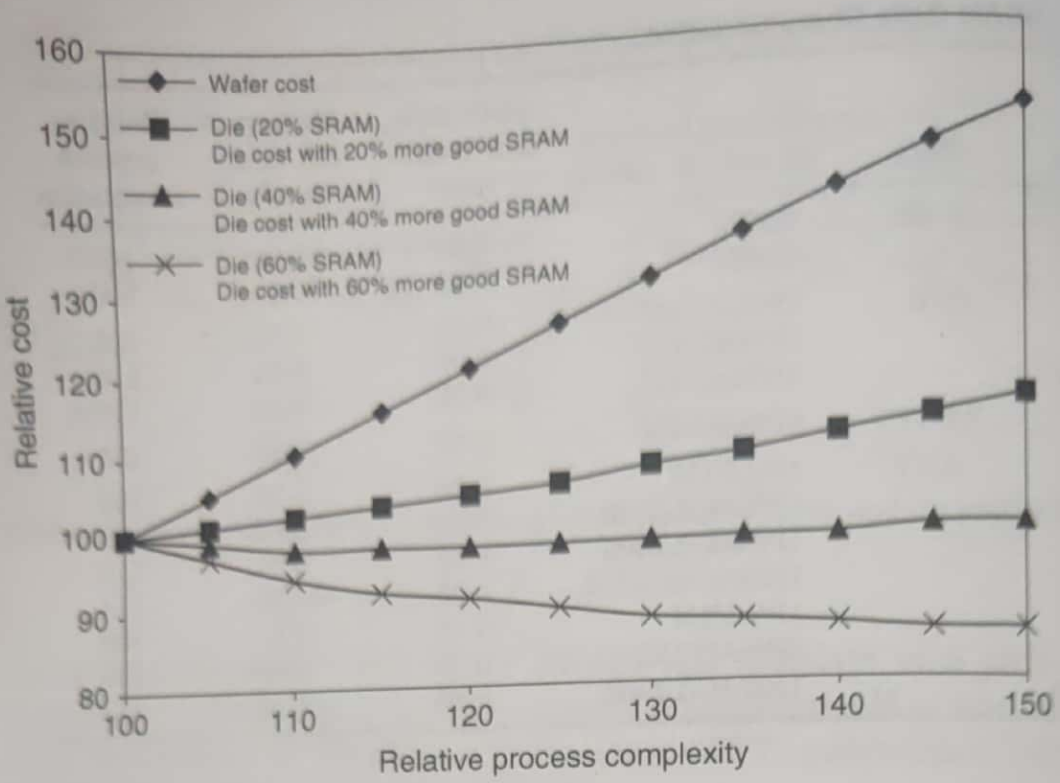



**Figure 6.8** Relative cost vs. process complexity.

area miniaturization far exceeds the shortcomings of good die diminution and higher process complexity. Therefore, these trade-offs will still reduce the manufacturing cost. Better still, this drawback of the more complex process will no longer be an issue if it results in an increase in the number of good dies, which, in turn, would more than compensate for the corresponding increase in wafer cost (Fig. 6.8). As a result, most major SRAM manufacturers resort to reducing the cell size by adding process steps.

The second major consideration in SRAM memory cell design is process compatibility. It is often advantageous to run SRAM and logic processes in a single wafer fabrication, which can be achieved only when both processes are compatible. The above mentioned cell size reduction technique by means of SRAM process enhancements (e.g., SAC) diminishes the cell's compatibility with logic processes. Hence, when effort is made to produce SRAM and logic in the same fabrication line, an incompatibility between the two will result in manufacturing problems. The A-6T cells aim to address this incompatibility. Even so, a majority of these cells have been confined to SAC use because of its ability to reduce cell size.

Memory cell stability is the third area of concern and it is characterized by Static Noise Margin (SNM). Figure 6.9 exhibits the SNM comparison between 4T and 6T cells at three different cell ratios, $\beta$. At lower $V_{DD}$, the 6T cells are especially more stable than 4T cells.

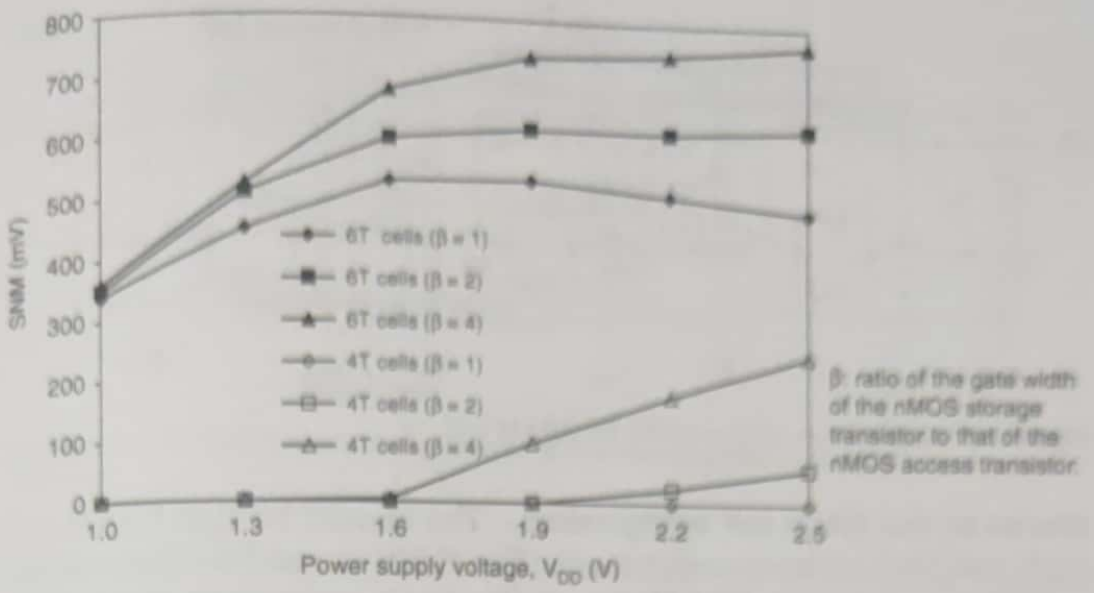



Figure 6.9 SNM vs. supply voltage.

However, due to inferior stability of 4T cells at low $V_{DD}$, they have to maintain higher cell ratios than 6T cells, thus distorting the cell size advantage that 4T cells have enjoyed.

The fourth area of concern is the Soft-Error Rate (SER). Historically, 6T SRAM cells have been more robust than the 4T cells in terms of the SER. Nonetheless, the 6T cells are normally larger and hence store more charge. As full CMOS 6T cells get reduced extensively to smaller sizes and scaled voltages, the SER then becomes an important issue. It's interesting to note that the 4T SRAM cells ruled the SRAM market for almost two decades, but because it seemed very difficult to further scale down their power supply to less than the 1.8 V boundary, 6T SRAM cells began to take over the stand-alone SRAM market. This was then followed by common process flows for advanced microprocessors and stand-alone SRAM products.

### 6.3.2 Low-voltage low-power (LVLP) SRAM cell designs

SRAM cells are indispensable sources of dissipation in many VLSI devices because they incorporate high-capacitance buses and they get accessed frequently [12]. LVLP SRAM can be achieved by modifying the structure of memory cells. Besides the 4T and 6T cells mentioned earlier, however, several other low-power SRAM cell designs have been developed that can perform well even with low supply voltage. This section describes four types of storage cells, including their circuit techniques and future prospects.

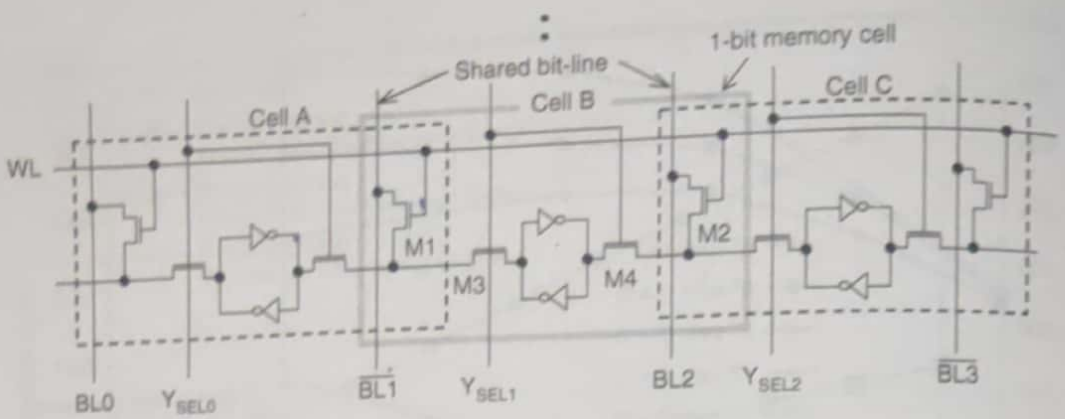



**Figure 6.10** Architecture of shared-BL 8T SRAM cell.

**Shared bit-line SRAM cell configuration.** The shared bit-line SRAM cell with modified address assignment (Fig. 6.10) is used for low-voltage word-bit configurable SRAM macrocells, having ultralow power dissipation [13]. A multiplexer (MUX)-merged charge-transfer sense amplifier is incorporated to ensure high-sensitivity read operation. As for the write operation, a bit-line precharge scheme with an equalizing line accomplishes high-speed write-recovery. Two supplementary access transistors, M1 and M2, are coupled to a conventional 6T cell. Cell A and Cell B are sharing a common BL and M1, while Cell B and Cell C are sharing BL2 and M2. By asserting WL and $Y_{SEL1}$, Cell B gets selected and M1, M2, M3 and M4 all get turned on. The data stored in Cell B then appears at $\overline{BL1}$ and BL2. Now, when $Y_{SEL0}$ and $Y_{SEL2}$ are not selected, this means that Cell A and Cell C are not activated and no memory cell current is being wasted. Alternatively, when Cell C is selected, M3 and M4 are off and the value stored in Cell C propagates to BL2 and $\overline{BL3}$. This way, we eliminate the memory cell's current waste without compromising the silicon area.

Conflicts of data will not occur in a bit-line shared by two adjacent cells because the cells are always selected at different times. Large parasitic capacitance at the bit-line is avoided since $Y_{SEL1}$ controls M3 and M4, which are not connected to bit-lines directly. Additionally, the parasitic capacitance at the word-line is decreased due to the fact that the number of transistors tied to the word-line is reduced by almost half as compared to the 6T cell. Thus, the word-line selecting delay and power dissipation due to charge and discharge is minimized.

Nonetheless, $Y_{SEL}$ has larger parasitic capacitance than that of a bit-line because two devices are adjoined to it (e.g., M3 and M4) and, due to all the charging and discharging, the power consumption will increase. To overcome this setback, the column address controlling $Y_{SEL}$ is assigned to more-significant bits in a memory address because a more-significant bit is generally less likely to change than a less-significant bit. Hence, the effective operating frequency of $Y_{SEL}$

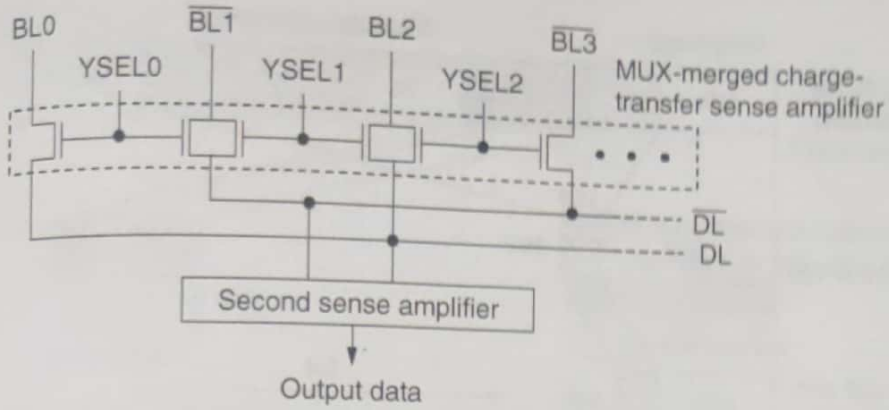



**Figure 6.11** The MUX-merged charge-transfer sense amplifier as the read circuitry.

will be reduced. As a result, the power dissipated by selecting $Y_{SEL}$ will also diminish.

Since the two access transistors are connected in series, the memory cell current is reduced. This in turn reduces the bit-line signal and increases the access time. To address this issue, a MUX-merged charge-transfer sense amplifier (Fig. 6.11) has been devised. When a transistor in the MUX is selected by the $Y_{SEL}$ signal, it acts as a charge-transfer amplifier and enlarges the bit-line signal. The second amplifier will then amplify the output signal to the CMOS level. This read circuitry increases the sensitivity of the read operation and hastens the access time previously affected by the reduced memory cell current. A drawback of this architecture is that the total cell area penalty is 15 percent compared to the conventional 6T cell.

The power consumption of this design is drastically reduced as compared to the 1 V operating word-bit configurable SRAM macrocell using the multi-threshold voltage CMOS (MTCMOS) technique [14]. The simulation result on power dissipation is presented in Fig. 6.12. Due to its shared bit-line cell configuration, there is a 93 percent power dissipation reduction in the memory cell. By taking into consideration the slight increase in power in the peripheral circuitry, the total power savings are about 75 percent as compared to the configurable SRAM macrocell. This configurable shared bit-line macrocells are of high demand in memory cell development where the low-power feature is an important metric of measurement. However, it remains unpopular when area is of prime concern.

**Power-efficient 7 T SRAM cell with current-mode read and write.** Unlike the traditional current-mode SRAMs, where only the read access is executed in current-mode [15–17], this section presents a memory cell utilizing the current-mode scheme in both read and write operations, namely the 7T cell [12]. The current-mode method is advantageous

## 6.9 Low-Power SRAM Technologies

### 6.9.1 Sources of SRAM power

The power present in an SRAM is categorized into two parts—active and standby (data retention) power [23]. The active power is the sum of the power consumed by the row and column decoders, memory array,

sense amplifiers, and peripheral circuits such as the input/output buffers. The effective data retention current of the unselected memory cells when the sense amplifiers are significantly disabled contributes to the standby power of an SRAM. Static current from other sources is negligible.

### 6.9.2 Development of low-power circuit techniques

The following circuit techniques for low-voltage low-power operation embrace both historically developed methods as well as improved versions introduced in recent years.

**Capacitance reduction.** Capacitance in memory is largely contributed by word-lines, bit-lines, and data-lines. Hence, low-power consumption can be achieved by reducing the size of these lines. An approach commonly employed is the partial activation of the Divided Word-Line (DWL), which features a two-stage hierarchical row-decoder structure, shown in Fig. 6.38 [27]. One main word-line in the data-line direction usually has four sub-word-lines coupled to it. This arrangement comprises the area of a main row decoder with the area of a local row decoder [28]. By utilizing the two-step decoding, the sub-word-line of each subarray is selected by the main word-line and the corresponding local row-decoder. Thus, only the memory cells in the particular sub-word-line within a selected subarray are accessed in a cycle [8]. This will greatly minimize the capacitance of the address-lines to a row decoder and the word-line RC delay.

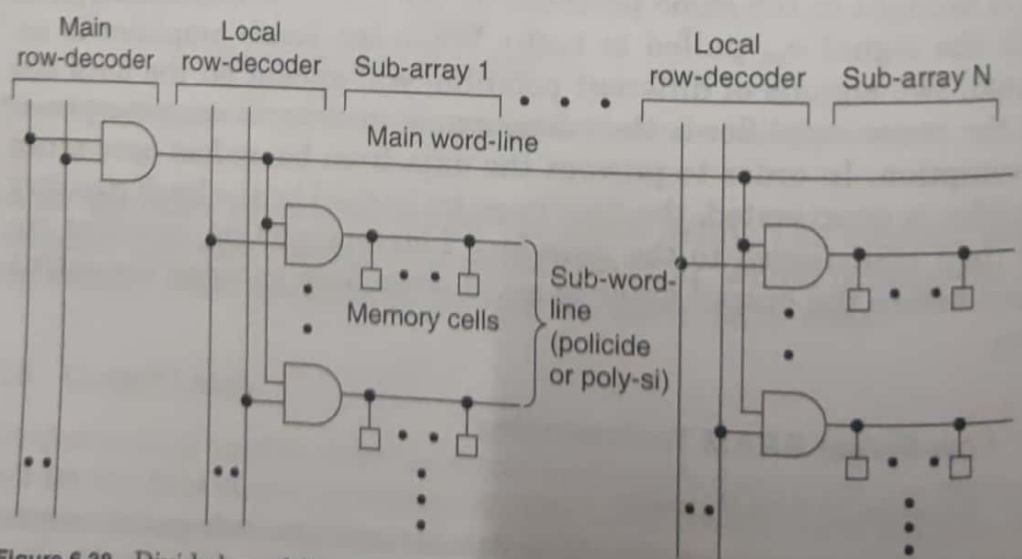



*Figure 6.38* Divided word-line structure.

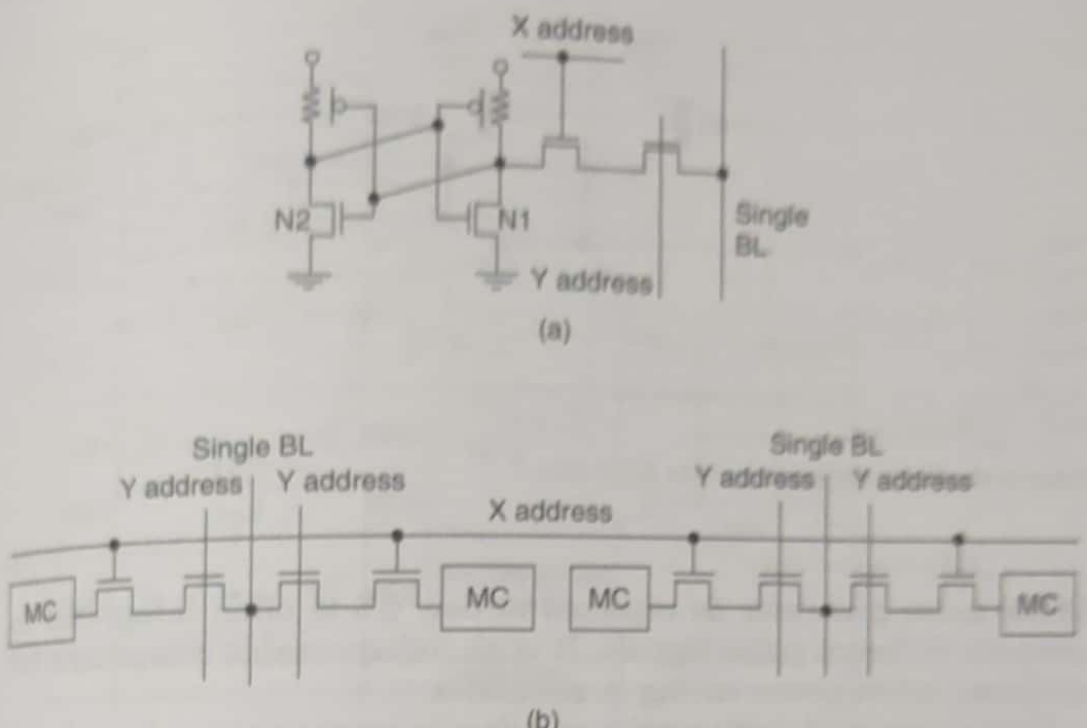



Figure 6.39 (*a*) Memory cell (MC) for the SCPA architecture. (*b*) Diagram of the SCPA.

To further reduce power consumption, the Single bit-line Cross-Point cell Activation (SCPA) illustrated in Fig. 6.39 has been developed [29]. The Y address controls both the access transistors and the X address. Since only one memory cell at the cross-point of X and Y is activated, a column current is drawn only by the accessed cell. This mechanism therefore requires the minimum possible column current without having to increase the block division of the cell array. Accordingly, the decoder area is reduced and the memory core area is 10 percent smaller than it is in the conventional DWL scheme. The drawback of this architecture is that the X and Y lines have to be boosted during the write "high" cycle.

**AC current reduction.** In memory applications, the commonly used technique to lower AC current is multistage decoding [23]. By using a two-stage decoding architecture as portrayed in Fig. 6.24 of Sec. 6.5, the number of transistors, fan-in, and the loading on the address input buffers are reduced as compared to a single-stage decoder. Subsequently, both speed and power characteristics are improved and optimized.

**Pulse operation.** Pulse operation techniques can be used to shorten the duration of active duty cycle and lower the amount of dissipated power [23]. To achieve this purpose, an on-chip Address Transition Detection

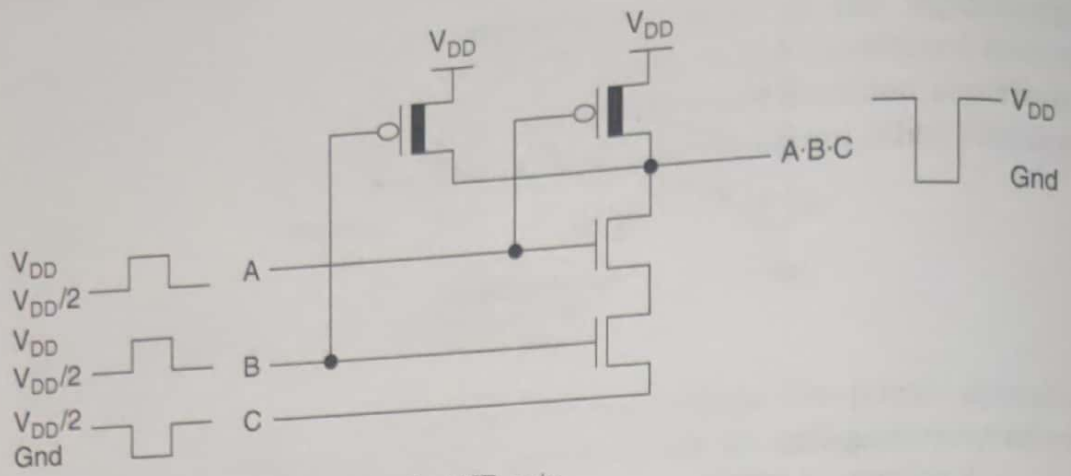



**Figure 6.40** Half-swing pulse-mode AND gate.

(ATD) pulse generator as reported in Sec. 6.6 is often integrated to generate different pulse signals. It is an indispensable component for attaining active power saving in memories.

Power consumed during write and decode operations can be reduced by lowering signal swing on high capacitance predecode lines, write-bus lines, and bit-lines without affecting performance [30]. The majority of the power savings is derived from operating the bit-lines at half-$V_{DD}$ rather than full $V_{DD}$. The typical shortcoming of using reduced swing signal is a reduced gate overdrive at the receiver end, which causes performance deterioration. To combat this drawback, a technique utilizing the half-swing pulse-mode scheme is introduced [30]. Its schematic diagram is shown in Fig. 6.40. This gate is actually a combination of a voltage-level converter with an AND logic. A positive half-swing and a negative half-swing combined with the receiver-gate logic style lead to a full gate overdrive for all of the forward transition driving transistors. Here, positive half-swing denotes swinging from a rest state of $V_{DD}/2$ to $V_{DD}$ and back to $V_{DD}/2$, whereas negative half-swing implies transition from a rest state of $V_{DD}/2$ to ground, and back to its original state. With this method, the low-swing inputs have negligible effects on the performance of the receiver. On the other hand, the usage of charge recycling between the positive and negative half-swing pulses can further reduce the power consumption.

**Low-power sensing.** As discussed in Sec. 6.7, the current-mode sense amplifier has gradually replaced the traditional current-mirror voltage amplifier in the sense circuit. This is because the current-mode sense amplifier facilitates low-voltage and high-speed operation by providing a very low bit-line and data-line voltage swings. An improved current-mode sense amplifier with a modified current conveyor for 1.5 V power supply was developed, as shown in Fig. 6.41 [31].

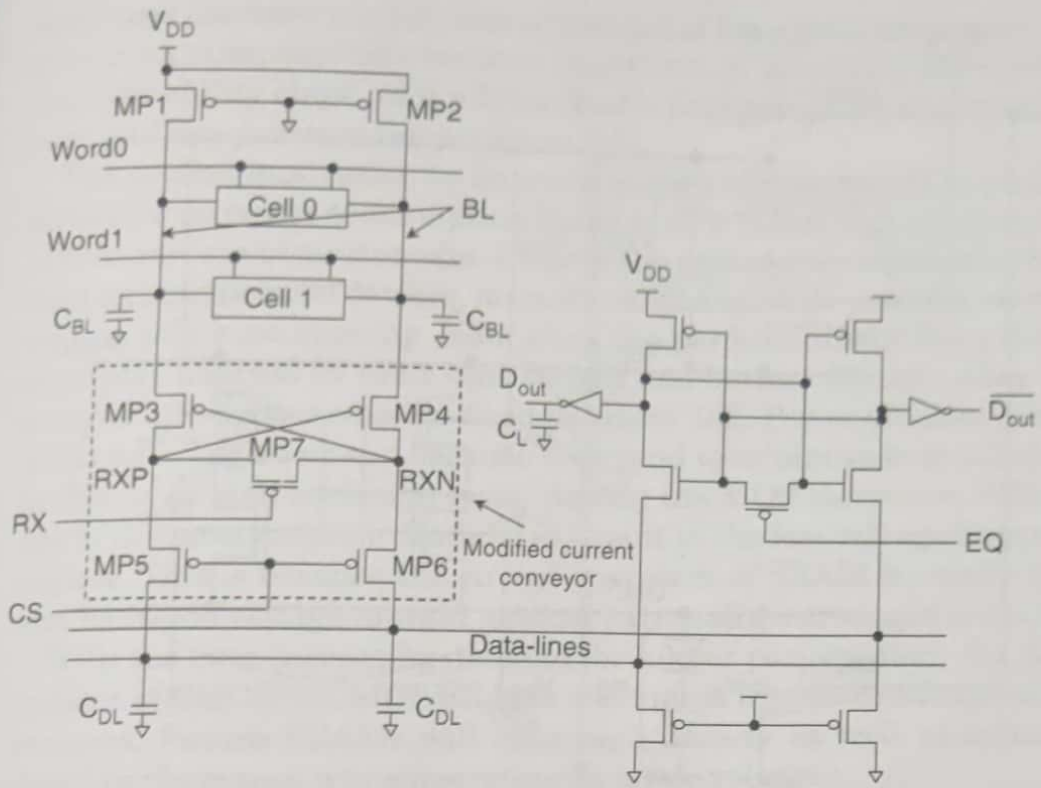



**Figure 6.41** An improved current-mode sense amplifier with a modified current conveyor.

The new circuit shows performance leverage over the conventional sense amplifier by resolving the problem associated with pattern dependency, which limits the scaling of the operating voltage. The modified current conveyor makes use of an additional pMOS transistor MP7. After every read cycle, this equalization transistor will be turned on, equalizing nodes RXP and RXN and hence eliminating any residual differential voltage between them.

**Leakage current reduction.** The most effective way of reducing the dynamic power consumption is to scale down the operating power supply voltage ($V_{DD}$). Nevertheless, the threshold voltage will be reduced together with the scaling of $V_{DD}$. The reduction of the threshold voltage will then amplify the leakage current during both active and standby modes. A number of techniques have been proposed to rectify this phenomenon [32–34]. The Auto-Backgate Controlled Multi-Threshold voltage method (ABC-MT-CMOS) is shown in Fig. 6.42 [34]. The internal circuitry is designed with low-$V$th (L-$V$th) devices. Meanwhile, its external transistors (MP1, MP2, MP3, and MN1) are of high Vth (H-Vth) and function as switches to cut off the leakage current. It is capable of tremendously reducing the leakage current during "sleep" mode.

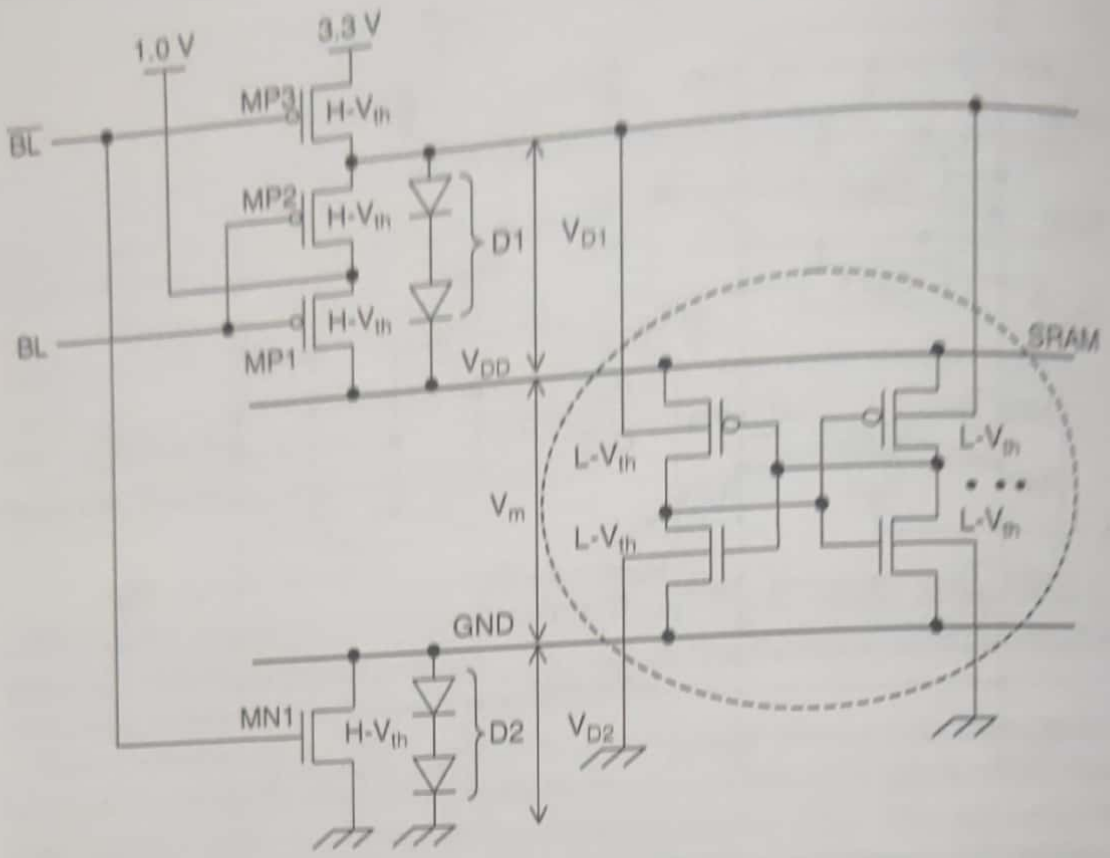



Figure 6.42 Schematic diagram of the ABC MT-CMOS.

## 6.10 Future Trend and Development of SRAM

Today's relentless development efforts to keep improving SRAMs is geared toward low-power applications because consumers are demanding higher speed performance and increased packing density. In other words, the continual demand for better SRAMs fuels the underlying motivation for relentless insistent development.

The reasons for these strong requirements lie mainly in two areas.

First, seeing as how the recent research trend is moving into developing high-performance Ultra-Large-Scale-Integration (ULSI) systems, the demand for greater scale through SRAM and logic LSI integration is escalating. SRAM technology has achieved noteworthy advancements in power reduction and, even still, circuit designers remain optimistic about its prospects for future growth. The rapid progress in ULSI fabrication has led to smaller device geometry and increased transistor density in integrated circuits [23]. Circuits with high complexities and very high frequencies have also been emerging. These circuits consume a great deal of power and they dissipate excessive power in the form of heat, which actually increases their vulnerability to run-time failures and can lead to serious reliability problems. In addition, with memories accommodating an integral share of power consumption in processors,

emphasis has been placed on development of low-power memories. Low-power consumption has become important in providing low-cost and high-reliability chips, as it allows plastic packaging, low operating current and low-junction temperature [28].

The second motivation to improve energy efficiency and to place restrictions on power consumption is due to new technology requirements in this new era of multimedia. Clearly, the demand for notebooks, handheld communication devices, memory cards and other portable electronics has only kept growing. Now, since the performance of these devices is greatly affected by their size, weight and battery lifetime, they have evolved into lighter smaller-sized products [23]. Power-efficient designs are thus in high demand because they tend to minimize both reliability problems as well as design costs. Among the MOS memories, SRAM is one of the most suitable candidates to suit in the low-voltage low-power regime. This is because the circuit operation of SRAM is totally static and has wide voltage margin against externally introduced noise [35].

With the ever-increasing demand for higher performance, the development of CMOS/BiCMOS SRAMs will march into more advanced generations. Future SRAMs will offer high density as well as enhanced speed performance while operating at lower voltages.

## 6.11 Conclusions

This chapter presents a thorough and wide-ranging review of the development and evolution of low-voltage low-power SRAM. It begins with some basic principles and then goes on to elaborate on fundamental components and devices essential for proper SRAM operation, such as the memory cell, row/column decoder and sense amplifier, to name a few. Multiple conferences and journal papers presenting the latest exploratory approaches to cell design and sense amplifier design are also included. The subsequent section reports on the sources of power that are present in SRAM and, furthermore, outlines the techniques suitable for achieving power savings and efficiency. Finally, the chapter concludes with a brief write-up on the future trend of SRAM memory. This knowledge is invariably essential as it stimulates new ideas that, according to today's research, should further reduce the operating power supply voltage, which is evidently the most eminent means to minimizing power consumption.

## References

1. N. Shibata, M. Watanabe, and Y. Tanabe, "A current-sensed high-speed and low-power first-in-first-out memory using a wordline/bitline-swapped dual-port SRAM cell," *IEEE J. Solid-State Circuits*, Vol. 37, No. 6, pp. 735–750, June 2002.
2. K. Koh, R. J. Hwang, G. H. Han, K. H. Kwak, Y. S. Son, J. H. Jang, H. S. Kim, D. Park, and K. Kim, "Ultra-low power and high speed SRAM for mobile applications using

explored in logic technology. To add difficulty to the integration process, this technology specially dedicated to the DRAM cell actually restricts the use of process modules, which are required for transistor performance in logic technology [20]. Nonetheless, possible solutions and remedies have been explored to address the above-mentioned obstacles [20–23].

## 7.3 Basics of DRAM

### 7.3.1 Basic architecture

Figure 7.5 presents a block diagram of the basic DRAM architecture for a 4 Mb structure [2]. It consists of a number of fundamental building blocks, namely the memory cells (MC), the row and column decoders, the word-line drivers, the sense amplifiers, the precharge and equalization circuits, the read/write circuitries, the Half-Voltage Generator (HVG), the Back-Bias Generator (BBG), the Boosted-Voltage Generator (BVG), the Voltage-Down Converter (VDC), and the In/Out (I/O) buffers.

Several select/control pins are required for proper operation of DRAM. The address logic is separated in time to cater for two types of fields, the row address and column address. The Row Address Select ($\overline{RAS}$) is an active low-control signal used to latch the row address and to initiate the memory cycle. When $\overline{RAS}$ is active, the data available at the address pins will determine the row to be selected. In the meantime, the Column Address Select ($\overline{CAS}$) is used to latch and activate the column address. Similarly, the column to be selected depends on the information on the address pins when $\overline{CAS}$ is active [24]. In other words, the memory cell is selected by the cross-point between the row and column addresses.

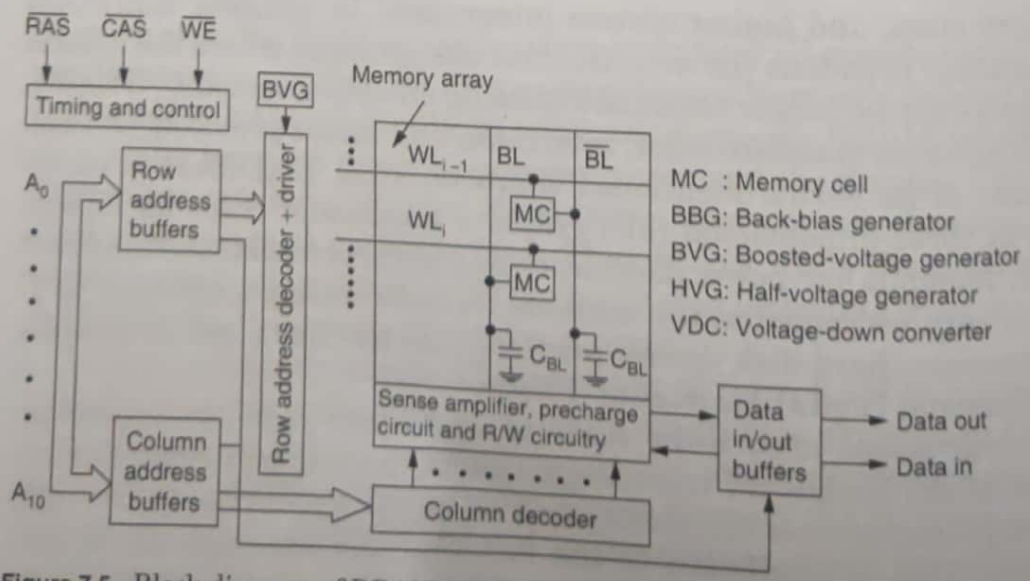



**Figure 7.5** Block diagram of DRAM architecture.

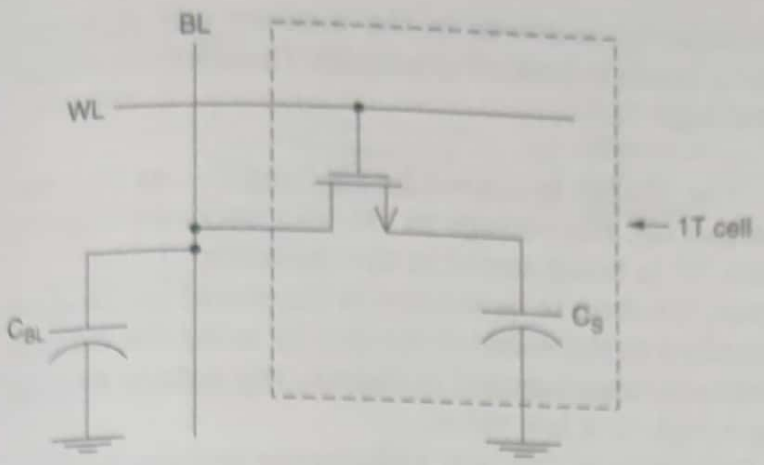



Figure 7.6 Schematic diagram of 1T DRAM cell.

While the Write Enable ($\overline{WE}$) signal is used to determine between the read and the write accesses, the Output Enable ($\overline{OE}$) permits data to arrive at the data input/output buffer during a read operation whenever it is switched to a low logic. $\overline{OE}$ is de-activated during a write operation. The data I/O pins are needed to interface the DRAM with the outside world.

Each DRAM memory cell consists of a single pair of nMOS transistor and capacitor. It is also commonly known as the one-transistor (1T) cell, which renders smaller chip size and lower cost. It can perform both read and write operations. Its schematic diagram is depicted in Fig. 7.6 [2]. The word-line (WL) is the output signal from row decoder and controls the transistor gate running perpendicular to the Bit-Line (BL). In contrast to the SRAM cell where two BLs are used, DRAM utilizes only one BL. Reading and writing is accomplished by turning on the access transistor with the word-line. The storage capacitor $C_S$ is used to store the data in the form of charge, whereas $C_{BL}$ denotes the capacitance at the BL, including the parasitic load of the connected circuits. The role of $C_S$ to accumulate charge necessitates periodical refreshing of the cell to avoid information loss.

### 7.3.2 Read and Write Operation

Prior to the read operation, the bit-line parasitic capacitance ($C_{BL}$) is precharged to logic half-$V_{DD}$, where $V_{DD}$ is the power supply voltage. Subsequently, WL is pulled high in order to activate the nMOS-type access transistor. In an array of memory cells, the appropriate cell to be read from is determined by the row and column coordinates [1].

**Read "1" Operation.** When WL is asserted, the voltage at BL tends to increase slightly if the charge stored in $C_S$ is an ideal logic "1."

This implies that a logic "1" is stored in the memory cell. However, the charge stored in $C_S$ tends to leak off gradually. Therefore its charge is usually not at ideal logic "1."

**Read "0" Operation.** The charge is shared by $C_{BL}$ and $C_S$. As $C_{BL}$ discharges and $C_S$ charges up, the voltage at BL will decrease slightly, indicating that a logic "0" is being stored in the capacitor.

As for the write case, the data to be written is imprinted on the data I/O lines and then applied to the selected bit-line by using the column decoder. When a particular memory cell is chosen, the data is written on the BL by forcing a high or a low value.

**Write "1" Operation.** The bit-line potential is pulled to logic "1" by charging $C_{BL}$ to either $V_{DD}$ or $(V_{DD} - V_{th})$ depending on the type of write circuitry used, while the selected word-line is pulled high by the row address decoder. $V_{th}$ implies the threshold voltage of the nMOS transistor. The access transistor will then turn on by asserting the WL, allowing the storage capacitor to be charged up to a logic "1" level.

**Write "0" Operation.** Bit-line is pulled to logic "0" by the write circuitry whereas the selected word-line is raised by the row address decoder. The access transistor turns on and storage capacitor discharges if it is initially storing a logic "1" or remains unchanged if it is already storing a logic "0."

## 7.4 Self-Refresh Circuit

DRAM makes use of capacitors to store charges but these charges tend to erode away as time goes by. As a result, refreshing of the DRAM to its pristine version is compulsory before the stored voltages reach unacceptable values [25]. Even though the read and write operations provide automatic refreshing to each and every memory cell in a selected row, periodic refreshing of the entire memory array must be executed every 5 to 10 ms, an interval as required by the particular chip [1]. A self-timed refreshing method accustomed for high-speed, low-power and high-density memory array is reported in this section.

### 7.4.1 General mechanism

The refresh operation of a memory array requires the use of a circuitry, which is able to offer a refreshing sequence. At the same time, this circuit must have the capability to indicate and single out the row that needs to be refreshed so that a timely refreshing can be provided to the DRAMs. Once the refreshing commences, it continues to sweep through the whole memory array until all the memory cells are duly refreshed.

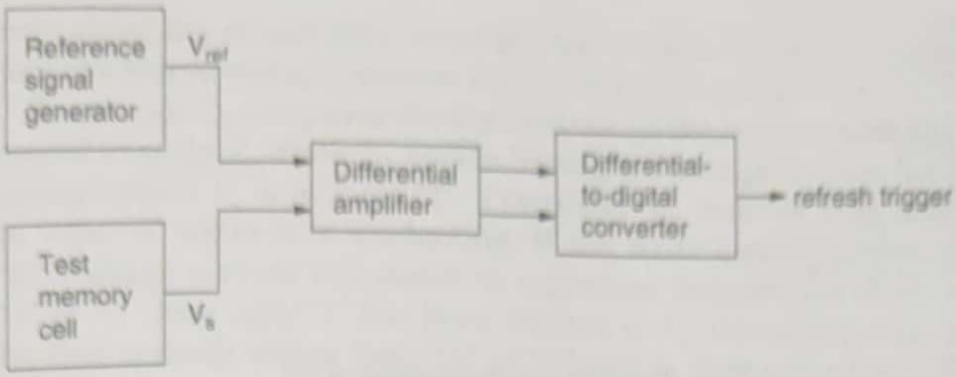



Figure 7.7 Block diagram of the refresh trigger signal generator.

There are two approaches that can be employed for realizing the refresh sequence circuit. The first method is to make use of an address decoder and a counter. The counter will start counting from 0 and proceed toward the maximum number of entries. The second approach is to utilize a shift register where only one row is selected and this selection will then be consecutively shifted to choose the subsequent row. This scheme requires clock, reset, and refresh trigger pins. As its name implies, the reset signal is used to set the circuit to point to the first row for refreshing. The refresh trigger signal acts as a control pin, which is activated only when there is a need to instigate the refresh operation. A block diagram outlining the generation of the refresh trigger signal is depicted in Fig. 7.7 [25]. In order to minimize static power dissipation within differential amplifiers, the clocked CMOS logic is used [26, 27].

A test memory cell, which characterizes closely the memory array's cells, functions as a test bed to monitor a degrading stored data, $V_s$. The Reference Signal Generator (RSG) produces a reference signal, $V_{ref}$, used to be compared with the $V_s$. The signal $V_{ref}$ is kept constant at about 2 V being the minimum acceptable value to which the stored "1" in memory array is allowed to degrade to, whereas $V_s$ represents a logic "1" that degrades with time being monitored by the differential amplifiers. A stored "0" has been determined not to increase in value when stored. Therefore, no test cell is used to monitor the behavior of a stored "0". The comparison and the difference between $V_s$ and $V_{ref}$ are provided by the nMOS differential amplifier. The CMOS differential-to-digital converter is then used to transform the difference between $V_{ref}$ and $V_s$ into a single-ended output that constitutes as the refresh trigger signal to initiate the refresh cycle.

### 7.4.2 Reference Signal Generator (RSG)

The RSG is shown in Fig. 7.8 [25]. It comprises four transistors and a write-line. When the write-line is asserted "high," $T_w$ is used to transfer a logic "1" to the storage device. In this case, the gate of the transistor

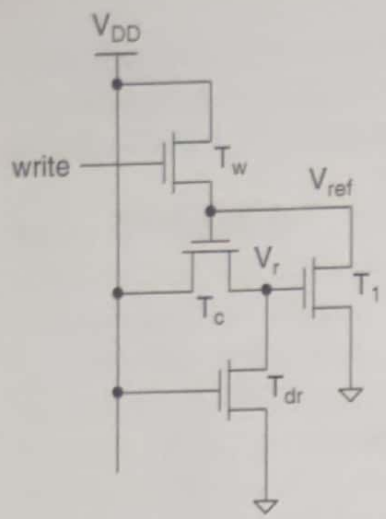



**Figure 7.8** Circuit diagram of the reference signal generator.

$T_c$ poises as the storage element and allows a logic "1" to be passed to node $V_r$ once a logic "1" has been written to $V_{ref}$.

$T_1$ comes into action as it is necessary to reduce $V_{ref}$ from $V_{DD} - V_{th}$ to the lowest possible value to which a stored "1" is allowed to degrade to. It provides a temporary leakage path from $V_{ref}$ to the ground. Transistor $T_{dr}$ is always on so as to drain the node $V_r$ and eventually turns $T_1$ off. To preserve $V_{ref}$ at a constant voltage level, the write-line signal is periodically raised to logic "1" in order to pass a "1" to storage node.

### 7.4.3 Test memory cell

A resemblance of logic "1" stored in any of the cells in a memory array, the test memory cell storing the degrading $V_s$ signal, is portrayed in Fig. 7.9 [25]. It is built from four transistors and connected to bit-write- and read-lines. The serial transistors $T_{w1}$ and $T_{w2}$ are used to pass data from the bit-line to node $V_s$. This series combination passes a voltage slightly smaller than the voltage in a memory cell where only one transistor is used. As a result, this voltage will decay faster than a "1" stored in a typical memory array. It prepares a safety margin in

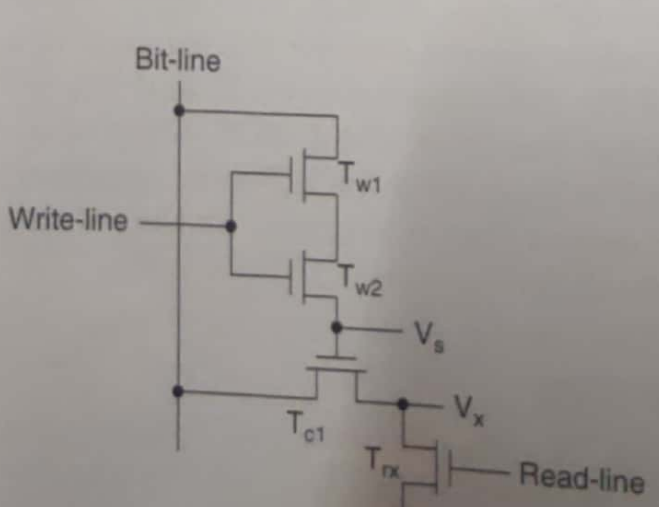



**Figure 7.9** Schematic diagram of the test memory cell.

refreshing the stored data, ensuring the refresh operation to occur before the stored voltage reaches an unacceptable level.

Transistor $T_{c1}$ compares the logic values on the bit-line with the value stored at node $V_s$ while $T_{rx}$ avoids having a direct path from bit-line to ground when $V_s$ is at logic "1". Transistor $T_{rx}$ assures that node $V_x$ is at logic "0" when it is conducting. If the write signal is kept at logic "0", leakage current will result in a gradual degradation of the stored "1" at $V_s$. Once logic "1" has been written to $V_s$, the information on the bit-line is made either logic "1" or "0" and is changed every now and then to represent the data presented on the bit-lines of the memory arrays. The voltages $V_s$ and $V_{ref}$ generated by the test memory cell and the RSG, respectively, will then be fed to the subsequent stage, which is the nMOS differential amplifier.

### 7.4.4 The nMOS differential amplifier

Figure 7.10 presents the schematic of the nMOS differential amplifier [25]. The always-conducting transistors $T_{L1}$ and $T_{L2}$ serve as the resistive active loads for this active amplifier. Transistor $T_{ref}$ is driven by reference signal $V_{ref}$ whereas transistor $T_s$ is driven by the degrading stored signal $V_s$. There is always a direct path from $V_{DD}$ to ground because $T_{ref}$, $T_{L1}$ and $T_{cs}$ are conducting at all times. Thus, the output voltage $V_{diffr}$ has to achieve a constant value. Initially, when $V_s$ is at a nondegrading "1", the voltage difference between $V_{diffr}$ and $V_{diffs}$ is at a maximum level. This is due to the fact that $T_s$ will cause $V_{diffs}$ to be discharged to logic "0" as long as $V_s$ is an acceptable logic "1", since it conducts better than $T_{ref}$. As $V_s$ decreases to be equal to $V_{ref}$ due to leakage current, $V_{diffs}$ will increase approaching $V_{diffr}$, causing the difference between $V_{diffs}$ and $V_{diffr}$ to be zero. This is the critical point of interest since the reference signal is the lowest acceptable value that the stored value $V_s$ is permitted to deplete to. The output plot showing $V_{diffs}$ and

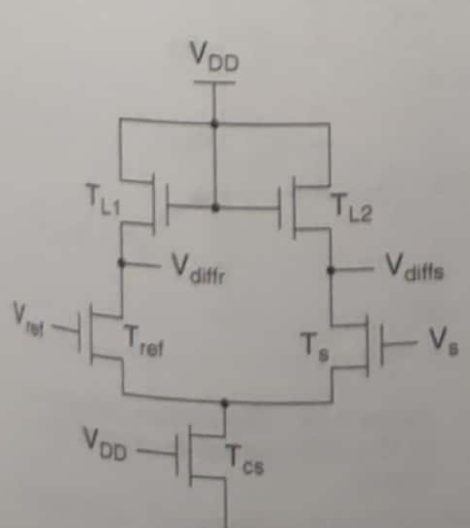



Figure 7.10 The nMOS active load differential amplifier.

$V_{outb}$, which is the complement of $V_{out}$, is the refresh trigger signal, which initiates the refresh cycle.

## 7.10 Future Trend and Development of DRAM

The standard/commodity DRAM has seen remarkable progress as its memory-chip capacity went from 4 Gb [4–7] at the R&D stage and 256 Mb at the mass production stage [38]. Moreover, the innovative memory subsystem scheme exemplified in Ref. [39] has immensely boosted the DRAM throughput to 1.6 Gbyte/s for a 1 Gb chip. Undeniably, DRAM technology is marching into the multigigabit era.

As we move into the multigigabit era, the reduction in power dissipation is of utmost importance for the purpose of extending the data holding time and containing internal noises [40]. Meanwhile, low-power circuits attain ultrahigh throughput for gigabit chips as they prolong the refresh time by reducing junction temperature [41]. Apart from that, low-voltage circuits overcome the reliability issue in miniaturized devices in spite of the fact that the power-supply standardization remains a critical issue. If the problem caused by sub-threshold current is successfully resolved, such circuits may create a niche in the markets for battery-, or even solar-cell-operated products, which will help to justify the massive investment expected in the multigigabit era.

DRAM has successfully created a niche for itself in today's semiconductor industry. The quadrupling of memory capacity through high-density technology such as the one-transistor one-capacitor (1T) cell, vertical cell capacitors, scaled CMOS devices, and multilevel metal wiring has contributed to this progress [41]. However, as chip size increases, the difficulty in achieving higher speed increases. This is due to the fact that as memory capacity increases, both wiring capacitance and wiring resistance will also upsurge [40]. Future improvements in speed will have to rely on genuine breakthroughs and developments in the circuit design of high-performance chips.

There are several DRAM architectures that render promising performance and are greatly recommended for future development. The Rambus DRAM is in direct competition with the industry standard synchronous DRAM, SyncLink DRAM, and DDR DRAM architectures. The SLDRAM Consortium was incorporated in 1998 and it is a nonprofit corporation that drives the development of an open specification for the SLDRAM. It combines cooperative engineering talents from most of the DRAM and system companies to focus on the challenge of satisfying high bandwidth DRAM requirements, and thus rivals Rambus. The SLDRAM solution is a natural evolution of the SDRAM. While many of today's SDRAM has function-assigned pins, SLDRAM uses a multiplexed command bus, which not only requires fewer pins but also achieves higher bandwidth via external speeds. For video, graphics, and telecommunications applications, SLDRAM provides multiple

independent banks, allowing fast read/write bus turnaround and the capability for small, fully pipelined bursts. In addition, the EmDRAM for system-on-chip (SOC) application has also exhibited significant potential in the DRAM architecture competition. Rather than emphasizing low cost and low voltage, the recent focus EmDRAM is high-speed operation [38].

## 7.11 Conclusions

This chapter provides a glimpse of the comprehensive and wide-ranging review of the past, present, and future development of low-voltage, low-power DRAM circuits. A myriad of eminent architectures of DRAM families ranging from the primitive Fast Page Mode DRAM (FPM DRAM) to the latest Embedded DRAM (EmDRAM) has been reviewed. An overview of the basic architecture of DRAM has been covered, followed by a detailed illustration on the additional peripheral circuits such as the Half-Voltage Generator (HVG), the Back-Bias Generator (BBG), and the Boosted Voltage Generator (BVG). Finally, the chapter ends with the future trend of DRAM, which presents the future limitations and challenges on further development of DRAM.

## References

1. A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, Toronto: Saunders College Publishing, 1998.
2. A. Bellaouar and M. I. Elmasry, *Low-Power Digital VLSI Design: Circuits and Systems*," The Netherlands: Kluwer Academic Publishers, 1995.
3. T. Sekiguchi, K. Itoh, T. Takahashi, M. Sugaya, H. Fujisawa, M. Nakamura, K. Kajigaya, and K. Kimura, "A Low-Impedance Open Bit-Line Array for Multigigabit DRAM," *IEEE J. Solid-State Circuits*, Vol. 37, No. 4, Apr. 2002, pp. 487–498.
4. K. N. Kim, H. S. Jeong, W. S. Yang, Y. S. Hwang, C. H. Cho, M. M. Jeong, S. Park, S. J. Ahn, Y. S. Chun, S. H. Shin, J. S. Park, S. H. Song, J. Y. Lee, S. M. Jang, C. H. Lee, J. H. Jeong, M. H. Cho, H. I. Yoon, and J. S. Jeon, "Highly Manufacturable and High-Performance SDR/DDR 4 Gb DRAM," *Symp. VLSI Technology Dig. Tech. Papers*, June 2001, pp. 7–8.
5. Y. Hongil, Y. S. Jae, S. L. Hyun, N. L. Kyu, Y. L. Jae, J. K. Nam, Y. K. Keum, M. B. Sang, S. Y. Won, H. C. Chang, S. J. Hong, H. Y. Jel, I. S. Dong, K. Kinam, I. B. Byung, and