CHAPTER 15

Memory Circuits

IN THIS CHAPTER YOU WILL LEARN:

1. How the basic bistable circuit, the latch, is realized by connecting two inverters in a positive-feedback loop.
2. How to augment the latch to obtain different types of flip-flops that are useful building blocks for digital systems.
3. How CMOS is particularly suited for the efficient implementation of a particular type of flip-flop, the *D* flip-flop.
4. How memory chips that contain as many as 4 gigabits are organized, as well as their various types and the terminology used to describe them.
5. The analysis and design of the six-transistor circuit that is almost universally used to implement the storage cell in static random access memory (SRAM) and the one-transistor circuit that is equally universal in the implementation of the storage cell in dynamic random access memory (DRAM).
6. Interesting circuit techniques for accessing a particular storage cell in a memory chip and for amplifying the signal readout from the cell.
7. How various types of read-only memory (ROM) are designed, programmed, erased, and reprogrammed.

Introduction

The logic circuits studied in Chapters 13 and 14 are called **combinational** circuits. Their output depends only on the present value of the input. Thus these circuits do *not* have memory. *Memory* is a very important part of digital systems. Its availability in digital computers allows for storing programs and data. Furthermore, it is important for temporary storage of the output produced by a combinational circuit for use at a later time in the operation of a digital system.

Logic circuits that incorporate memory are called **sequential circuits**; that is, their output depends not only on the present value of the input but also on the input's previous values. Such circuits require a timing generator (a *clock*) for their operation.

There are basically two approaches for providing memory to a digital circuit. The first relies on the application of positive feedback that, as will be seen shortly, can be arranged to provide a circuit with two stable states. Such a *bistable* circuit can then be used to store one bit of information: One stable state would correspond to a stored 0, and the other to a stored 1. A bistable circuit can remain in either state indefinitely, and thus it belongs to the category of *static sequential circuits*. The other approach to realizing memory utilizes the storage of charge on a capacitor: When the capacitor is charged, it would be regarded as storing a 1; when it is discharged, it would be storing a 0. Since the inevitable leakage effects will cause the capacitor to discharge, such a form of memory requires the periodic recharging of the capacitor, a process known as *refresh*. Thus, like dynamic logic (Section 14.3), memory based on charge storage is known as *dynamic memory* and the corresponding sequential circuits as *dynamic sequential circuits*.

This chapter is concerned with the study of memory circuits. We begin in Section 15.1 with the basic bistable circuit, the latch, and its application in flip-flops, an important class of building blocks for digital systems. After an overview of memory-chip types, organization, and nomenclature in Section 15.2, we study the circuit of the static memory cell (SRAM) and that of the dynamic memory cell (DRAM) in Section 15.3. Besides the array of storage cells, memory chips require circuits for selecting and accessing a particular cell in the array (address decoders) and for amplifying the signal that is retrieved from a particular cell (sense amplifiers). A sampling of these peripheral circuits is presented in Section 15.4. The chapter concludes with an important class of memories, the read-only memory (ROM) in Section 15.5.

15.1 Latches and Flip-Flops

In this section, we shall study the basic memory element, the latch, and consider a sampling of its applications. Both static and dynamic circuits will be considered.

15.1.1 The Latch

The basic memory element, the latch, is shown in Fig. 15.1(a). It consists of two cross-coupled logic inverters, G_1 and G_2. The inverters form a positive-feedback loop. To investigate the operation of the latch we break the feedback loop at the input of one of the inverters, say G_1, and apply an input signal, v_W, as shown in Fig. 15.1(b). Assuming that the input impedance of G_1 is large, breaking the feedback loop will not change the loop voltage transfer characteristic, which can be determined from the circuit of Fig. 15.1(b) by plotting v_Z versus v_W. This is the voltage transfer characteristic of two cascaded inverters and thus takes the shape shown in Fig. 15.1(c). Observe that the transfer characteristic consists of three segments, with the middle segment corresponding to the transition region of the inverters.

Also shown in Fig. 15.1(c) is a straight line with unity slope. This straight line represents the relationship $v_W = v_Z$ that is realized by reconnecting Z to W to close the feedback loop and thus to return it to its original form. As indicated, the straight line intersects the loop transfer curve at three points, A, B, and C. Thus any of these three points can serve as the operating point for the latch. We shall now show that while points A and C are stable operating points in the sense that the circuit can remain at either indefinitely, point B is an unstable operating point; the latch cannot operate at B for any significant period of time.

The reason point B is unstable can be seen by considering the latch circuit in Fig. 15.1(a) to be operating at point B, and taking account of the electrical interference (or noise) that is inevitably present in any circuit. Let the voltage v_W increase by a small increment v_w. The voltage at X will increase (in magnitude) by a larger increment, equal to the product of v_w and the incremental gain of G_1 at point B. The resulting signal v_x is applied to G_2 and gives rise to an even larger signal at node Z. The voltage v_z is related to the original increment v_w by the loop gain at point B, which is the slope of the curve of v_Z versus v_W at point B. This gain is usually much

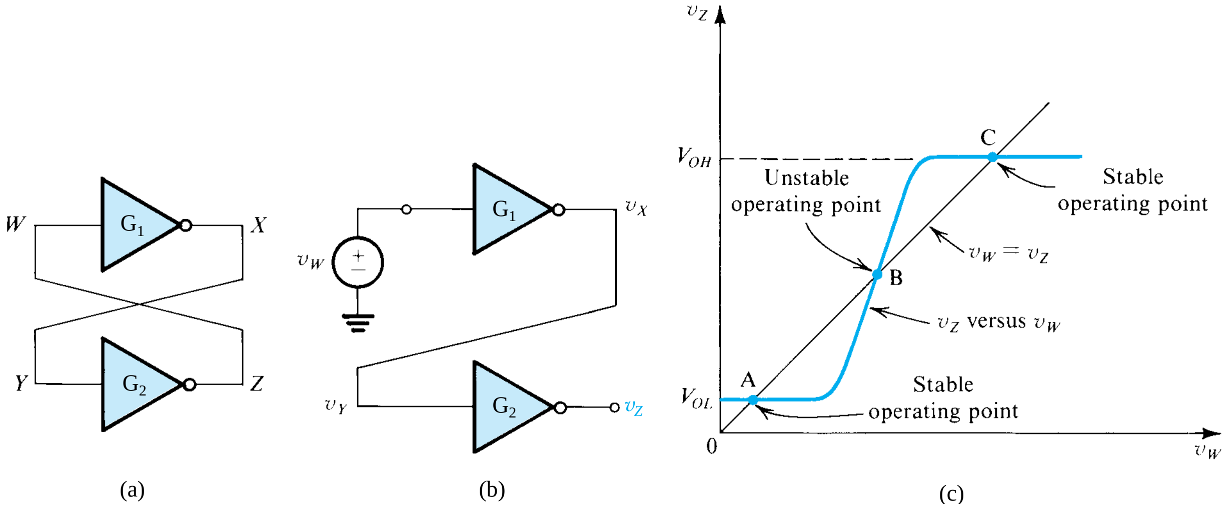


Figure 15.1 **(a)** Basic latch. **(b)** The latch with the feedback loop opened. **(c)** Determining the operating point(s) of the latch.

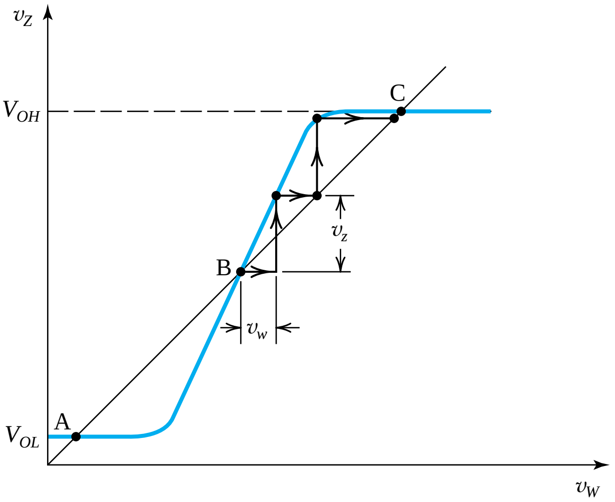


Figure 15.2 Point B is an unstable operating point for the latch: A small positive increment v_w gets amplified around the loop and causes the operating point to shift to the stable operating point C. Had v_w been negative, the operating point would have shifted to the other stable point, A.

greater than unity. Since v_z is coupled to the input of G_1, it becomes the new value of v_W and is further amplified by the loop gain. This regenerative process continues, shifting the operating point from B upward to point C, as illustrated in Fig. 15.2. Since at C the loop gain is zero (or almost zero), no regeneration can take place.

In the description above, we assumed arbitrarily an initial positive voltage increment at *W*. Had we instead assumed a negative voltage increment, we would have seen that the operating point moves downward from B to A. Again, since at point A the slope of the transfer curve is zero (or almost zero), no regeneration can take place. In fact, for regeneration to occur, the loop gain must be greater than unity, which is the case at point B.

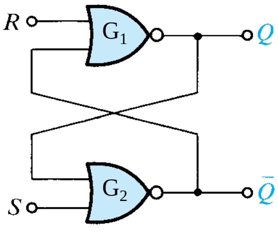
The discussion above leads us to conclude that the latch has two stable operating points, A and C. At point C, v_W is high, v_X is low, v_Y is low, and v_Z is high. The reverse is true at point A. If we consider X and Z as the latch outputs, we see that in one of the stable states (say that corresponding to operating point A), v_X is high (at V_{OH}) and v_Z is low (at V_{OL}). In the other state (corresponding to operating point C), v_X is low (at V_{OL}) and v_Z is high (at V_{OH}). Thus the latch is a **bistable** circuit having two complementary outputs. The stable state in which the latch operates depends on the external excitation that forces it to the particular state. The latch then *memorizes* this external action by staying indefinitely in the acquired state. As a memory element the latch is capable of storing one bit of information. For instance, we can arbitrarily designate the state in which v_X is high and v_Z is low as corresponding to a stored logic 1. The other complementary state then is designated by a stored logic 0. Finally, we note that the latch circuit described is of the static variety.

It now remains to devise a mechanism by which the latch can be *triggered* to change state. The latch together with the triggering circuitry forms a *flip-flop*. This will be discussed next. Analog bistable circuits utilizing op amps will be presented in Chapter 17.

15.1.2 The SR Flip-Flop

The simplest type of flip-flop is the set/reset (SR) flip-flop shown in Fig. 15.3(a). It is formed by cross-coupling two NOR gates, and thus it incorporates a latch. The second inputs of G$_1$ and G$_2$ together serve as the trigger inputs of the flip-flop. These two inputs are labeled S (for set) and R (for reset). The outputs are labeled Q and $\overline{Q}$, emphasizing their complementarity. The flip-flop is considered to be set (i.e., storing a logic 1) when Q is high and $\overline{Q}$ is low. When the flip-flop is in the other state (Q low, $\overline{Q}$ high), it is considered to be reset (storing a logic 0).

In the *rest* or *memory state* (i.e., when we do not wish to change the state of the flip-flop), both the S and R inputs should be low. Consider the case when the flip-flop is storing a logic 0. Since Q will be low, both inputs to the NOR gate G$_2$ will be low. Its output will therefore be high. This high is applied to the input of G$_1$, causing its output Q to be low, satisfying the original assumption. To set the flip-flop we raise S to the logic-1 level while leaving R at 0. The 1 at the S terminal will force the output of G$_2$, $\overline{Q}$, to 0. Thus the two inputs to G$_1$ will be 0 and its output Q will go to 1. Now even if S returns to 0, the $Q = 1$ signal fed to the input of G$_2$ will keep $\overline{Q} = 0$, and the flip-flop will remain in the newly acquired set state. Note that if we raise S to 1 again (with R remaining at 0), no change will occur. To reset the flip-flop we need to raise R to 1 while leaving $S = 0$. We can readily show that this forces the flip-flop into the reset state ($Q = 0$, $\overline{Q} = 1$) and that the flip-flop remains in this state even after R has returned to 0. It should be observed that the trigger signal merely starts the regenerative action of the positive-feedback loop of the latch.



(a)

R	S	Q_{n+1}
0	0	Q_n
0	1	1
1	0	0
1	1	Not used

(b)

Figure 15.3 **(a)** The set/reset (SR) flip-flop and **(b)** its truth table.

Finally, we inquire into what happens if both S and R are simultaneously raised to 1. The two NOR gates will cause both Q and $\overline{Q}$ to become 0 (note that in this case the complementary labeling of these two variables is incorrect). However, if R and S return to the rest state ($R = S = 0$) simultaneously, the state of the flip-flop will be undefined. In other words, it will be impossible to predict the final state of the flip-flop. For this reason, this input combination is usually disallowed (i.e., not used). Note, however, that this situation arises only in the idealized case, when both R and S return to 0 precisely simultaneously. In actual practice one of the two will return to 0 first, and the final state will be determined by the input that remains high longest.

The operation of the flip-flop is summarized by the *truth table* in Fig. 15.3(b), where Q_n denotes the value of Q at time t_n just before the application of the R and S signals, and Q_{n+1} denotes the value of Q at time t_{n+1} after the application of the input signals.

Rather than using two NOR gates, one can also implement an SR flip-flop by cross-coupling two NAND gates, in which case the set and reset functions are active when low (see Problem 15.2).

15.1.3 CMOS Implementation of SR Flip-Flops

The SR flip-flop of Fig. 15.3 can be directly implemented in CMOS by simply replacing each of the NOR gates by its CMOS circuit realization. We encourage the reader to sketch the resulting circuit (see Problem 15.1). Although the CMOS circuit thus obtained works well, it is somewhat complex. As an alternative, we consider a simplified circuit that furthermore implements additional logic. Specifically, Fig. 15.4 shows a *clocked* version of an SR flip-flop. Since the clock inputs form AND functions with the set and reset inputs, the flip-flop can be set or reset only when the clock ϕ is high. Observe that although the two cross-coupled inverters at the heart of the flip-flop are of the standard CMOS type, only NMOS transistors are used for the set–reset circuitry. Nevertheless, since there is no conducting path between V_{DD} and ground (except during switching), the circuit does not dissipate any static power.

Except for the addition of clocking, the SR flip-flop of Fig. 15.4 operates in exactly the same way as its logic antecedent in Fig. 15.3: To illustrate, consider what happens when the flip-flop is in the reset state ($Q = 0$, $\overline{Q} = 1$, $v_Q = 0$, $v_{\overline{Q}} = V_{DD}$), and assume that we wish to set

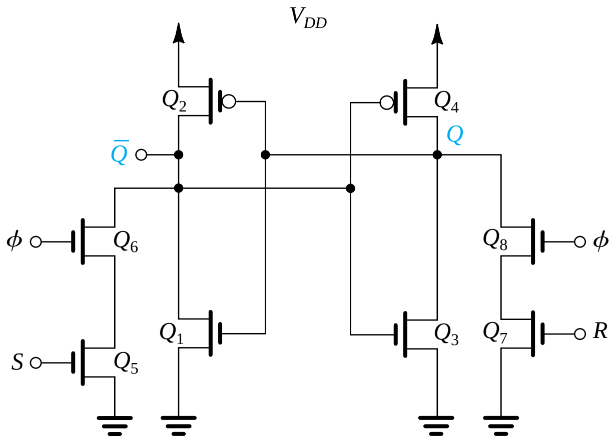


Figure 15.4 CMOS implementation of a clocked SR flip-flop. The clock signal is denoted by ϕ.

it. To do so, we arrange for a high (V_{DD}) signal to appear on the S input while R is held low at 0 V. Then, when the clock ϕ goes high, both Q_5 and Q_6 will conduct, pulling the voltage $v_{\overline{Q}}$ down. If $v_{\overline{Q}}$ goes below the threshold V_M of the (Q_3, Q_4) inverter, the inverter will switch states (or at least begin to switch states), and its output v_Q will rise. This increase in v_Q is fed back to the input of the (Q_1, Q_2) inverter, causing its output $v_{\overline{Q}}$ to go down even further; the regeneration process, characteristic of the positive-feedback latch, is now in progress.

The preceding description of flip-flop switching is predicated on two assumptions:

1. Transistors Q_5 and Q_6 supply sufficient current to pull the node $\overline{Q}$ down to a voltage at least slightly below the threshold of the (Q_3, Q_4) inverter. This is essential for the regenerative process to begin. Without this initial trigger, the flip-flop will fail to switch. In Example 15.1, we shall investigate the *minimum* W/L ratios that Q_5 and Q_6 must have to meet this requirement.
2. The set signal remains high for an interval long enough to cause regeneration to take over the switching process. An estimate of the minimum width required for the set pulse can be obtained as the sum of the interval during which $v_{\overline{Q}}$ is reduced from V_{DD} to $V_{DD}/2$, and the interval for the voltage v_Q to respond and rise to $V_{DD}/2$. This point also will be illustrated in Example 15.1.

Finally, note that the symmetry of the circuit indicates that all the preceding remarks apply equally well to the reset process.

Example 15.1

The CMOS SR flip-flop in Fig. 15.4 is fabricated in a 0.18-µm process for which $\mu_n C_{ox} = 4\ \mu_p C_{ox} = 300\ \mu\text{A/V}^2$, $V_{tn} = |V_{tp}| = 0.5$ V, and $V_{DD} = 1.8$ V. The inverters have $(W/L)_n = 0.27$ µm/0.18 µm and $(W/L)_p = 4\,(W/L)_n$. The four NMOS transistors in the set–reset circuit have equal W/L ratios.

(a) Determine the minimum value required for this ratio to ensure that the flip-flop will switch.

(b) Also, determine the minimum width the set pulse must have for the case in which the W/L ratio of each of the four transistors in the set–reset circuit is selected at twice the minimum value found in (a). Assume that the total capacitance between each of the Q and $\overline{Q}$ nodes and ground is 20 fF.

Solution

(a) Figure 15.5(a) shows the relevant portion of the circuit for our present purposes. Observe that since the circuit is in the reset state and regeneration has not yet begun, we assume that $v_Q = 0$ and thus Q_2 will be conducting. The circuit is in effect a pseudo-NMOS gate, and our task is to select the W/L ratios for Q_5 and Q_6 so that V_{OL} of this inverter is lower than $V_{DD}/2$ (the threshold of the Q_3, Q_4 inverter whose Q_N and Q_P are matched). The minimum required W/L for Q_5 and Q_6 can be found by equating the current supplied by Q_5 and Q_6 to the current supplied by Q_2 at $v_{\overline{Q}} = V_{DD}/2$. To simplify matters, we assume that the series connection of Q_5 and Q_6 is equivalent to a single transistor whose W/L is half the W/L of each of Q_5 and Q_6 (Fig. 15.5b). Now, since at $v_{\overline{Q}} = V_{DD}/2$ both this equivalent transistor and Q_2 will be operating in the triode region, we can write

$$I_{D\text{eq}} = I_{D2}$$

$$300 \times \frac{1}{2}\left(\frac{W}{L}\right)_5 \left[(1.8 - 0.5)\left(\frac{1.8}{2}\right) - \frac{1}{2}\left(\frac{1.8}{2}\right)^2\right]$$

$$= 75 \times \frac{1.08}{0.18}\left[(1.8 - 0.5)\left(\frac{1.8}{2}\right) - \frac{1}{2}\left(\frac{1.8}{2}\right)^2\right]$$

which yields

$$\left(\frac{W}{L}\right)_5 = \frac{0.54\ \mu\text{m}}{0.18\ \mu\text{m}}$$

and thus

$$\left(\frac{W}{L}\right)_6 = \frac{0.54\ \mu\text{m}}{0.18\ \mu\text{m}}$$

(b) The value calculated for $(W/L)_5$ and $(W/L)_6$ is the absolute minimum needed for switching to occur. To guarantee that the flip-flop will switch, the value selected for $(W/L)_5$ and $(W/L)_6$ is usually somewhat larger than the minimum. Selecting a value twice the minimum,

$$(W/L)_5 = (W/L)_6 = 1.08\ \mu\text{m}/0.18\ \mu\text{m}$$

The minimum required width of the set pulse is composed of two components: the time for $v_{\bar{Q}}$ in the circuit of Fig. 15.5(a) to fall from V_{DD} to $V_{DD}/2$, where $V_{DD}/2$ is the threshold voltage of the inverter formed by Q_3 and Q_4 in Fig. 15.4, and the time for the output of the Q_3–Q_4 inverter to rise from 0 to $V_{DD}/2$. At the end of the second time interval, the feedback signal will have traveled around the feedback loop, and regeneration can continue without the presence of the set pulse. We will denote the first component t_{PHL} and the second t_{PLH}, and will calculate their values as follows.

To determine t_{PHL} refer to the circuit in Fig. 15.6 and note that the capacitor discharge current i_C is the difference between the current of the equivalent transistor Q_{eq} and the current of Q_2,

$$i_C = i_{D\text{eq}} - i_{D2}$$

To determine the average discharge current i_C, we calculate $i_{D\text{eq}}$ and i_{D2} at $t = 0$ and $t = t_{PHL}$. At $t = 0$, $v_{\bar{Q}} = V_{DD}$, thus Q_2 is off,

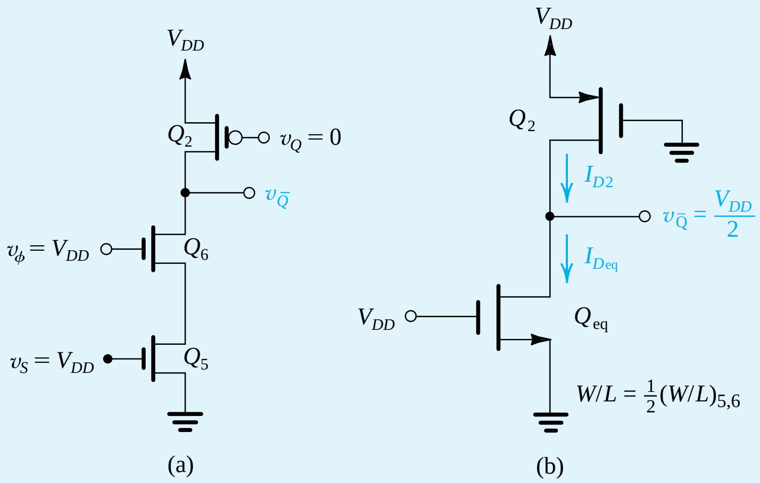


Figure 15.5 (**a**) The relevant portion of the flip-flop circuit of Fig. 15.4 for determining the minimum W/L ratios of Q_5 and Q_6 needed to ensure that the flip-flop will switch. (**b**) The circuit in (**a**) with Q_5 and Q_6 replaced with their equivalent transistor Q_{eq}, at the point of switching.

Example 15.1 *continued*

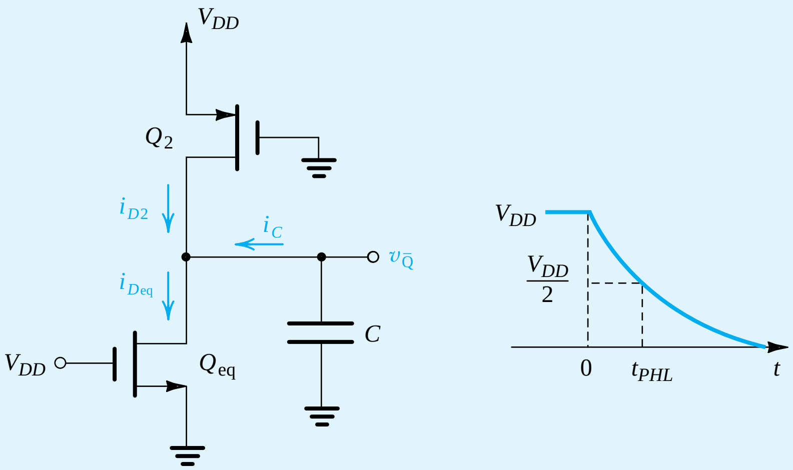


Figure 15.6 Determining the time t_{PHL} for $v_{\bar{Q}}$ to fall from V_{DD} to $V_{DD}/2$.

$$i_{D2}(0) = 0$$

and Q_{eq} is in saturation,

$$\begin{aligned} i_{Deq} &= \frac{1}{2} \times 300 \times \frac{1}{2} \times \frac{1.08}{0.18} \times (1.8 - 0.5)^2 \\ &= 760.5\ \mu\text{A} \end{aligned}$$

Thus,

$$i_C(0) = 760.5 - 0 = 760.5\ \mu\text{A}$$

At $t = t_{PHL}$, $v_{\bar{Q}} = V_{DD}/2$, thus both Q_2 and Q_{eq} will be in the triode region,

$$\begin{aligned} i_{D2}(t_{PHL}) &= 75 \times \frac{1.08}{0.18} \times \left[(1.8 - 0.5) - 0.5\left(\frac{1.8}{2}\right)^2\right] \\ &= 344.25\ \mu\text{A} \end{aligned}$$

and

$$\begin{aligned} i_{Deq}(t_{PHL}) &= 300 \times \frac{1}{2} \times \frac{1.08}{0.18}\left[(1.8 - 0.5)\left(\frac{1.8}{2}\right) - 0.5\left(\frac{1.8}{2}\right)^2\right] \\ &= 688.5\ \mu\text{A} \end{aligned}$$

Thus,

$$i_C(t_{PHL}) = 688.5 - 344.25 = 344.25\ \mu\text{A}$$

and the average value of i_C over the interval $t = 0$ to $t = t_{PHL}$ is

$$\begin{aligned} i_C\big|_{\text{av}} &= \frac{i_C(0) + i_C(t_{PHL})}{2} \\ &= \frac{760.5 + 344.25}{2} = 552.4\ \mu\text{A} \end{aligned}$$

We now can calculate t_{PHL} as

$$t_{PHL} = \frac{C(V_{DD}/2)}{i_C|_{\text{av}}} = \frac{20 \times 10^{-15} \times 0.9}{552.4 \times 10^{-6}} = 32.6 \text{ ps}$$

Next we consider the time t_{PHL} for the output of the Q_3–Q_4 inverter, v_Q, to rise from 0 to $V_{DD}/2$. The value of t_{PLH} can be calculated using the propagation delay formula derived in Chapter 13 (Eq. 13.66), which is also listed in Table 13.3, namely,

$$t_{PLH} = \frac{\alpha_p C}{k'_p (W/L)_p V_{DD}}$$

where

$$\alpha_p = 2 \Bigg/ \left[\frac{7}{4} - \frac{3|V_{tp}|}{V_{DD}} + \left(\frac{|V_{tp}|}{V_{DD}}\right)^2\right]$$

Substituting numerical values we obtain,

$$\alpha_p = \frac{2}{1.75 - \frac{3 \times 0.5}{1.8} + \left(\frac{0.5}{1.8}\right)^2} = 2.01$$

and

$$t_{PLH} = \frac{2.01 \times 20 \times 10^{-15}}{75 \times 10^{-6} \times (1.08/0.18) \times 1.8} = 49.7 \text{ ps}$$

Finally, the minimum required width of the set pulse can be calculated as

$$T_{\min} = t_{PHL} + t_{PLH}$$

EXERCISE

15.1 For the SR flip-flop specified in Example 15.1, find the minimum W/L for both Q_5 and Q_6 so that switching is achieved when inputs S and ϕ are at $(V_{DD}/2)$.
Ans. 14.3

15.1.4 A Simpler CMOS Implementation of the Clocked SR Flip-Flop

A simpler implementation of a clocked SR flip-flop is shown in Fig. 15.7. Here, pass-transistor logic is employed to implement the clocked set–reset functions. This circuit is very popular in the design of static random-access memory (SRAM) chips, where it is used as the basic memory cell (Section 15.4.1).

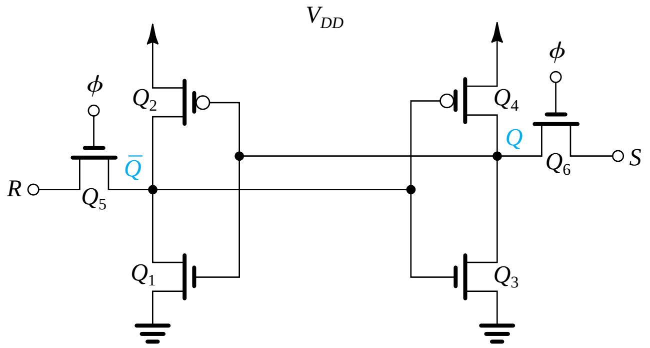


Figure 15.7 A simpler CMOS implementation of the clocked SR flip-flop. This circuit is popular as the basic cell in the design of static random-access memory (SRAM) chips.

15.1.5 D Flip-Flop Circuits

A variety of flip-flop types exist and can be synthesized using logic gates. CMOS circuit implementations can be obtained by simply replacing the gates with their CMOS circuit realizations. This approach, however, usually results in rather complex circuits. In many cases, simpler circuits can be found by taking a circuit-design viewpoint, rather than a logic-design one. To illustrate this point, we shall consider the CMOS implementation of a very important type of flip-flop, the data, or D, flip-flop.

The D flip-flop is shown in block diagram form in Fig. 15.8. It has two inputs, the data input D and the clock input ϕ. The complementary outputs are labeled Q and $\overline{Q}$. When the clock is low, the flip-flop is in the memory, or rest, state; signal changes on the D input line have no effect on the state of the flip-flop. As the clock goes high, the flip-flop acquires the logic level that existed on the D line just before the rising edge of the clock. Such a flip-flop is said to be **edge triggered**. Some implementations of the D flip-flop include direct set and reset inputs that override the clocked operation just described.

A simple implementation of the D flip-flop is shown in Fig. 15.9. The circuit consists of two inverters connected in a positive-feedback loop, just as in the static latch of Fig. 15.1(a), except that here the loop is closed for only part of the time. Specifically, the loop is closed when the clock is low $(\phi = 0, \overline{\phi} = 1)$. The input D is connected to the flip-flop through a switch that closes when the clock is high. Operation is straightforward: When ϕ is high, the loop is opened, and the input D is connected to the input of inverter G_1. The capacitance at the input node of G_1 is charged to the value of D, and the capacitance at the input node of G_2 is charged to the value of $\overline{D}$. Then, when the clock goes low, the input line is isolated from

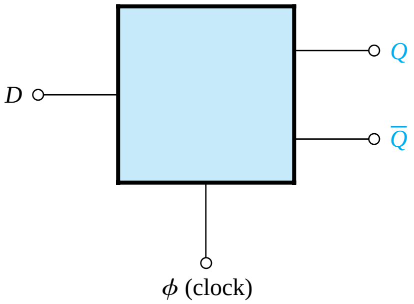


Figure 15.8 A block diagram representation of the D flip-flop.

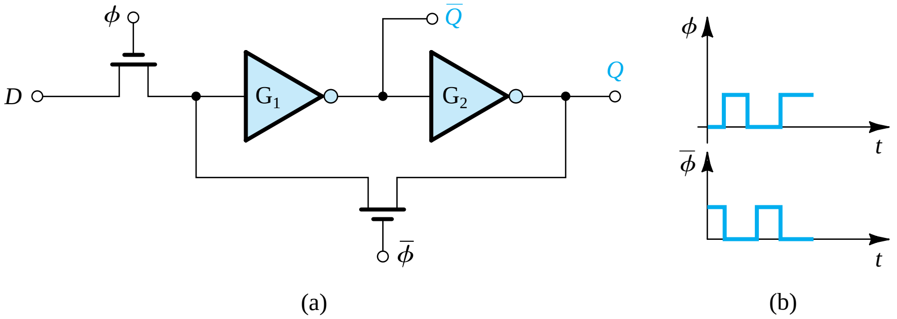


Figure 15.9 A simple implementation of the D flip-flop. The circuit in **(a)** utilizes the two-phase **non-overlapping clock** whose waveforms are shown in **(b).**

the flip-flop, the feedback loop is closed, and the latch acquires the state corresponding to the value of D just before ϕ went down, providing an output $Q = D$.

From the preceding, we observe that the circuit in Fig. 15.9 combines the positive-feedback technique of static bistable circuits and the charge-storage technique of dynamic circuits. It is important to note that the proper operation of this circuit, and of many circuits that use clocks, is predicated on the assumption that *ϕ and $\overline{\phi}$ will not be simultaneously high at any time*. This condition is defined by referring to the two clock phases as being *nonoverlapping*.

An inherent drawback of the D flip-flop implementation of Fig. 15.9 is that during ϕ, the output of the flip-flop simply follows the signal on the D input line. This can cause problems in certain logic-design situations. The problem is solved very effectively by using the **master–slave** configuration shown in Fig. 15.10(a). Before discussing its circuit operation, we note that although the switches are shown implemented with single NMOS transistors, CMOS transmission gates are employed in many applications. We are simply using the single MOS transistor as a "shorthand notation" for a series switch.

The master–slave circuit consists of a pair of circuits of the type shown in Fig. 15.9, operated with alternate clock phases. Here, to emphasize that the two clock phases must be nonoverlapping, we denote them ϕ_1 and ϕ_2, and clearly show the nonoverlap interval in the waveforms of Fig. 15.10(b). Operation of the circuit is as follows:

1. When ϕ_1 is high and ϕ_2 is low, the input is connected to the master latch whose feedback loop is opened, while the slave latch is isolated. Thus, the output Q remains at the value stored previously in the slave latch whose loop is now closed. The node capacitances of the master latch are charged to the appropriate voltages corresponding to the present value of D.
2. When ϕ_1 goes low, the master latch is isolated from the input data line. Then, when ϕ_2 goes high, the feedback loop of the master latch is closed, locking in the value of D. Further, its output is connected to the slave latch whose feedback loop is now open. The node capacitances in the slave are appropriately charged so that when ϕ_1 goes high again, the slave latch locks in the new value of D and provides it at the output, $Q = D$.

From this description, we note that at the positive transition of clock ϕ_2 the output Q adopts the value of D that existed on the D line at the end of the preceding clock phase, ϕ_1. This output value remains constant for one clock period. Finally, note that during the nonoverlap interval both latches have their feedback loops open, and we are relying on the node capacitances to maintain most of their charge. It follows that the nonoverlap interval should be kept reasonably short (perhaps one-tenth or less of the clock period, and of the order of 1 ns or so in current practice).

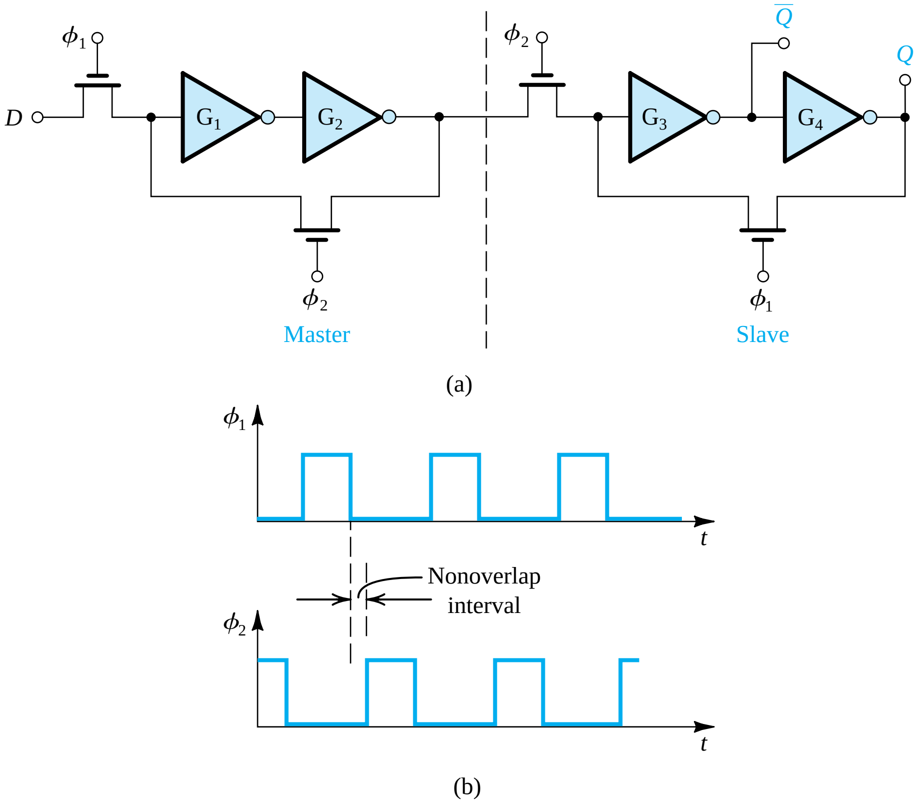


Figure 15.10 **(a)** A master–slave D flip-flop. The switches can be, and usually are, implemented with CMOS transmission gates. **(b)** Waveforms of the two-phase nonoverlapping clock required.

15.2 Semiconductor Memories: Types and Architectures

A computer system, whether a large machine or a microcomputer, requires memory for storing data and program instructions. Furthermore, within a given computer system there usually are various types of memory utilizing a variety of technologies and having different *access times*. Broadly speaking, computer memory can be divided into two types: **main memory** and **mass-storage** memory. The main memory is usually the most rapidly accessible memory and the one from which most, often all, instructions in programs are executed. The main memory is usually of the random-access type. A **random-access memory** (RAM) is one in which the time required for storing (writing) information and for retrieving (reading) information is independent of the physical location (within the memory) in which the information is stored.

Random-access memories should be contrasted with *serial* or *sequential* memories, such as disks and tapes, from which data are available only in the sequence in which the data were originally stored. Thus, in a serial memory the time to access particular information depends on the memory location in which the required information is stored, and the average access time is longer than the access time of random-access memory. In a computer system, serial memory is used for mass storage. Items not frequently accessed, such as large parts of the computer operating system, are usually stored in a *moving-surface memory* such as magnetic disk.

Another important classification of memory relates to whether it is a **read/write** or a **read-only memory**. Read/write (R/W) memory permits data to be stored and retrieved at comparable speeds. Computer systems require random-access read/write memory for data and program storage.

Read-only memories (**ROM**) permit reading at the same high speeds as R/W memories (or perhaps higher) but restrict the writing operation. ROMs can be used to store a microprocessor operating-system program. They are also employed in operations that require table lookup, such as finding the values of mathematical functions. A popular application of ROMs is their use in video game cartridges. It should be noted that read-only memory is usually of the random-access type. Nevertheless, in the digital circuit jargon, the acronym RAM usually refers to read/write, random-access memory, while ROM is used for read-only memory.

The regular structure of memory circuits has made them an ideal application for the design of circuits of the very-large-scale integrated (VLSI) type. Indeed, at any moment, memory chips represent the state of the art in packing density and hence integration level. Beginning with the introduction of the 1-Kbit chip in 1970, memory-chip density has quadrupled about every 3 years. At the present time (2009), chips containing 4 Gbit[1] are available. In this and the next two sections, we shall study some of the basic circuits employed in VLSI RAM chips. Read-only memory circuits are studied in Section 15.5.

15.2.1 Memory-Chip Organization

The bits on a memory chip are addressable either individually or in groups of 4 to 16. As an example, a 64-Mbit chip in which all bits are individually addressable is said to be organized as 64M words × 1 bit (or simply 64M × 1). Such a chip needs a 26-bit address (2^{26} = 67,108,864 = 64M). On the other hand, the 64-Mbit chip can be organized as 16M words × 4 bits (16M × 4), in which case a 24-bit address is required. For simplicity we shall assume in our subsequent discussion that all the bits on a memory chip are individually addressable.

The bulk of the memory chip consists of the cells in which the bits are stored. Each **memory cell** is an electronic circuit capable of storing one bit. We shall study memory-cell circuits in Section 15.3. For reasons that will become clear shortly, it is desirable to physically organize the storage cells on a chip in a square or a nearly square matrix. Figure 15.11 illustrates such an organization. The cell matrix has 2^M rows and 2^N columns, for a total storage capacity of 2^{M+N}. For example, a 1M-bit square matrix would have 1024 rows and 1024 columns ($M = N = 10$). Each cell in the array is connected to one of the 2^M row lines, known rather loosely, but universally, as **word lines**, and to one of the 2^N column lines, known as **digit lines** or, more commonly, **bit lines**. A particular cell is **selected** for reading or writing by activating its word line and its bit line.

Activating one of the 2^M word lines is performed by the **row decoder**, a combinational logic circuit that selects (raises the voltage of) the particular word line whose M-bit address is applied to the decoder input. The address bits are denoted $A_0, A_1, \ldots, A_{M-1}$. When the Kth word line is activated for, say, a **read operation**, all 2^N cells in row K will provide their contents to their respective bit lines. Thus, if the cell in column L (Fig. 15.11) is storing a 1, the voltage of bit-line number L will be raised, usually by a small voltage, say 0.1 V to 0.2 V. The readout voltage

[1] The capacity of a memory chip to hold binary information as binary digits (or bits) is measured in kilobit (Kbit), megabit (Mbit), and gigabit (Gbit) units, where 1 Kbit = 1024 bits, 1 Mbit = 1024 × 1024 = 1,048,576 bits, and, 1 Gbit = 1024^3 bits. Thus a 64-Mbit chip contains 67,108,864 bits of memory.

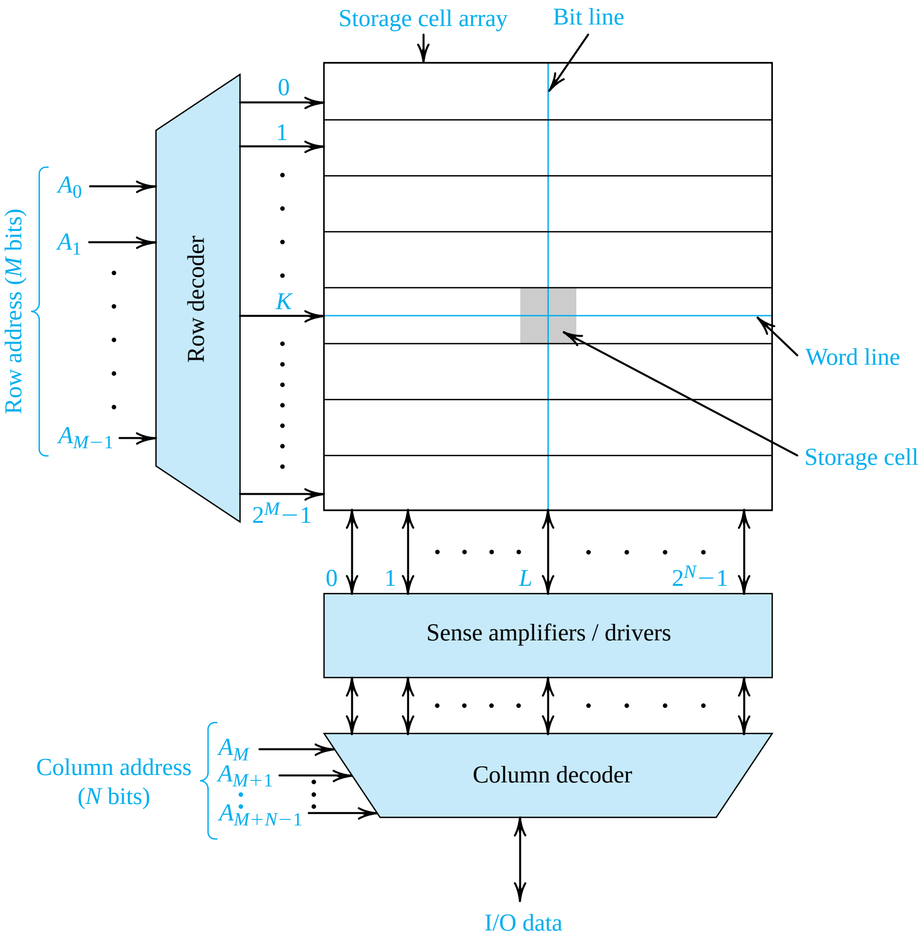


Figure 15.11 A 2^{M+N}-bit memory chip organized as an array of 2^M rows × 2^N columns.

is small because the cell is small, a deliberate design decision, since the number of cells is very large. The small readout signal is applied to a **sense amplifier** connected to the bit line. As Fig. 15.11 indicates, there is a sense amplifier for every bit line. The sense amplifier provides a full-swing digital signal (from 0 to V_{DD}) at its output. This signal, together with the output signals from all the other cells in the selected row, is then delivered to the **column decoder**. The column decoder selects the signal of the particular column whose N-bit address is applied to the decoder input (the address bits are denoted $A_M, A_{M+1}, \ldots, A_{M+N-1}$) and causes this signal to appear on the chip input/output (I/O) data line.

A **write operation** proceeds in a similar manner: The data bit to be stored (1 or 0) is applied to the I/O line. The cell in which the data bit is to be stored is selected through the combination of its row address and its column address. The sense amplifier of the selected column acts as a **driver** to write the applied signal into the selected cell. Circuits for sense amplifiers and address decoders will be studied in Section 15.4.

Before leaving the topic of memory organization (or memory-chip architecture), we wish to mention a relatively recent innovation in organization dictated by the exponential increase in chip density. To appreciate the need for a change, note that as the number of cells in the array increases, the physical lengths of the word lines and the bit lines increase. This has occurred even though for each new generation of memory chips, the transistor size has decreased (currently, CMOS process technologies with 45-nm feature size are utilized). The net increase in word-line and bit-line lengths increases their total resistance and capacitance, and thus slows down their transient response. That is, as the lines lengthen, the exponential rise of the voltage of the word line becomes slower, and it takes longer for the cells to be activated. This problem has been solved by partitioning the memory chip into a number of blocks. Each of the blocks has an organization identical to that in Fig. 15.11. The row and column addresses are broadcast to all blocks, but the data selected come from only one of the blocks. Block selection is achieved by using an appropriate number of the address bits as a block address. Such an architecture can be thought of as three-dimensional: rows, columns, and blocks.

15.2.2 Memory-Chip Timing

The **memory access time** is the time between the initiation of a read operation and the appearance of the output data. The **memory cycle time** is the minimum time allowed between two consecutive memory operations. To be on the conservative side, a memory operation is usually taken to include both read and write (in the same location). MOS memories have access and cycle times in the range of a few to a few hundred nanoseconds.

EXERCISES

15.2 A 4-Mbit memory chip is partitioned into 32 blocks, with each block having 1024 rows and 128 columns. Give the number of bits required for the row address, column address, and block address.
Ans. 10; 7; 5

15.3 The word lines in a particular MOS memory chip are fabricated using polysilicon (see Appendix A). The resistance of each word line is estimated to be 5 kΩ, and the total capacitance between the line and ground is 2 pF. Find the time for the voltage on the word line to reach $V_{DD}/2$, assuming that the line is driven by a voltage V_{DD} provided by a low-impedance inverter. (*Note:* The line is actually a distributed network that we are approximating by a lumped circuit consisting of a single resistor and a single capacitor.)
Ans. 6.9 ns

15.3 Random-Access Memory (RAM) Cells

As mentioned in Section 15.2, the major part of the memory chip is taken up by the storage cells. It follows that to be able to pack a large number of bits on a chip, it is imperative that the cell size be reduced to the smallest possible. The power dissipation per cell should be minimized also. Thus, many of the flip-flop circuits studied in Section 15.1 are too complex to be suitable for implementing the storage cells in a RAM chip.

There are basically two types of MOS RAM: static and dynamic. **Static RAMs** (called **SRAMs** for short) utilize static latches as the storage cells. Dynamic RAMs (called **DRAMs**), on the other hand, store the binary data on capacitors, resulting in further reduction in cell area, but at the expense of more complex read and write circuitry. In particular, while static RAMs can hold their stored data indefinitely, provided the power supply remains on, dynamic RAMs require *periodic refreshing* to regenerate the data stored on capacitors. This is because the storage capacitors will discharge, though slowly, as a result of the leakage currents inevitably present. By virtue of their smaller cell size, dynamic memory chips are usually four times as dense as their contemporary static chips. Thus while the state of the art in 2009 is a 4-Gbit DRAM chip, the highest-density SRAM chip has 1 Gbit capacity. Both static and dynamic RAMs are *volatile;* that is, they require the continuous presence of a power supply. By contrast, most ROMs are of the nonvolatile type, as we shall see in Section 15.5. In the following subsections, we shall study basic SRAM and DRAM storage cells.

15.3.1 Static Memory (SRAM) Cell

Figure 15.12 shows a typical static memory cell in CMOS technology. The circuit, which we encountered in Section 15.1, is a flip-flop comprising two cross-coupled inverters and two **access transistors**, Q_5 and Q_6. The access transistors are turned on when the word line is selected and its voltage raised to V_{DD}, and they connect the flip-flop to the column (bit or B) line and $\overline{\text{column}}$ ($\overline{\text{bit}}$ or $\overline{B}$) line. Note that although in principle only the B or the $\overline{B}$ line suffices, most often both are utilized, as shown in Fig. 15.12. This both provides a *differential data path* between the cell and the memory-chip output and increases the circuit reliability. The access transistors act as transmission gates allowing bidirectional current flow between the flip-flop and the B and $\overline{B}$ lines. Finally, we note that this circuit is known as the **six-transistor** or **6T cell**.

The Read Operation Consider first a read operation, and assume that the cell is storing a 1. In this case, Q will be high at V_{DD}, and $\overline{Q}$ will be low at 0 V. Before the read operation begins, the B and $\overline{B}$ lines are raised to a voltage in the range $V_{DD}/2$ to V_{DD}. This process, known as **precharging**, is performed using circuits we shall discuss in the next section in conjunction with the study of sense amplifiers. To simplify matters, we shall assume here that the precharge voltage of B and $\overline{B}$ is V_{DD}.

When the word line is selected and the access transistors Q_5 and Q_6 are turned on, examination of the circuit reveals that the only portion that will be conducting is that shown in Fig. 15.13. Noting that the initial value of $v_{\overline{Q}}$ is 0 V, we can see that current will flow from the $\overline{B}$

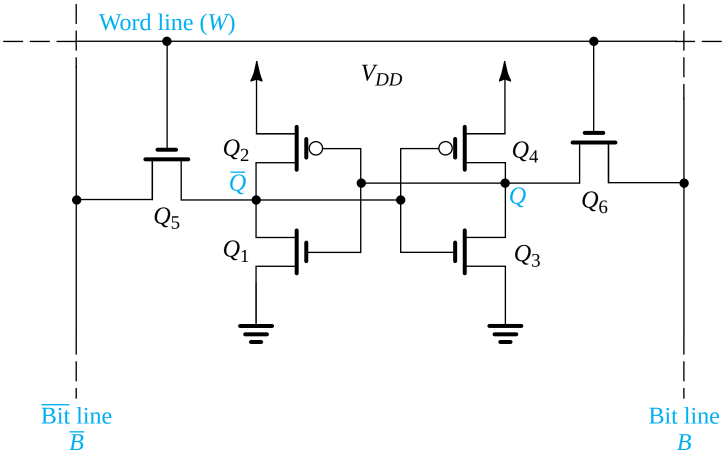


Figure 15.12 A CMOS SRAM memory cell.

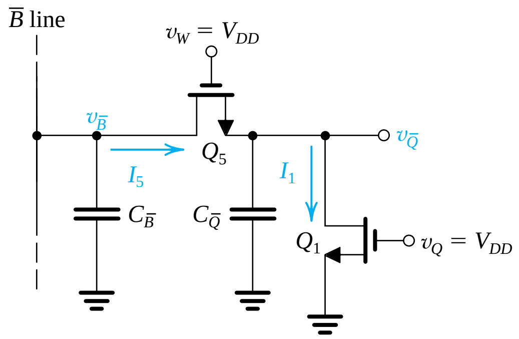


Figure 15.13 Relevant parts of the SRAM cell circuit during a read operation when the cell is storing a logic 1. Note that initially $v_Q = V_{DD}$ and $v_{\overline{Q}} = 0$. Also note that the B and $\overline{B}$ lines are precharged to a voltage V_{DD}.

line (actually, from the $\overline{B}$-line capacitance $C_{\overline{B}}$) through Q_5 and into capacitor $C_{\overline{Q}}$, which is the small equivalent capacitance between the $\overline{Q}$ node and ground. This current charges $C_{\overline{Q}}$ and thus $v_{\overline{Q}}$ rises and Q_1 conducts, sinking some of the current supplied by Q_5. Equilibrium will be reached when $C_{\overline{Q}}$ is charged to a voltage $V_{\overline{Q}}$ at which I_1 equals I_5, and no current flows through $C_{\overline{Q}}$. Here it is extremely important to note that to avoid changing the state of the flip-flop, that is, for our read operation to be **nondestructive**, $V_{\overline{Q}}$ must not exceed the threshold voltage of the inverter Q_3–Q_4. In fact, SRAM designers usually impose a more stringent requirement on the value of $V_{\overline{Q}}$, namely, that it should be lower than the threshold voltage of Q_3, V_{tn}. Thus, the design problem we shall now solve is as follows: Determine the ratio of $(W/L)_5/(W/L)_1$ so that $V_{\overline{Q}} \leq V_{tn}$.

Noting that Q_5 will be operating in saturation and neglecting, for simplicity, the body effect, we can write

$$I_5 = \frac{1}{2}(\mu_n C_{ox})\left(\frac{W}{L}\right)_5 (V_{DD} - V_{tn} - V_{\overline{Q}})^2 \tag{15.1}$$

Transistor Q_1 will be operating in the triode region, and its current I_1 can be written as

$$I_1 = (\mu_n C_{ox})\left(\frac{W}{L}\right)_1 \left[(V_{DD} - V_{tn})V_{\overline{Q}} - \frac{1}{2}V_{\overline{Q}}^2\right] \tag{15.2}$$

Equating I_5 and I_1 gives a quadratic equation in $V_{\overline{Q}}$, which can be solved to obtain

$$\frac{V_{\overline{Q}}}{V_{DD} - V_{tn}} = 1 - 1\Big/\sqrt{1 + \frac{(W/L)_5}{(W/L)_1}} \tag{15.3}$$

This is an attractive relationship, since it provides $V_{\overline{Q}}$ in normalized form and thus always applies, independent of the process technology utilized. Figure 15.14 shows a universal plot of $[V_{\overline{Q}}/(V_{DD} - V_{tn})]$ versus $(W/L)_5/(W/L)_1$. For a given process technology, V_{DD} and V_{tn} are determined, and the plot in Fig. 15.14 can be used to determine the maximum value permitted for $(W/L)_5/(W/L)_1$ while keeping $V_{\overline{Q}}$ below a desired value. Alternatively, we

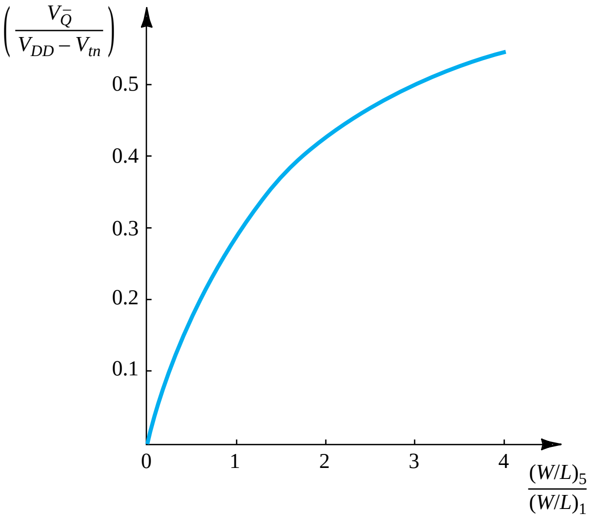


Figure 15.14 The normalized value of $V_{\bar{Q}}$ versus the ratio $(W/L)_5/(W/L)_1$ for the circuit in Fig. 15.13. This graph can be used to determine the maximum value permitted for $(W/L)_5/(W/L)_1$ so that $V_{\bar{Q}}$ is kept below a desired level.

can derive a formula for this purpose. For instance, if $V_{\bar{Q}}$ is to be kept below V_{tn}, the ratio of $(W/L)_5$ to $(W/L)_1$ must be kept below the value obtained from Eq. (15.3), that is,

$$\frac{(W/L)_5}{(W/L)_1} \leq \frac{1}{\left(1-\frac{V_{tn}}{V_{DD}-V_{tn}}\right)^2} - 1 \tag{15.4}$$

This is an important design constraint that can be expressed in a slightly more general form by replacing $(W/L)_5$ with $(W/L)_a$, where the subscript *a* denotes access transistors Q_5 and Q_6, and $(W/L)_1$ with $(W/L)_n$, which is the *W/L* ratio of Q_N in each of the two inverters; thus,

$$\frac{(W/L)_a}{(W/L)_n} \leq \frac{1}{\left(1-\frac{V_{tn}}{V_{DD}-V_{tn}}\right)^2} - 1 \tag{15.5}$$

EXERCISE

15.4 Find the maximum allowable *W/L* for the access transistors of the SRAM cell in Fig. 15.12 so that in a read operation, the voltages at Q and $\bar{Q}$ do not change by more than $|V_t|$. Assume that the SRAM is fabricated in a 0.18-μm technology for which $V_{DD} = 1.8$ V, $V_{tn} = |V_{tp}| = 0.5$ V and that $(W/L)_n = 1.5$.
Ans. $(W/L)_a \leq 2.5$

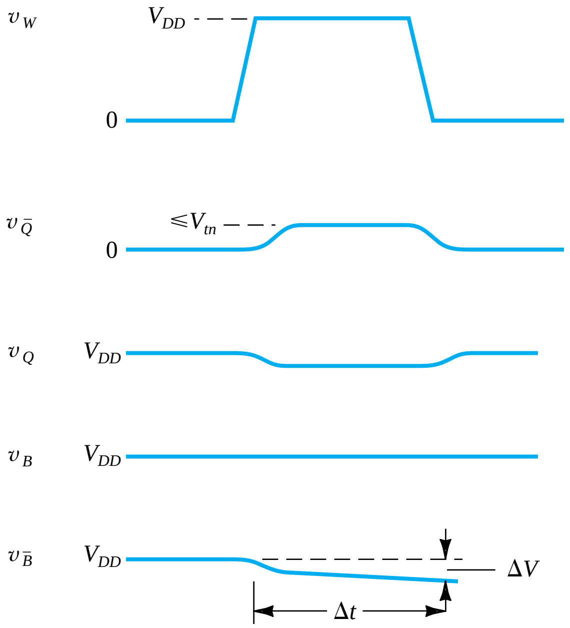


Figure 15.15 Voltage waveforms at various nodes in the SRAM cell during a read-1 operation.

Having determined the constraint imposed by the read operation on the *W*/*L* ratios of the access transistors, we now return to the circuit in Fig. 15.13, and show in Fig. 15.15 the voltage waveforms at various nodes during a read-1 operation. Observe that as we have already discussed, $v_{\bar{Q}}$ rises from zero to a voltage $V_{\bar{Q}} \leq V_{tn}$. Correspondingly, the change in v_Q will be very small, justifying the assumption implicit in the analysis above that v_Q remains constant at V_{DD}. Most important, note that the voltage of the $\bar{B}$ line, $v_{\bar{B}}$, decreases by a small amount ΔV. This is a result of the discharge of the capacitance of the $\bar{B}$ line, $C_{\bar{B}}$, by the current I_5. Assuming that I_5 reaches its equilibrium value in Eq. (15.1) relatively quickly, capacitor $C_{\bar{B}}$ is in effect discharged by a constant current I_5 and the change in its voltage, ΔV, obtained in a time interval Δt, can be found by writing a charge-balance equation,

$$I_5 \ \Delta t = C_{\bar{B}} \ \Delta V$$

Thus,

$$\Delta V = \frac{I_5 \Delta t}{C_{\bar{B}}} \tag{15.6}$$

Here we note that $C_{\bar{B}}$ is usually relatively large (1–2 pF) because a large number of cells are connected to the $\bar{B}$ line. The incremental change ΔV is therefore rather small (0.1–0.2 V), necessitating the use of a sense amplifier. If the sense amplifier requires a minimum decrement ΔV in $v_{\bar{B}}$ to detect the presence of a "1", then the read delay time can be found from Eq. (15.6) as

$$\Delta t = \frac{C_{\bar{B}} \ \Delta V}{I_5} \tag{15.7}$$

This equation indicates the need for a relatively large I_5 to reduce the delay time Δt. A large I_5, however, implies selecting $(W/L)_a$ near the upper bound given by Eq. (15.5), which in turn means an increase in the silicon area occupied by the access transistors and hence the cell area, an interesting design trade-off.

EXERCISE

15.5 For the SRAM cell considered in Exercise 15.4 whose $(W/L)_n = 1.5$ and $(W/L)_a \leq 2.5$, use Eq. (15.7) to determine the read delay Δt in two cases: (a) $(W/L)_a = 2.5$ and (b) $(W/L)_a = 1.5$. Let $\mu_n C_{ox} = 300\ \mu\text{A/V}^2$. In both cases, assume that $C_{\bar{B}} = 2$ pF and that the sense amplifier requires a ΔV of minimum magnitude of 0.2 V. [*Hint*: Use Eq. (15.1) to determine I_5, and recall that $V_{\bar{Q}} = V_{tn}$.]
Ans. 1.7 ns; 2.8 ns

We conclude our discussion of the read operation with two remarks:

1. Although we considered only the read-1 operation, the read-0 operation is identical; it involves Q_2 and Q_6 with the analysis resulting in an upper bound on $(W/L)_6/(W/L)_2$ equal to that we have found for $(W/L)_5/(W/L)_1$. This, of course, is entirely expected, since the circuit is symmetrical. The read-0 operation results in a decrement ΔV in the voltage of the B line, which is interpreted by the sense amplifier as a stored 0.
2. The component Δt of the read delay is relatively large because C_B and $C_{\bar{B}}$ are relatively large (in the picofarad range). Also, Δt is not the only component of the read delay; another significant component is due to the finite rise time of the voltage on the word line. Indeed, even the calculation of Δt is optimistic, since the word line will have reached a voltage lower than V_{DD} only, when the process of discharging $C_{\bar{B}}$ takes place. As will be seen shortly, the write operation is faster.

The Write Operation We next consider the write operation. Let the SRAM cell of Fig. 15.12 be storing a logic 1, thus $V_Q = V_{DD}$ and $V_{\bar{Q}} = 0$ V, and assume that we wish to write a 0; that is, we wish to have the flip-flop switch states. To write a zero, the B line is lowered to 0 V, and the $\bar{B}$ line is raised to V_{DD} and, of course, the cell is selected by raising the word line to V_{DD}. The objective now is to pull node Q down and node $\bar{Q}$ up and have the voltage of at least one of these two nodes pass by the inverter threshold voltage. Thus, if v_Q decreases below the threshold voltage of inverter Q_1–Q_2, the regenerative action of the latch will start and the flip-flop will switch to the stored-0 state. Alternatively, or in addition, if we manage to raise $v_{\bar{Q}}$ above the threshold voltage of the Q_3–Q_4 inverter, the regenerative action will be engaged and the latch will eventually switch state. Either one of the two actions is sufficient to engage the regenerative mechanism of the latch.

Figure 15.16 shows the relevant parts of the SRAM circuit during the interval when $v_{\bar{Q}}$ is being pulled up (Fig. 15.16a) and v_Q is being pulled down (Fig. 15.16b). Since **toggling** (i.e., state change) has not yet taken place, we assume that the voltage feeding the gate of Q_1 is still equal to V_{DD} and the voltage at the gate of Q_4 is still equal to 0 V. These voltages will of course be changing as $v_{\bar{Q}}$ goes up and v_Q goes down, but this assumption is nevertheless reasonable for hand analysis.

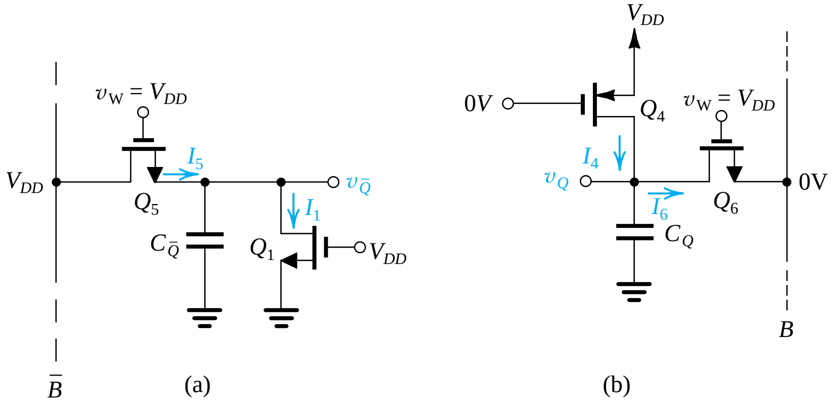


Figure 15.16 Relevant parts of the 6T SRAM circuit of Fig. 15.12 during the process of writing a 0. It is assumed that the cell is originally storing a 1 and thus initially $v_Q = V_{DD}$ and $v_{\bar{Q}} = 0$ V.

Consider first the circuit in Fig. 15.16(a). This is the same circuit we analyzed in detail in the study of the read operation above. Recall that to make the read process nondestructive, we imposed an upper bound on $(W/L)_5$. That upper bound ensured that $v_{\bar{Q}}$ will not rise above V_{tn}. Thus, this circuit is not capable of raising $v_{\bar{Q}}$ to the point that it can start the regenerative action. We must therefore rely solely on the circuit of Fig. 15.16(b). That is, our write-0 operation will be accomplished by pulling node Q down in order to initiate the regenerative action of the latch. To ensure that the latch will in fact switch state, SRAM designers impose a more stringent requirement on the voltage v_Q, namely, that it must fall below not just V_M of the $Q_1 - Q_2$ inverter but below V_{tn} of Q_1.

Let's now look more closely at the circuit of Fig. 15.16(b). Initially, v_Q is at V_{DD}. However, as Q_6 turns on, I_6 quickly discharges the small capacitance C_Q, and v_Q begins to fall. This will enable Q_4 to conduct, and equilibrium is reached when $I_4 = I_6$. To ensure toggling, we design the circuit so that this equilibrium occurs at a value of v_Q less than V_{tn}. At such a value V_Q, Q_4 will be operating in saturation and Q_6 will be operating in the triode region, thus

$$I_4 = \frac{1}{2}(\mu_p C_{ox})\left(\frac{W}{L}\right)_4 (V_{DD} - |V_{tp}|)^2 \tag{15.8}$$

and

$$I_6 = (\mu_n C_{ox})\left(\frac{W}{L}\right)_6 \left[(V_{DD} - V_{tn})V_Q - \frac{1}{2}V_Q^2\right] \tag{15.9}$$

Substituting $|V_{tp}| = V_{tn}$, which is usually the case, and equating I_4 and I_6 results in a quadratic equation in V_Q whose solution is

$$\frac{V_Q}{V_{DD} - V_{tn}} = 1 - \sqrt{1 - \left(\frac{\mu_p}{\mu_n}\right)\frac{(W/L)_4}{(W/L)_6}} \tag{15.10}$$

This relationship is not as convenient as that in Eq. (15.3) because the right-hand side includes a process-dependent quantity, namely, μ_p/μ_n. Thus we do not have a universally

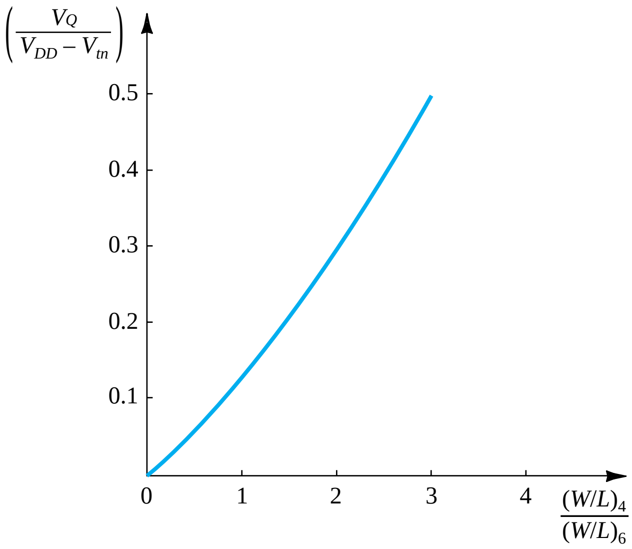


Figure 15.17 The normalized value of V_Q versus the ratio $(W/L)_4/(W/L)_6$ for the circuit in Fig. 15.16(b). The graph applies for process technologies for which $\mu_n \simeq 4\mu_p$. It can be used to determine the maximum $(W/L)_4/(W/L)_6$ for which V_Q is guaranteed to fall below a desired value.

applicable relationship. Nevertheless, for a number of CMOS process technologies, including the 0.25-µm, the 0.18-µm, and the 0.13-µm processes, $\mu_n/\mu_p \simeq 4$. Thus, upon substituting $\mu_p/\mu_n = 0.25$ in Eq. (15.10), we obtain the semiuniversal graph shown in Fig. 15.17. We can use this graph to determine the maximum allowable value of the ratio $(W/L)_4/(W/L)_6$ that will ensure a value of $V_Q \leq V_{tn}$ for given process parameters V_{DD} and V_{tn}. Alternatively, substituting $V_Q = V_{tn}$, $(W/L)_4 = (W/L)_p$, and $(W/L)_6 = (W/L)_a$, we can obtain the upper bound analytically as

$$\frac{(W/L)_p}{(W/L)_a} \leq \left(\frac{\mu_n}{\mu_p}\right)\left[1-\left(1-\frac{V_{tn}}{V_{DD}-V_{tn}}\right)^2\right] \tag{15.11}$$

Observe that this relationship provides an upper bound on $(W/L)_p$ in terms of $(W/L)_a$ and that the relationship in Eq. (15.5) provides an upper bound on $(W/L)_a$ in terms of $(W/L)_n$. Thus, the two relationships can be used together to design the SRAM cell.

EXERCISE

15.6 For the SRAM cell considered in Exercise 15.4, where $(W/L)_n = 1.5$ and $(W/L)_a \leq 2.5$, use Eq. (15.11) to find the maximum allowable value of $(W/L)_p$. Recall that for this 0.18-µm process, $\mu_p \simeq 4\mu_n$. For all transistors having $L = 0.18$ µm, find W_n, W_p, and W_a that result in a minimum-area cell. Assume that the minimum allowable width is 0.18 µm.
Ans. $(W/L)_p \leq 2.5\,(W/L)_a$, thus $(W/L)_p \leq 6.25$; for minimum area select $W_n = W_p = W_a = 0.18$ µm.

We conclude our study of the write process by noting that it is fast because it does not require discharging the large capacitance of the bit lines. The voltages of the B and $\bar{B}$ lines are driven to their required values of 0 or V_{DD} by powerful driver circuits and thus achieve their desired voltages very quickly. The write delay is determined roughly by the time for the regenerating signal to propagate around the feedback loop of the latch; thus it is about twice the propagation delay of the inverter. Of course, the write cycle time is still lengthened by the word-line delay.

15.3.2 Dynamic Memory Cell

Although a variety of DRAM storage cells have been proposed over the years, a particular cell, shown in Fig. 15.18, has become the industry standard. The cell consists of a single *n*-channel MOSFET, known as the **access transistor**, and a **storage capacitor** C_S. The cell is appropriately known as the **one-transistor cell**.[2] The gate of the transistor is connected to the word line, and its source (drain) is connected to the bit line. Observe that only one bit line is used in DRAMs, whereas in SRAMs both the bit and $\overline{\text{bit}}$ lines are utilized. The DRAM cell stores its bit of information as charge on the cell capacitor C_S. When the cell is storing a 1, the capacitor is charged to V_{DD}; when a 0 is stored, the capacitor is discharged to zero volts.

Some explanation is needed to appreciate how the capacitor can be charged to the full supply voltage V_{DD}. Consider a write-1 operation. The word line is at V_{DD} and the bit line is at V_{DD} and the transistor is conducting, charging C_S. The transistor will cease conduction when the voltage on C_S reaches $(V_{DD} - V_t)$. This is the same problem we encountered with pass-transistor logic (PTL) in Section 14.2. The problem is overcome in DRAM design by boosting the word line to a voltage equal to $V_{DD} + V_t$. In this case the capacitor voltage for a stored 1 will be equal to the full V_{DD}. However, because of leakage effects, the capacitor charge will leak off, and hence the cell must be refreshed periodically. During **refresh**, the cell content is read and the data bit is rewritten, thus *restoring* the capacitor voltage to its proper value. Typically, the refresh operation must be performed every 5 ms to 10 ms.

Let us now consider the DRAM operation in more detail. As in the static RAM, the row decoder selects a particular row by raising the voltage of its word line. This causes all the

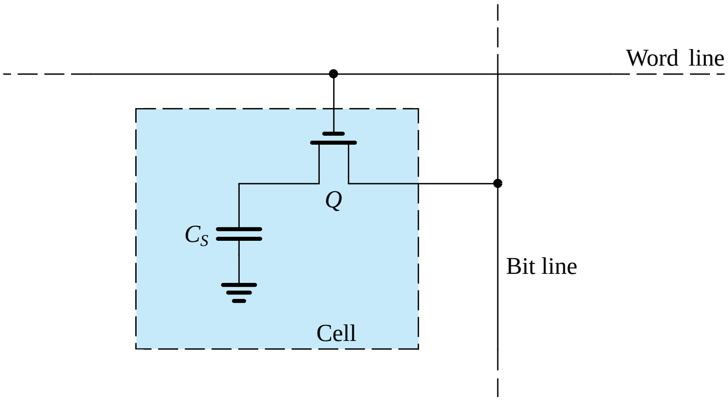


Figure 15.18 The one-transistor dynamic RAM (DRAM) cell.

[2] The name was originally used to distinguish this cell from earlier ones utilizing three transistors.

Figure 15.19 When the voltage of the selected word line is raised, the transistor conducts, thus connecting the storage capacitor C_S to the bit-line capacitance C_B.

access transistors in the selected row to become conductive, thereby connecting the storage capacitors of all the cells in the selected row to their respective bit lines. Thus the cell capacitor C_S is connected in parallel with the bit-line capacitance C_B, as indicated in Fig. 15.19. Here, it should be noted that C_S is typically 20 fF to 30 fF, whereas C_B is 10 times larger. Now, if the operation is a read, the bit line is precharged to $V_{DD}/2$. To find the change in the voltage on the bit line resulting from connecting a cell capacitor C_S to it, let the initial voltage on the cell capacitor be V_{CS} ($V_{CS} = V_{DD}$ when a 1 is stored, and $V_{CS} = 0$ V when a 0 is stored). Using charge conservation, we can write

$$C_S V_{CS} + C_B \frac{V_{DD}}{2} = (C_B + C_S)\left(\frac{V_{DD}}{2} + \Delta V\right)$$

from which we can obtain for ΔV

$$\Delta V = \frac{C_S}{C_B + C_S}\left(V_{CS} - \frac{V_{DD}}{2}\right) \tag{15.12}$$

and since $C_B \gg C_S$,

$$\Delta V \simeq \frac{C_S}{C_B}\left(V_{CS} - \frac{V_{DD}}{2}\right) \tag{15.13}$$

Now, if the cell is storing a 1, $V_{CS} = V_{DD}$, and

$$\Delta V(1) \simeq \frac{C_S}{C_B}\left(\frac{V_{DD}}{2}\right) \tag{15.14}$$

whereas if the cell is storing a 0, $V_{CS} = 0$, and

$$\Delta V(0) \simeq -\frac{C_S}{C_B}\left(\frac{V_{DD}}{2}\right) \tag{15.15}$$

Since usually C_B is much greater than C_S, these readout voltages are very small. For example, for $C_B = 10\ C_S$, $V_{DD} = 1.8$ V, $\Delta V(0)$ will be about –90 mV, and $\Delta V(1)$ will be +90 mV. This is a best-case scenario, for the 1 level in the cell might very well be below V_{DD}. Furthermore, in modern memory chips, V_{DD} is 1.2 V or even lower. In any case, we see that a stored 1 in the cell results in a small positive increment in the bit-line voltage, whereas a stored zero results in a small negative increment. Observe also that the readout process is *destructive,* since the resulting voltage across C_S will no longer be V_{DD} or 0.

The change of voltage on the bit line is detected and amplified by the column sense amplifier causing the bit line to be driven to the full scale value (0 or V_{DD}) of the detected signal. This amplified signal is then impressed on the storage capacitor, thus restoring its signal to the proper level (V_{DD} or 0). In this way, all the cells in the selected row are refreshed. Simultaneously, the signal at the output of the sense amplifier of the selected column is fed to the data-output line of the chip through the action of the column decoder.

The write operation proceeds similarly to the read operation, except that the data bit to be written, which is impressed on the data input line, is applied by the column decoder to the selected bit line. Thus, if the data bit to be written is a 1, the *B*-line voltage is raised to V_{DD} (i.e., C_B is charged to V_{DD}). When the access transistor of the particular cell is turned on, its capacitor C_S will be charged to V_{DD}; thus a 1 is written in the cell. Simultaneously, all the other cells in the selected row are simply refreshed.

Although the read and write operations result in automatic refreshing of all the cells in the selected row, provision must be made for the periodic refreshing of the entire memory, typically every 5 to 10 ms, as specified for the particular chip. The refresh operation is carried out in a *burst mode,* one row at a time. During refresh, the chip will not be available for read or write operations. This is not a serious matter, however, since the interval required to refresh the entire chip is typically less than 2% of the time between refresh cycles. In other words, the memory chip is available for normal operation more than 98% of the time.

EXERCISES

15.7 In a particular dynamic memory chip, $C_S = 30$ fF, $C_B = 0.3$ pF, and $V_{DD} = 1.2$ V. Find the output readout voltage for a stored 1 and a stored 0. Recall that in a read operation, the bit lines are precharged to $V_{DD}/2$.
Ans. 60 mV; –60 mV

15.8 A 64-Mbit DRAM chip fabricated in a 0.4-μm CMOS technology requires 2 μm^2 per cell. If the storage array is square, estimate its dimensions. Further, if the peripheral circuitry (e.g., sense amplifiers, decoders) adds about 30% to the chip area, estimate the dimensions of the resulting chip.
Ans. 11.6 mm × 11.6 mm; 13.2 mm × 13.2 mm

15.4 Sense Amplifiers and Address Decoders

Having studied the circuits commonly used to implement the storage cells in SRAMs and DRAMs, we now consider some of the other important circuit blocks in a memory chip. The design of these circuits, commonly referred to as the **memory peripheral circuits**, presents exciting challenges and opportunities to integrated-circuit designers: Improving the performance of peripheral circuits can result in denser and faster memory chips that dissipate less power.

15.4.1 The Sense Amplifier

Next to the storage cells, the sense amplifier is the most critical component in a memory chip. Sense amplifiers are essential to the proper operation of DRAMs, and their use in SRAMs results in speed and area improvements.

A variety of sense-amplifier designs are in use, some of which closely resemble the active-load MOS differential amplifier studied in Chapter 8. Here, we first describe a differential sense amplifier that employs positive feedback. Because the circuit is differential, it can be employed directly in SRAMs, where the SRAM cell utilizes both the B and $\overline{B}$ lines. On the other hand, the one-transistor DRAM circuit we studied in Section 15.3.2 is a single-ended circuit, utilizing one bit line only. The DRAM circuit, however, can be made to resemble a differential signal source through the use of the "dummy-cell" technique, which we shall discuss shortly. Therefore, we shall assume that the memory cell whose output is to be amplified develops a difference output voltage between the B and $\overline{B}$ lines. This signal, which can range from 30 mV to 500 mV depending on the memory type and cell design, will be applied to the input terminals of the sense amplifier. The sense amplifier in turn responds by providing a full-swing (0 to V_{DD}) signal at its output terminals. The particular amplifier circuit we shall discuss here has a rather unusual property: *Its output and input terminals are the same!*

A Sense Amplifier with Positive Feedback Figure 15.20 shows the sense amplifier together with some of the other column circuitry of a RAM chip. Note that the sense amplifier is nothing but the familiar latch formed by cross-coupling two CMOS inverters: One inverter

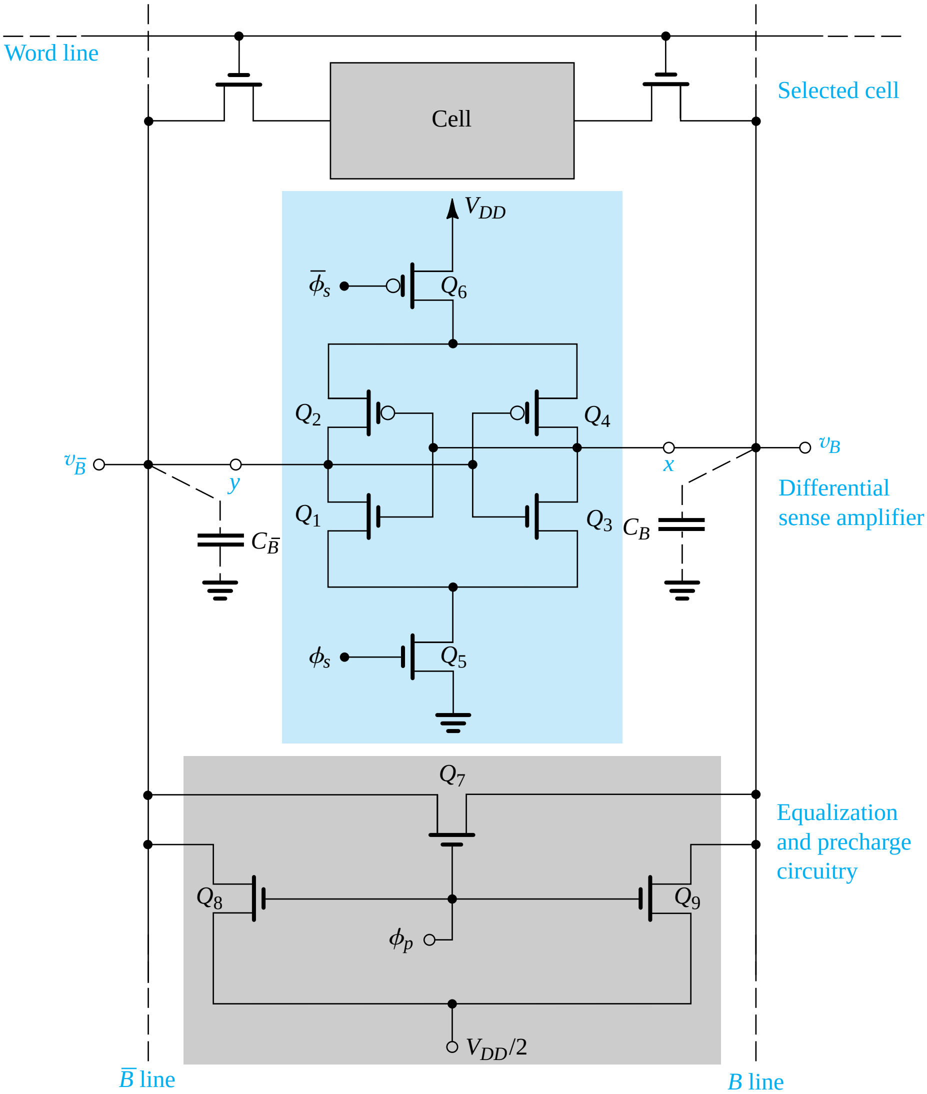


Figure 15.20 A differential sense amplifier connected to the bit lines of a particular column. This arrangement can be used directly for SRAMs (which utilize both the B and $\overline{B}$ lines). DRAMs can be turned into differential circuits by using the "dummy-cell" arrangement shown later (Fig. 15.22).

is implemented by transistors Q_1 and Q_2, and the other by transistors Q_3 and Q_4. Transistors Q_5 and Q_6 act as switches that connect the sense amplifier to ground and V_{DD} only when data-sensing action is required. Otherwise, ϕ_s is low and the sense amplifier is turned off. This conserves power, an important consideration because usually there is one sense amplifier per column, resulting in *thousands of sense amplifiers per chip*. Note, again, that terminals x and y are both the input and the output terminals of the amplifier. As indicated, these I/O terminals are connected to the B and $\overline{B}$ lines. The amplifier is required to detect a small signal appearing between B and $\overline{B}$, and to amplify it to provide a full-swing signal at B and $\overline{B}$. For instance, if during a read operation, the cell has a stored 1, then a small positive voltage will develop between B and $\overline{B}$, with v_B higher than $v_{\overline{B}}$. The amplifier will then cause v_B to rise to V_{DD} and $v_{\overline{B}}$ to fall to 0 V. This 1 output is then directed to the chip I/O pin by the column decoder (not shown) and at the same time is used to rewrite a 1 in the DRAM cell, thus performing the restore operation that is required because the DRAM readout process is destructive.

Figure 15.20 also shows the precharge and equalization circuit. Operation of this circuit is straightforward: When ϕ_p goes high (to V_{DD}) prior to a read operation, all three transistors conduct. While Q_8 and Q_9 precharge the $\overline{B}$ and B lines to $V_{DD}/2$, transistor Q_7 helps speed up this process by equalizing the initial voltages on the two lines. This equalization is critical to the proper operation of the sense amplifier. Any voltage difference present between B and $\overline{B}$ prior to commencement of the read operation can result in erroneous interpretation by the sense amplifier of its input signal. In Fig. 15.20, we show only one of the cells in this particular column, namely, the cell whose word line is activated. The cell can be either an SRAM or a DRAM cell. All other cells in this column will not be connected to the B and $\overline{B}$ lines (because their word lines will remain low).

Let us now consider the sequence of events during a read operation:

1. The precharge and equalization circuit is activated by raising the control signal ϕ_p. This will cause the B and $\overline{B}$ lines to be at equal voltages, equal to $V_{DD}/2$. The clock ϕ_p then goes low, and the B and $\overline{B}$ lines are left to float for a brief interval.

2. The word line goes up, connecting the cell to the B and $\overline{B}$ lines. A voltage then develops between B and $\overline{B}$, with v_B higher than $v_{\overline{B}}$ if the accessed cell is storing a 1, or v_B lower than $v_{\overline{B}}$ if the cell is storing a 0. To keep the cell area small, and to facilitate operation at higher speeds, the readout signal, which the cell is required to provide between B and $\overline{B}$, is kept small (typically, 30–500 mV).

3. Once an adequate difference voltage signal has been developed between B and $\overline{B}$ by the storage cell, the sense amplifier is turned on by connecting it to ground and V_{DD} through Q_5 and Q_6, activated by raising the sense-control signal ϕ_s. Because initially the input terminals of the inverters are at $V_{DD}/2$, the inverters will be operating in their transition region where the gain is high (Section 13.2). It follows that initially the latch will be operating at its unstable equilibrium point. Thus, depending on the signal between the input terminals, the latch will quickly move to one of its two stable equilibrium points (refer to the description of the latch operation in Section 15.1). This is achieved by the regenerative action, inherent in positive feedback. Figure 15.21 clearly illustrates this point by showing the waveforms of the signal on the bit line for both a read-1 and a read-0 operation. Observe that once activated, the sense amplifier causes the small initial difference, $\Delta V(1)$ or $\Delta V(0)$, provided by the cell, to grow exponentially to either V_{DD} (for a read-1 operation) or 0 (for a read-0 operation). The waveforms of the signal on the $\overline{B}$ line will be complementary to those shown in Fig. 15.21 for the B line. In the following, we quantify the process of exponential growth of v_B and $v_{\overline{B}}$.

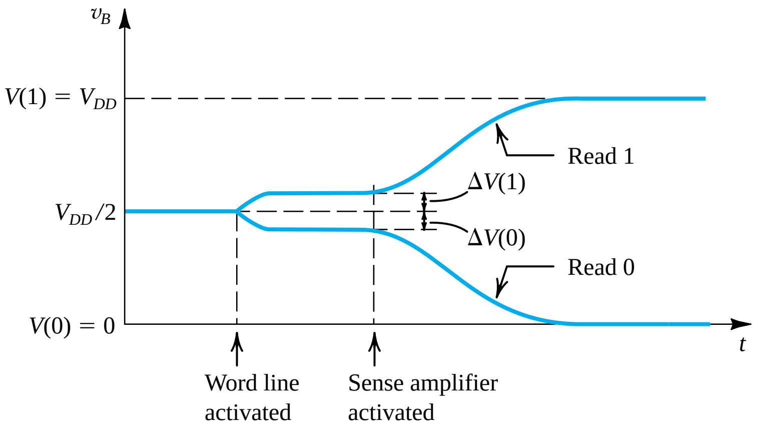


Figure 15.21 Waveforms of v_B before and after the activation of the sense amplifier. In a read-1 operation, the sense amplifier causes the initial small increment $\Delta V(1)$ to grow exponentially to V_{DD}. In a read-0 operation, the negative $\Delta V(0)$ grows to 0. Complementary signal waveforms develop on the $\overline{B}$ line.

A Closer Look at the Operation of the Sense Amplifier Developing a precise expression for the output signal of the sense amplifier shown in Fig. 15.20 is a rather complex task requiring the use of large-signal (and thus nonlinear) models of the inverter voltage transfer characteristic, as well as taking the positive feedback into account. We will not do this here; rather, we shall consider the operation in a semiquantitative way.

Recall that at the time the sense amplifier is activated, each of its two inverters is operating in the transition region near $V_{DD}/2$. Thus, for small-signal operation, each inverter can be modeled using g_{mn} and g_{mp}, the transconductances of Q_N and Q_P, respectively, evaluated at an input bias of $V_{DD}/2$. Specifically, a small-signal v_i superimposed on $V_{DD}/2$ at the input of one of the inverters gives rise to an inverter output current signal of $(g_{mn} + g_{mp})\, v_i \equiv G_m v_i$. This output current is delivered to one of the capacitors, C_B or $C_{\overline{B}}$. The voltage thus developed across the capacitor is then fed back to the other inverter and is multiplied by its G_m, which gives rise to an output current feeding the other capacitor, and so on, in a regenerative process. The positive feedback in this loop will mean that the signal around the loop, and thus v_B and $v_{\overline{B}}$, will *rise or decay exponentially* (see Fig. 15.21) with a time constant of (C_B/G_m) [or $(C_{\overline{B}}/G_m)$, since we have been assuming $C_B = C_{\overline{B}}$]. Thus, for example, in a read-1 operation we obtain

$$v_B = \frac{V_{DD}}{2} + \Delta V(1) e^{G_m/C_B t}, \qquad v_B \le V_{DD} \tag{15.16}$$

whereas in a read-0 operation,

$$v_B = \frac{V_{DD}}{2} - \Delta V(0) e^{(G_m/C_B)t} \tag{15.17}$$

Because these expressions have been derived assuming small-signal operation, they describe the exponential growth (decay) of v_B reasonably accurately only for values close to $V_{DD}/2$. Nevertheless, they can be used to obtain a reasonable estimate of the time required to develop a particular signal level on the bit line.

Example 15.2

Consider the sense-amplifier circuit of Fig. 15.20 during the reading of a 1. Assume that the storage cell provides a voltage increment on the *B* line of $\Delta V(1) = 0.1$ V. If the NMOS devices in the amplifiers have $(W/L)_n = 0.54\ \mu\text{m}/0.18\ \mu\text{m}$ and the PMOS devices have $(W/L)_p = 2.16\ \mu\text{m}/0.18\ \mu\text{m}$, and assuming that $V_{DD} = 1.8$ V, $V_{tn} = |V_{tp}| = 0.5$ V, and $\mu_n C_{ox} = 4\,\mu_p C_{ox} = 300\ \mu\text{A/V}^2$, find the time required for v_B to reach 0.9 V_{DD}. Assume $C_B = 1$ pF.

Solution

First, we determine the transconductances g_{mn} and g_{mp}

$$g_{mn} = \mu_n C_{ox}\left(\frac{W}{L}\right)_n (V_{GS} - V_t)$$

$$= 300 \times \frac{0.54}{0.18}(0.9 - 0.5)$$

$$= 0.36\ \text{mA/V}$$

$$g_{mp} = \mu_p C_{ox}\left(\frac{W}{L}\right)_p (V_{GS} - |V_t|)$$

$$= 75 \times \frac{2.16}{0.18}(0.9 - 0.5) = 0.36\ \text{mA/V}$$

Thus, the inverter G_m is

$$G_m = g_{mn} + g_{mp} = 0.72\ \text{mA/V}$$

and the time constant τ for the exponential growth of v_B will be

$$\tau \equiv \frac{C_B}{G_m} = \frac{1 \times 10^{-12}}{0.72 \times 10^{-3}} = 1.4\ \text{ns}$$

Now, the time, Δt, for v_B to reach 0.9 V_{DD} can be determined from

$$0.9 \times 1.8 = 0.9 + 0.1 e^{\Delta t/1.4}$$

resulting in

$$\Delta t = 2.8\ \text{ns}$$

Obtaining Differential Operation in Dynamic RAMs The sense amplifier described earlier responds to difference signals appearing between the bit lines. Thus, it is capable of rejecting interference signals that are common to both lines, such as those caused by capacitive coupling from the word lines. For this *common-mode rejection* to be effective, great care has to be taken to match both sides of the amplifier, taking into account the circuits that feed each side. This is an important consideration in any attempt to make the inherently single-ended output of the DRAM cell appear differential. We shall now discuss an ingenious scheme for accomplishing this task. Although the technique has been around for many years (see the first edition of this book, published in 1982), it is still in use today. The method is illustrated in Fig. 15.22.

Basically, each bit line is split into two identical halves. Each half-line is connected to half the cells in the column and to an additional cell, known as a *dummy cell,* having a storage capacitor $C_D = C_S$. When a word line on the left side is selected for reading, the dummy cell on the right side (controlled by $\bar{\phi}_D$) is also selected, and vice versa; that is, when a word line on

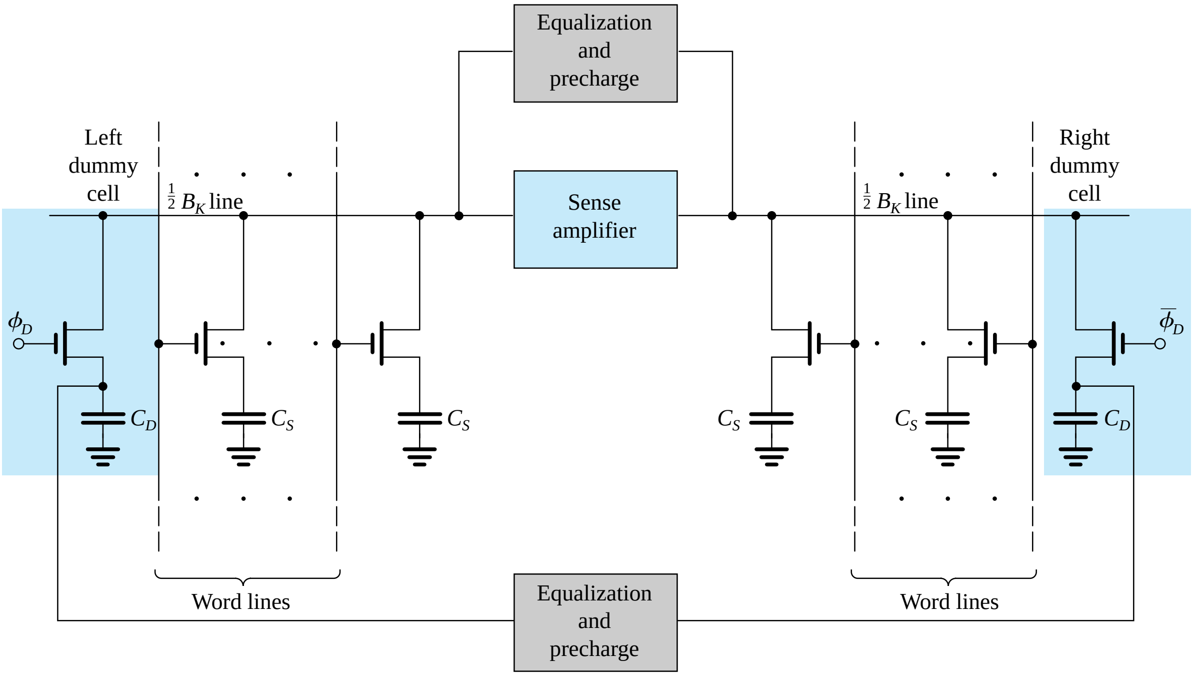


Figure 15.22 An arrangement for obtaining differential operation from the single-ended DRAM cell. Note the dummy cells at the far right and far left.

the right side is selected, the dummy cell on the left (controlled by ϕ_D) is also selected. In effect, then, the dummy cell serves as the other half of a differential DRAM cell. When the left-half bit line is in operation, the right-half bit line acts as its complement (or $\overline{B}$ line) and vice versa.

Operation of the circuit in Fig. 15.22 is as follows: The two halves of the line are precharged to $V_{DD}/2$ and their voltages are equalized. At the same time, the capacitors of the two dummy cells are precharged to $V_{DD}/2$. Then a word line is selected, and the dummy cell on the other side is enabled (with ϕ_D or $\overline{\phi}_D$ raised to V_{DD}). Thus the half-line connected to the selected cell will develop a voltage increment (around $V_{DD}/2$) of $\Delta V(1)$ or $\Delta V(0)$ depending on whether a 1 or a 0 is stored in the cell. Meanwhile, the other half of the line will have its voltage held equal to that of C_D (i.e., $V_{DD}/2$). The result is a differential signal of $\Delta V(1)$ or $\Delta V(0)$ that the sense amplifier detects and amplifies when it is enabled. As usual, by the end of the regenerative process, the amplifier will cause the voltage on one half of the line to become V_{DD} and that on the other half to become 0.

EXERCISES

15.9 It is required to reduce the time Δt of the sense-amplifier circuit in Example 15.2 by a factor of 2 by increasing g_m of the transistors (while retaining the matched design of each inverter). What must the *W/L* ratios of the *n*- and *p*-channel devices become?
Ans. $(W/L)_n = 6$; $(W/L)_p = 24$

15.10 If in the sense amplifier of Example 15.2, the signal available from the cell is only half as large (i.e., only 50 mV), what will Δt become?
Ans. 3.7 ns, an increase of 32%

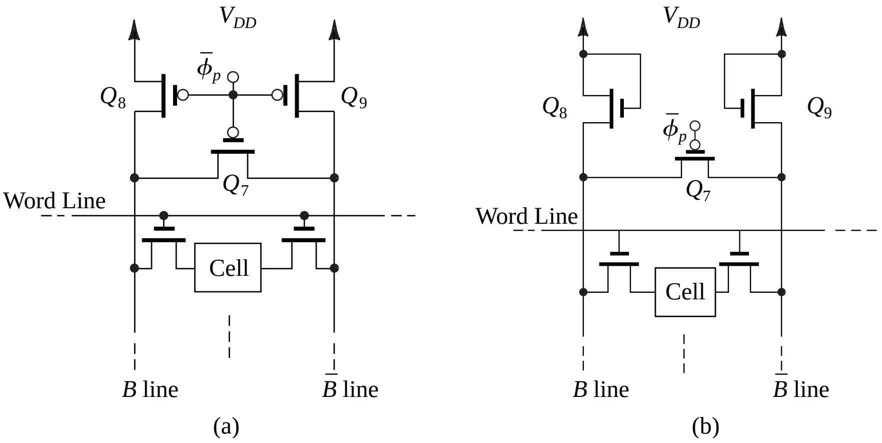


Figure 15.23 Two alternative arrangements for precharging the bit line: (**a**) The B and $\bar{B}$ lines are precharged to V_{DD}; (**b**) the B and $\bar{B}$ lines are charged to $(V_{DD} - V_t)$.

Alternative Precharging Arrangements If it is desired to precharge the B and $\bar{B}$ lines to V_{DD}, the arrangement in Fig. 15.23(a) can be utilized. Here precharging and equalization occur when $\bar{\phi}_P$ is low. Then, just prior to the activation of the word line, $\bar{\phi}_P$ goes high. Another precharging arrangement using diode-connected NMOS transistors is shown in Fig. 15.23(b). In this case, the B and $\bar{B}$ lines are charged to $(V_{DD} - V_t)$, and Q_7 equalizes their voltages.

An Alternative Sense Amplifier Another popular implementation of the sense amplifier is the differential MOS amplifier with a current-mirror load, studied in detail in Section 8.5. Here, we present a brief overview of the operation of this versatile circuit as a sense amplifier.

The amplifier circuit is shown in Fig. 15.24 fed from the bit and $\overline{\text{bit}}$ lines (voltages v_B and $v_{\bar{B}}$). Transistors Q_1 and Q_2 are connected in the differential-pair configuration and are biased by a constant current I supplied by current source Q_5. Transistors Q_3 and Q_4 form a current mirror, which acts as the load circuit for the amplifying transistors Q_1 and Q_2. The differential nature of the amplifier aids significantly in its effectiveness as a sense amplifier: It rejects noise or interference signals that are coupled equally to the B and $\bar{B}$ lines, and amplifies only the small difference signals that appear between B and $\bar{B}$ as a result of the read operation of a cell connected to the B and $\bar{B}$ lines.

The amplifier is designed so that in normal small-signal operation, all transistors operate in the saturation region. Figure 15.24(b) shows the amplifier in its equilibrium state with $v_B = v_{\bar{B}} = V_{DD} - V_t$. Note that we have assumed that the B and $\bar{B}$ lines are precharged to $(V_{DD} - V_t)$ using the circuit in Fig. 15.23(b). It turns out that this voltage is particularly convenient for the operation of this amplifier type as a sense amplifier. As indicated in Fig. 15.24(b), the bias current I divides equally between Q_1 and Q_2; thus each conducts a current $I/2$. The current of Q_1 is fed to the input side of the current mirror, transistor Q_3; thus the mirror provides an equal output current $I/2$ in the drain of Q_4. At the output node, we see that we have two equal and opposite currents, leaving a zero current to flow into the load capacitor. Thus, in an ideal situation of perfect matching, v_O will be equal to the voltage at the drain of Q_1.

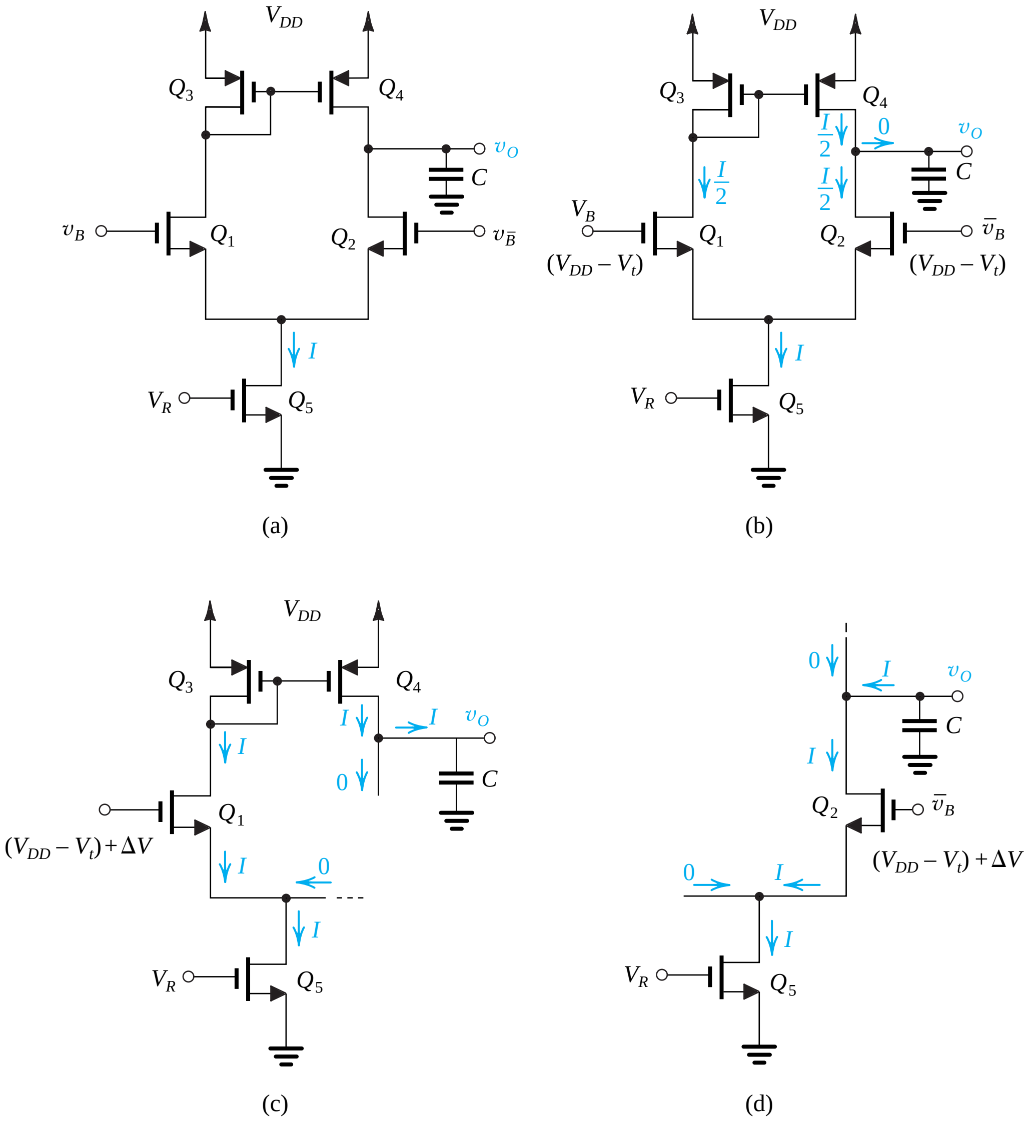


Figure 15.24 The active-loaded MOS differential amplifier as a sense amplifier.

Next consider the situation when the B line shows an incremental voltage ΔV above the voltage of the $\bar{B}$ line. As shown in Fig. 15.24(c), if ΔV is sufficiently large, Q_2 will turn off and all the bias current I will flow through Q_1 and on to Q_3. Thus the mirror output current becomes I and flows through the amplifier output terminal to the equivalent output capacitance C. Thus C will charge to V_{DD} in time Δt,

$$\Delta t = \frac{CV_{DD}}{I} \tag{15.18}$$

The complementary situation when $v_{\bar{B}}$ exceeds v_B by ΔV is illustrated in Fig. 15.24(d). Here Q_1, Q_3, and Q_4 are turned off, and Q_2 conducts all the current I. Thus capacitor C is discharged to ground by a constant current I.

An important question to answer before leaving this amplifier circuit is how large is ΔV that causes the current I to switch from one side of the differential pair to the other? The answer is given in Section 8.5 (see Fig. 8.32), namely,

$$\Delta V = \sqrt{2}V_{OV} \tag{15.19}$$

where V_{OV} is the overdrive voltage at which Q_1 and Q_2 are operating in equilibrium, that is,

$$\frac{I}{2} = \frac{1}{2}(\mu_n C_{ox})\left(\frac{W}{L}\right)_{1,2} V_{OV}^2 \tag{15.20}$$

Finally, we note that this sense amplifier dissipates static power given by

$$P = V_{DD}I$$

Observe that increasing I reduces the time Δt in Eq. (15.18) at the expense of increased power dissipation.

EXERCISE

D15.11 It is required to design the sense amplifier in Fig. 15.24 to detect an input signal $\Delta V = 100$ mV and to provide a full output in 0.5 ns. If $C = 50$ fF and $V_{DD} = 1.8$ V , find the required current I and the power dissipation.
Ans. 180 μA ; 324 μW

15.4.2 The Row-Address Decoder

As described in Section 15.2, the row-address decoder is required to select one of the 2^M word lines in response to an M-bit address input. As an example, consider the case $M = 3$ and denote the three address bits A_0, A_1, and A_2, and the eight word lines $W_0, W_1, \ldots, W_7$. Conventionally, word line W_0 will be high when $A_0 = 0$, $A_1 = 0$, and $A_2 = 0$; thus we can express W_0 as a Boolean function of A_0, A_1, and A_2,

$$W_0 = \bar{A}_0\bar{A}_1\bar{A}_2 = \overline{A_0 + A_1 + A_2}$$

Thus the selection of W_0 can be accomplished by a three-input NOR gate whose three inputs are connected to A_0, A_1, and A_2 and whose output is connected to word line 0. Word line W_3 will be high when $A_0 = 1$, $A_1 = 1$, and $A_2 = 0$; thus,

$$W_3 = A_0A_1\bar{A}_2 = \overline{\bar{A}_0 + \bar{A}_1 + A_2}$$

Thus the selection of W_3 can be realized by a three-input NOR gate whose three inputs are connected to $\bar{A}_0$, $\bar{A}_1$, and A_2, and whose output is connected to word line 3. We can thus see that this address decoder can be realized by eight three-input NOR gates. Each NOR gate is fed with the appropriate combination of address bits and their complements, corresponding to the word line to which its output is connected.

A simple approach to realizing these NOR functions is provided by the matrix structure shown in Fig. 15.25. The circuit shown is a dynamic one (Section 14.3). Attached to each

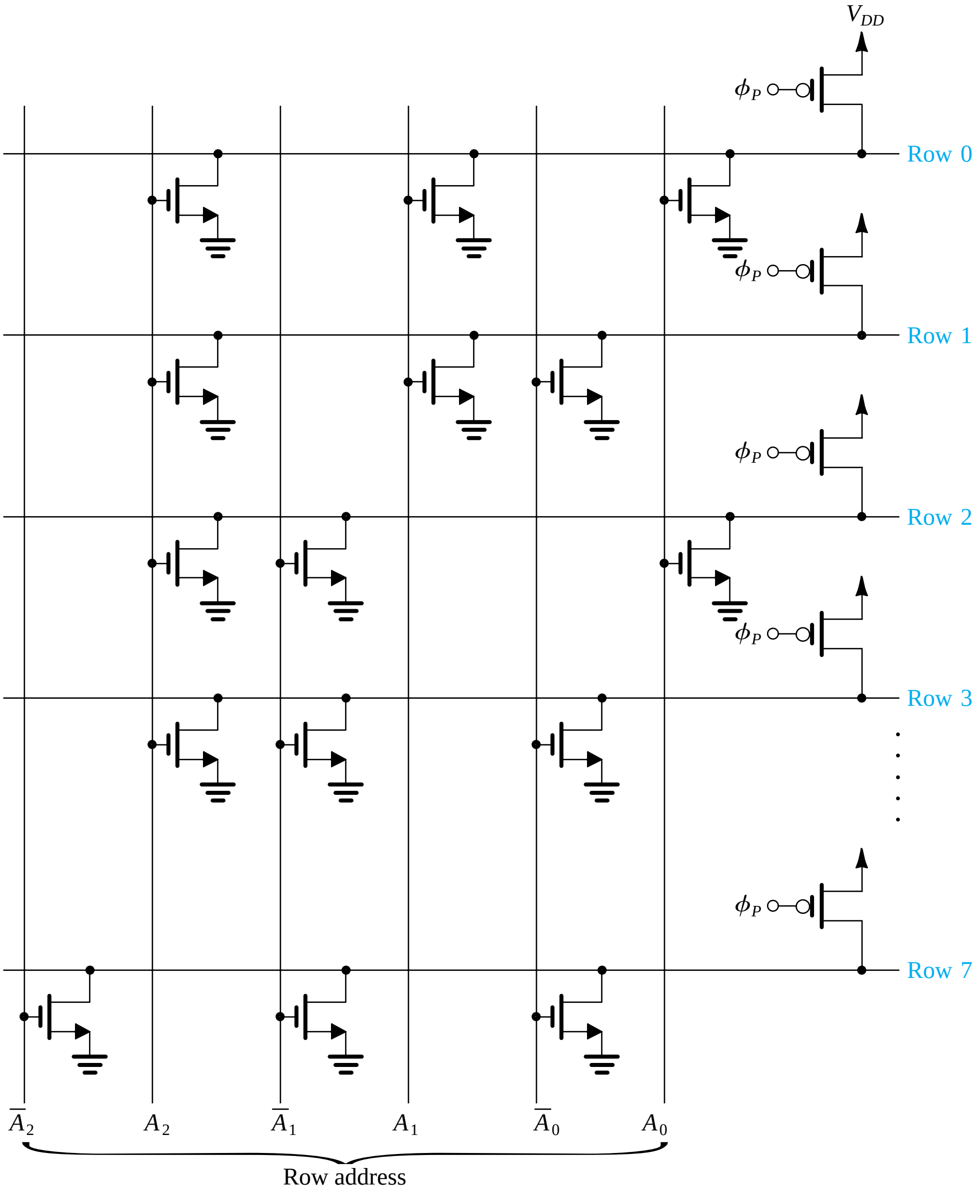


Figure 15.25 A NOR address decoder in array form. One out of eight lines (row lines) is selected using a 3-bit address.

row line is a *p*-channel device that is activated, prior to the decoding process, using the precharge control signal ϕ_P. During precharge (ϕ_P low), all the word lines are pulled high to V_{DD}. It is assumed that at this time the address input bits have not yet been applied and all the inputs are low; hence there is no need for the circuit to include the evaluation transistor utilized in dynamic logic gates. Then, the decoding operation begins when the address bits and their complements are applied. Observe that the NMOS transistors are placed so that the word lines not selected will be discharged. For any input combination, only one word line will not be discharged, and thus its voltage remains high at V_{DD}. For instance, row 0 will be high only when $A_0 = 0$, $A_1 = 0$, and $A_2 = 0$; this is the only combination that will result in all three transistors connected to row 0 being cut off. Similarly, row 3 has transistors connected to $\bar{A}_0$, $\bar{A}_1$, and A_2, and thus it will be high when $A_0 = 1$, $A_1 = 1$, $A_2 = 0$, and so on. After the decoder outputs have stabilized, the output lines are connected to the word lines of the array, usually via clock-controlled transmission gates. This decoder is known as a NOR decoder. Observe that because of the precharge operation, the decoder circuit does not dissipate static power.

EXERCISE

15.12 How many transistors are needed for a NOR row decoder with an M-bit address?
Ans. $M2^M$ NMOS + 2^M PMOS = $2^M(M + 1)$

15.4.3 The Column-Address Decoder

From the description in Section 15.2, the function of the column-address decoder is to connect one of the 2^N bit lines to the data I/O line of the chip. As such, it is a multiplexer and can be implemented using pass-transistor logic (Section 14.2) as shown in Fig. 15.26. Here, each bit line is connected to the data I/O line through an NMOS transistor. The gates of the pass transistors are controlled by 2^N lines, one of which is selected by a NOR decoder similar to that used for decoding the row address. Finally, note that better performance can be obtained by utilizing transmission gates in place of NMOS transistors (Section 14.2). In such a case, however, the decoder needs to provide complementary output signals.

An alternative implementation of the column decoder that uses a smaller number of transistors (but at the expense of slower speed of operation) is shown in Fig. 15.27. This circuit, known as a *tree decoder,* has a simple structure of pass transistors. Unfortunately, since a relatively large number of transistors can exist in the signal path, the resistance of the bit lines increases, and the speed decreases correspondingly.

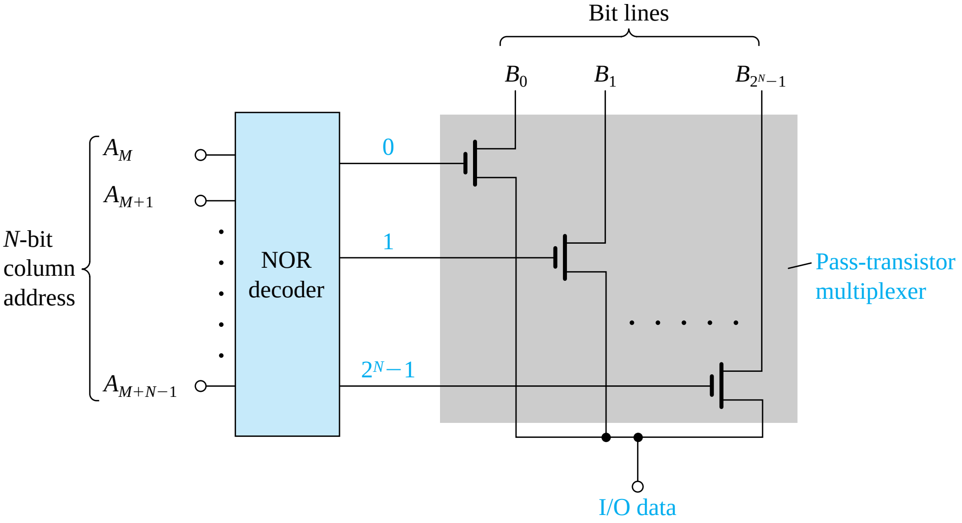


Figure 15.26 A column decoder realized by a combination of a NOR decoder and a pass-transistor multiplexer.

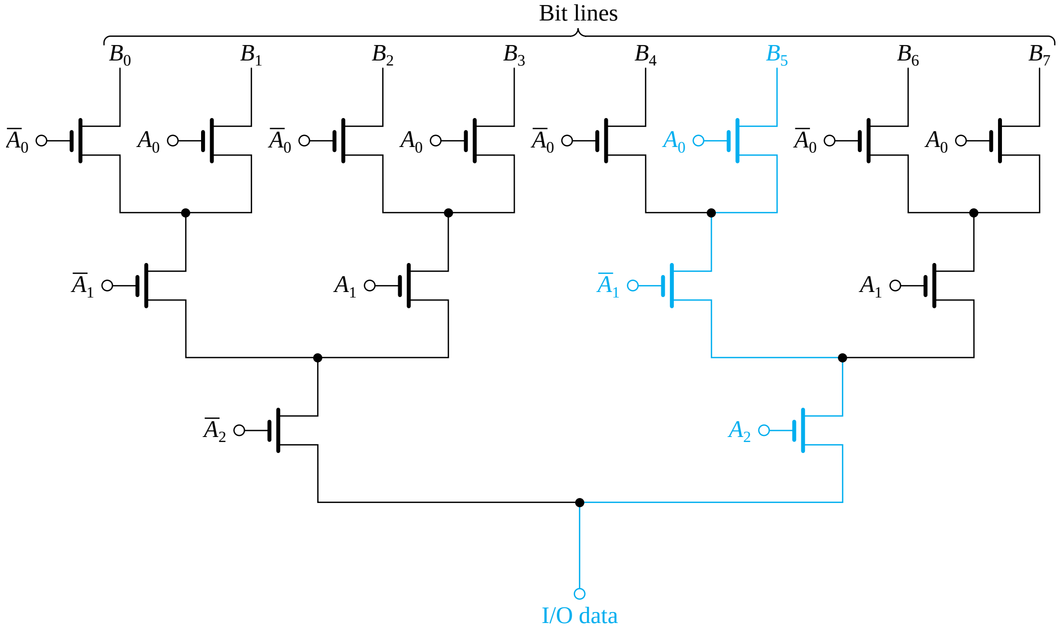


Figure 15.27 A tree column decoder. Note that the colored path shows the transistors that are conducting when $A_0 = 1$, $A_1 = 0$, and $A_2 = 1$, the address that results in connecting B_5 to the data line.

EXERCISE

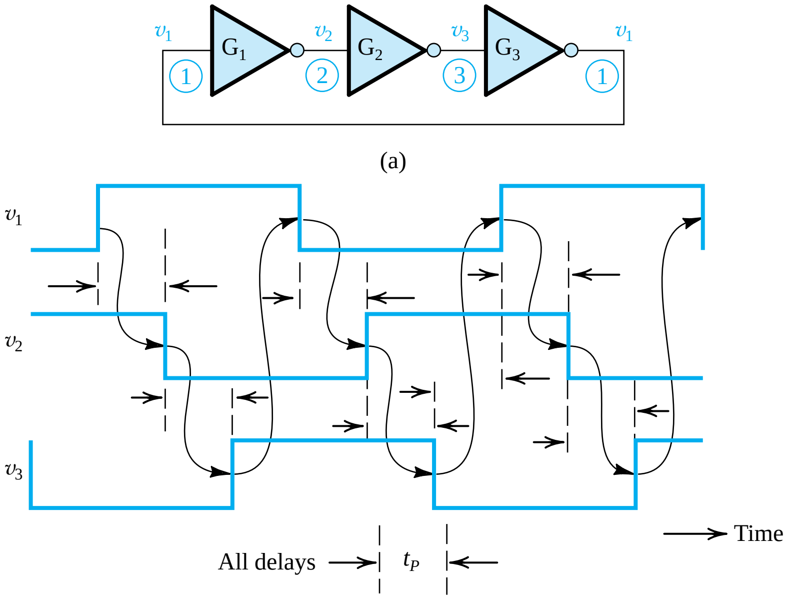
15.13 How many transistors are needed for a tree decoder when there are 2^N bit lines?
Ans. $2(2^N - 1)$

15.4.4 Pulse-Generation Circuits

Memory chips require a large number of pulse signals, sometimes with intricate timing relationships among them. It is not our purpose here to study this important subject; rather, we present two simple circuits that find widespread applicability in memory-chip timing as well as in other digital-system components, such as microprocessors.

The Ring Oscillator The ring oscillator is formed by connecting an *odd* number of inverters in a loop. Although usually at least five inverters are used, we illustrate the principle of operation using a ring of three inverters, as shown in Fig. 15.28(a). Figure 15.28(b) shows the waveforms obtained at the outputs of the three inverters. These waveforms are idealized in the sense that their edges have zero rise and fall times. Nevertheless, they will serve to explain the circuit operation.

Observe that a rising edge at node 1 propagates through gates 1, 2, and 3 to return inverted after a delay of $3t_P$. This falling edge then propagates, and returns with the original (rising) polarity after another $3t_P$ interval. It follows that the circuit oscillates with a period



(a)

(b)

Figure 15.28 **(a)** A ring oscillator formed by connecting three inverters in cascade. (Normally at least five inverters are used.) **(b)** The resulting waveform. Observe that the circuit oscillates with frequency $1/6t_P$.

of $6t_P$ or correspondingly with frequency $1/6t_P$. In general, a ring with N inverters (where N must be odd) will oscillate with a period of $2Nt_P$ and frequency $1/2Nt_P$.

As a final remark, we note that the ring oscillator provides a relatively simple means for measuring the inverter propagation delay.

EXERCISE

15.14 Find the frequency of oscillation of a ring of five inverters if the inverter propagation delay is specified to be 1 ns.

Ans. 100 MHz

A One-Shot or Monostable Multivibrator Circuit The one-shot or monostable multivibrator circuit provides, when triggered, a single output pulse with a predetermined width.[3] A variety of circuits exist for implementing the one-shot function, and some using op amps will be studied in Section 17.6. Here, in Fig. 15.29(a), we show a circuit commonly used in digital IC design. The circuit utilizes an exclusive-OR (XOR) gate together with a delay circuit. Recalling that the XOR gate provides a high output only when its two inputs are dissimilar, we see that prior to the arrival of the input positive step, the output will be

[3] The name "monostable" arises because this class of circuits has one stable state, which is the quiescent state. When a trigger is applied, the circuit moves to its quasi-stable state and stays in it for a predetermined length of time (the width of the output pulse). It then switches back automatically to the stable state.

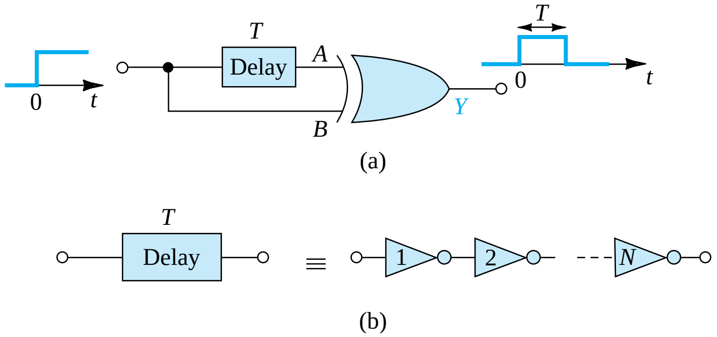


Figure 15.29 (**a**) A one-shot or monostable circuit. Utilizing a delay circuit with a delay T and an XOR gate, this circuit provides an output pulse of width T. (**b**) The delay circuit can be implemented as the cascade of N inverters where N is even, in which case $T = Nt_P$.

low. When the input goes high, only the B input of the XOR will be high and thus its output will go high. The high input will reach input A of the XOR T seconds later, at which time both inputs of the XOR will be high and thus its output will go low. We thus see that the circuit produces an output pulse with a duration T equal to the delay of the delay block for each transition of the input signal. The latter can be implemented by connecting an even number of inverters in cascade as shown in Fig. 15.29(b).

15.5 Read-only Memory (ROM)

As mentioned in Section 15.2, read-only memory (ROM) is memory that contains fixed data patterns. It is used in a variety of digital-system applications. Currently, a very popular application is the use of ROM in microprocessor systems to store the instructions of the system's basic operating program. ROM is particularly suited for such an application because it is nonvolatile; that is, it retains its contents when the power supply is switched off.

A ROM can be viewed as a combinational logic circuit for which the input is the collection of address bits of the ROM and the output is the set of data bits retrieved from the addressed location. This viewpoint leads to the application of ROMs in code conversion—that is, in changing the code of the signal from one system (say, binary) to another. Code conversion is employed, for instance, in secure communication systems, where the process is known as *scrambling*. It consists of feeding the code of the data to be transmitted to a ROM that provides corresponding bits in a (supposedly) secret code. The reverse process, which also uses a ROM, is applied at the receiving end.

In this section we will study various types of read-only memory. These include fixed ROM, which we refer to simply as ROM, programmable ROM (PROM), and erasable programmable ROM (EPROM).

15.5.1 A MOS ROM

Figure 15.30 shows a simplified 32-bit (or 8-word × 4-bit) MOS ROM. As indicated, the memory consists of an array of n-channel MOSFETs whose gates are connected to the word lines, whose sources are grounded, and whose drains are connected to the bit lines. Each bit line is connected to the power supply via a PMOS load transistor, in the manner of pseudo-NMOS logic

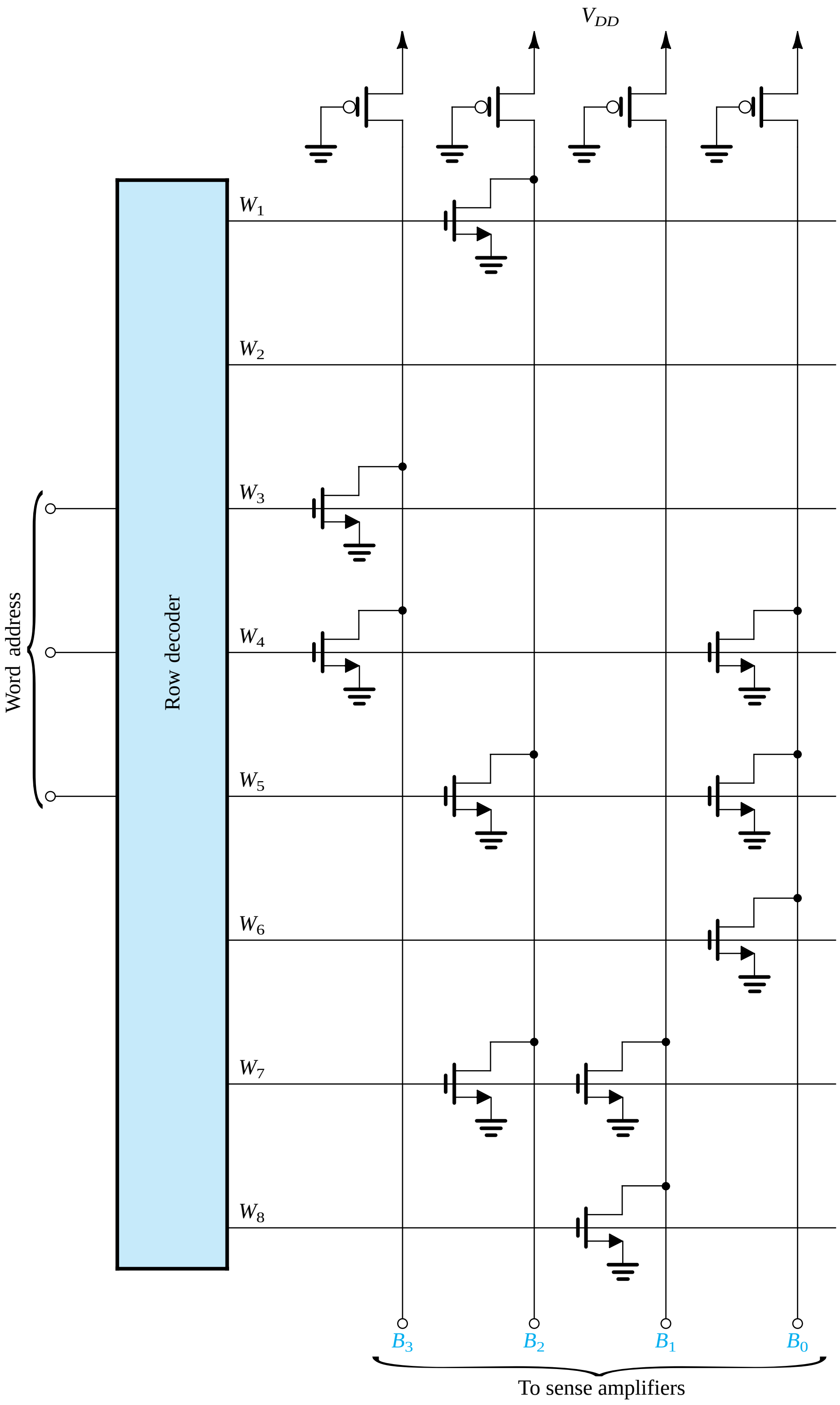


Figure 15.30 A simple MOS ROM organized as 8 words × 4 bits.

(Section 14.1). An NMOS transistor exists in a particular cell if the cell is storing a 0; a cell storing a 1 has no MOSFET. This ROM can be thought of as 8 words of 4 bits each. The row decoder selects one of the 8 words by raising the voltage of the corresponding word line. The cell transistors connected to this word line will then conduct, thus pulling the voltage of the bit lines (to which transistors in the selected row are connected) down from V_{DD} to a voltage close to ground voltage (the logic-0 level). The bit lines that are connected to cells (of the selected word) without transistors (i.e., the cells that are storing a logic 1) will remain at the power-supply voltage (logic 1) because of the action of the pull-up PMOS load devices. In this way, the bits of the addressed word can be read.

A disadvantage of the ROM circuit in Fig. 15.30 is that it dissipates static power. Specifically, when a word is selected, the transistors in this particular row will conduct static current that is supplied by the PMOS load transistors. Static power dissipation can be eliminated by a simple change. Rather than grounding the gate terminals of the PMOS transistors, we can connect these transistors to a precharge line ϕ that is normally high. Just before a read operation, ϕ is lowered and the bit lines are precharged to V_{DD} through the PMOS transistors. The precharge signal ϕ then goes high, and the word line is selected. The bit lines that have transistors in the selected word are then discharged, thus indicating stored zeros, whereas those lines for which no transistor is present remain at V_{DD}, indicating stored ones.

EXERCISE

15.15 The purpose of this exercise is to estimate the various delay times involved in the operation of a ROM. Consider the ROM in Fig. 15.30 with the gates of the PMOS devices disconnected from ground and connected to a precharge control signal ϕ. Let all the NMOS devices have W/L = 6 μm/2 μm and all the PMOS devices have W/L = 24 μm/2 μm. Assume that $\mu_n C_{ox} = 50\ \mu\text{A/V}^2$, $\mu_p C_{ox} = 20\ \mu\text{A/V}^2$, $V_{tn} = -V_{tp} = 1$ V, and $V_{DD} = 5$ V.

(a) During the precharge interval, ϕ is lowered to 0 V. Estimate the time required to charge a bit line from 0 V to 5 V. Use, as an average charging current, the current supplied by a PMOS transistor at a bit-line voltage halfway through the 0-V to 5-V excursion (i.e., 2.5 V). The bit-line capacitance is 2 pF. Note that all NMOS transistors are cut off at this time.

(b) After completion of the precharge interval and the return of ϕ to V_{DD}, the row decoder raises the voltage of the selected word line. Because of the finite resistance and capacitance of the word line, the voltage rises exponentially toward V_{DD}. If the resistance of each of the polysilicon word lines is 3 kΩ and the capacitance between the word line and ground is 3 pF, what is the (10% to 90%) rise time of the word-line voltage? What is the voltage reached at the end of one time constant?

(c) We account for the exponential rise of the word-line voltage by approximating the word-line voltage by a step equal to the voltage reached in one time constant. Find the interval Δt required for an NMOS transistor to discharge the bit line and lower its voltage by 0.5 V. (It is assumed that the sense amplifier needs a 0.5-V change at its input to detect a low bit value.)

Ans. (a) 6.1 ns; (b) 19.8 ns, 3.16 V; (c) 2.9 ns

15.5.2 Mask-Programmable ROMs

The data stored in the ROMs discussed thus far is determined at the time of fabrication, according to the user's specifications. However, to avoid having to custom-design each ROM from scratch (which would be extremely costly), ROMs are manufactured using a process known as **mask programming**. As explained in Appendix A, integrated circuits are fabricated on a wafer of silicon using a sequence of processing steps that include photomasking, etching, and diffusion. In this way, a pattern of junctions and interconnections is created on the surface of the wafer. One of the final steps in the fabrication process consists of coating the surface of the wafer with a layer of aluminum and then selectively (using a mask) etching away portions of the aluminum, leaving aluminum only where interconnections are desired. This last step can be used to program (i.e., to store a desired pattern in) a ROM. For instance, if the ROM is made of MOS transistors as in Fig. 15.30, MOSFETs can be included at all bit locations, but only the gates of those transistors where 0s are to be stored are connected to the word lines; the gates of transistors where 1s are to be stored are not connected. This pattern is determined by the mask, which is produced according to the user's specifications.

The economic advantages of the mask programming process should be obvious: All ROMs are fabricated similarly; customization occurs only during one of the final steps in fabrication.

15.5.3 Programmable ROMs (PROMs and EPROMs)

PROMs are ROMs that can be programmed by the user, but only once. A typical arrangement employed in BJT PROMs involves using polysilicon fuses to connect the emitter of each BJT to the corresponding digit line. Depending on the desired content of a ROM cell, the fuse can be either left intact or blown out using a large current. The programming process is obviously irreversible.

An erasable programmable ROM, or EPROM, is a ROM that can be erased and reprogrammed as many times as the user wishes. It is therefore the most versatile type of read-only memory. It should be noted, however, that the process of erasure and reprogramming is time-consuming and is intended to be performed only infrequently.

State-of-the-art EPROMs use variants of the memory cell whose cross section is shown in Fig. 15.31(a). The cell is basically an enhancement-type *n*-channel MOSFET with two gates made of polysilicon material.[4] One of the gates is not electrically connected to any

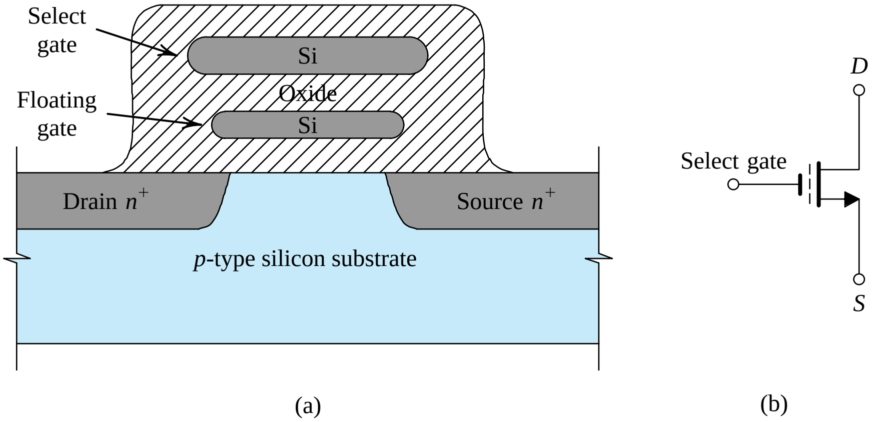


Figure 15.31 **(a)** Cross section and **(b)** circuit symbol of the floating-gate transistor used as an EPROM cell.

[4] See Appendix A for a description of silicon-gate technology.

other part of the circuit; rather, it is left floating and is appropriately called a **floating gate**. The other gate, called a **select gate**, functions in the same manner as the gate of a regular enhancement MOSFET.

The MOS transistor of Fig. 15.31(a) is known as a **floating-gate transistor** and is given the circuit symbol shown in Fig. 15.31(b). In this symbol the broken line denotes the floating gate. The memory cell is known as the **stacked-gate cell**.

Let us now examine the operation of the floating-gate transistor. Before the cell is programmed (we will shortly explain what this means), no charge exists on the floating gate and the device operates as a regular *n*-channel enhancement MOSFET. It thus exhibits the i_D–v_{GS} characteristic shown as curve (*a*) in Fig. 15.32. Note that in this case the threshold voltage (V_t) is rather low. This state of the transistor is known as the **not-programmed state**. It is one of two states in which the floating-gate transistor can exist. Let us arbitrarily take the not-programmed state to represent a stored 1. That is, a floating-gate transistor whose i_D–v_{GS} characteristic is that shown as curve (*a*) in Fig. 15.32 will be said to be storing a 1.

To program the floating-gate transistor, a large voltage (16–20 V) is applied between its drain and source. Simultaneously, a large voltage (about 25 V) is applied to its select gate. Figure 15.33 shows the floating-gate MOSFET during programming. In the absence of any charge on the floating gate, the device behaves as a regular *n*-channel enhancement MOSFET: An *n*-type inversion layer (channel) is created at the wafer surface as a result of the large positive voltage applied to the select gate. Because of the large positive voltage at the drain, the channel has a tapered shape.

The drain-to-source voltage accelerates electrons through the channel. As these electrons reach the drain end of the channel, they acquire large kinetic energy and are referred to as *hot electrons*. The large positive voltage on the select gate (greater than the drain voltage) establishes an electric field in the insulating oxide. This electric field attracts the hot electrons and accelerates them (through the oxide) toward the floating gate. In this way the floating gate is charged, and the charge that accumulates on it becomes trapped.

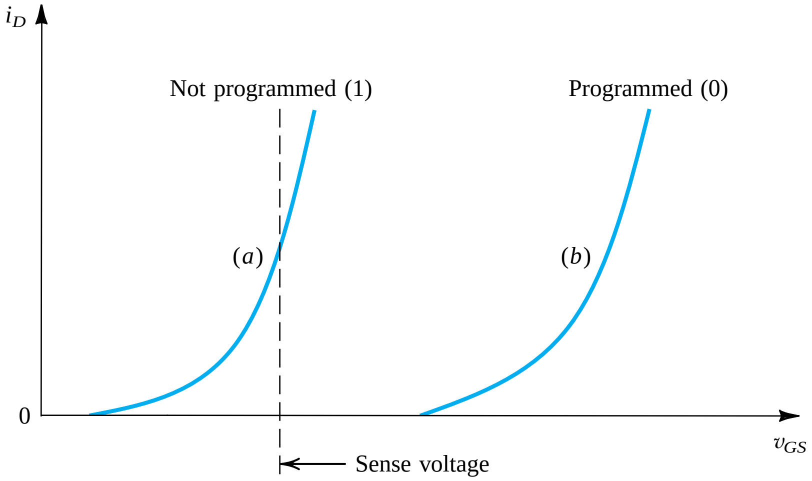


Figure 15.32 Illustrating the shift in the i_D–v_{GS} characteristic of a floating-gate transistor as a result of programming.

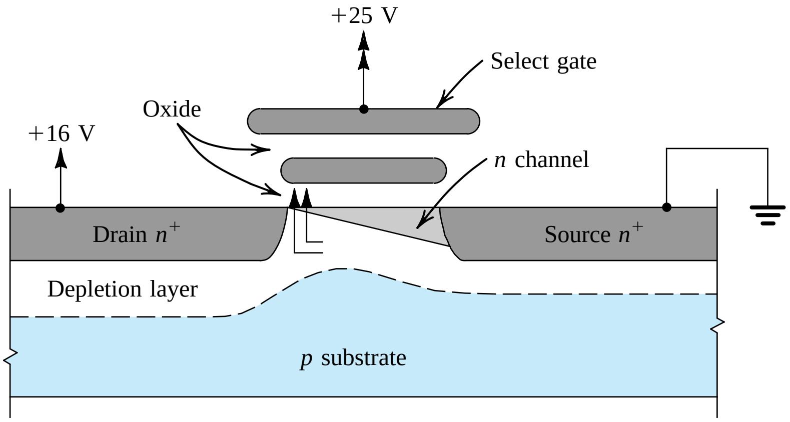


Figure 15.33 The floating-gate transistor during programming.

Fortunately, the process of charging the floating gate is self-limiting. The negative charge that accumulates on the floating gate reduces the strength of the electric field in the oxide to the point that it eventually becomes incapable of accelerating any more of the hot electrons.

Let us now inquire about the effect of the floating gate's negative charge on the operation of the transistor. The negative charge trapped on the floating gate will cause electrons to be repelled from the surface of the substrate. This implies that to form a channel, the positive voltage that has to be applied to the select gate will have to be greater than that required when the floating gate is not charged. In other words, the threshold voltage V_t of the programmed transistor will be higher than that of the not-programmed device. In fact, programming causes the i_D–v_{GS} characteristic to shift to the curve labeled (b) in Fig. 15.32. In this state, known as the *programmed state,* the cell is said to be storing a 0.

Once programmed, the floating-gate device retains its shifted i–v characteristic (curve b) even when the power supply is turned off. In fact, extrapolated experimental results indicate that the device can remain in the programmed state for as long as 100 years!

Reading the content of the stacked-gate cell is easy: A voltage V_{GS} somewhere between the low and high threshold values (see Fig. 15.32) is applied to the selected gate. While a programmed device (one that is storing a 0) will not conduct, a not-programmed device (one that is storing a 1) will conduct heavily.

To return the floating-gate MOSFET to its not-programmed state, the charge stored on the floating gate has to be returned to the substrate. This *erasure* process can be accomplished by illuminating the cell with ultraviolet light of the correct wavelength (2537 Å) for a specified duration. The ultraviolet light imparts sufficient photon energy to the trapped electrons to allow them to overcome the inherent energy barrier, and thus be transported through the oxide, back to the substrate. To allow this erasure process, the EPROM package contains a quartz window. Finally, it should be noted that the device is extremely durable, and can be erased and programmed many times.

A more versatile programmable ROM is the electrically erasable PROM (or EEPROM). As the name implies, an EEPROM can be erased and reprogrammed electrically without the need for ultraviolet illumination. EEPROMs utilize a variant of the floating-gate MOSFET. An important class of EEPROMs using a floating gate variant and implementing block erasure are referred to as **flash memories**.

Summary

- Flip-flops employ one or more latches. The basic static latch is a bistable circuit implemented using two inverters connected in a positive-feedback loop. The latch can remain in either stable state indefinitely.

- As an alternative to the positive-feedback approach, memory can be provided through the use of charge storage. A number of CMOS flip-flops are realized this way, including some master–slave D flip-flops.

- A random-access memory (RAM) is one in which the time required for storing (writing) information and for retrieving (reading) information is independent of the physical location (within the memory) in which the information is stored.

- The major part of a memory chip consists of the cells in which the bits are stored and that are typically organized in a square matrix. A cell is selected for reading or writing by activating its row, via the row-address decoder, and its column, via the column-address decoder. The sense amplifier detects the content of the selected cell and provides a full-swing version of it to the data-output terminal of the chip.

- There are two kinds of MOS RAM: static and dynamic. Static RAMs (SRAMs) employ flip-flops as the storage cells. In a dynamic RAM (DRAM), data is stored on a capacitor and thus must be periodically refreshed. DRAM chips provide the highest possible storage capacity for a given chip area.

- Two circuits have emerged as the near-universal choice in implementing the storage cell: the six-transistor SRAM cell and the one-transistor DRAM cell.

- Although sense amplifiers are utilized in SRAMs to speed up operation, they are essential in DRAMs. A particular type of sense amplifier is a differential circuit that employs positive feedback to obtain an output signal that grows exponentially toward either V_{DD} or 0.

- Read-only memory (ROM) contains fixed data patterns that are stored at the time of fabrication and cannot be changed by the user. On the other hand, the contents of an erasable programmable ROM (EPROM) can be changed by the user. The erasure and reprogramming is a time-consuming process and is performed only infrequently.

- Some EPROMS utilize floating-gate MOSFETs as the storage cells. The cell is programmed by applying (to the selected gate) a high voltage, which in effect changes the threshold voltage of the MOSFET. Erasure is achieved by illuminating the chip by ultraviolet light. Even more versatile, EEPROMs can be erased and reprogrammed electrically.

Problems involving design are marked with D throughout the text. As well, problems are marked with asterisks to describe their degree of difficulty. Difficult problems are marked with an asterisk (*); more difficult problems with two asterisks (**); and very challenging and/or time-consuming problems with three asterisks (***).

Section 15.1: Latches and Flip-Flops

D 15.1 Sketch the standard CMOS circuit implementation of the SR flip-flop shown in Fig. 15.3.

D 15.2 Sketch the logic-gate implementation of an SR flip-flop utilizing two cross-coupled NAND gates. Clearly label the output terminals and the input trigger terminals. Provide the truth table and describe the operation.

D 15.3 For the SR flip-flop of Fig. 15.4, show that if each of the two inverters utilizes matched transistors, that is, $(W/L)_p = (\mu_n/\mu_p)(W/L)_n$, then the minimum *W*/*L* that each of Q_5–Q_8 must have so that switching occurs is $2(W/L)_n$. Give the sizes of all eight transistors if the flip-flop is fabricated in a 0.13-µm process for which $\mu_n = 4\mu_p$. Use the minimum channel length for all transistors and the minimum size ($W/L = 1$) for Q_1 and Q_3.

D 15.4 In this problem we investigate the effect of velocity saturation (Section 13.5) on the design of the SR flip-flop in Example 15.1. Specifically, answer part (a) of the question in Example 15.1, taking into account the fact that for this technology, $V_{DS\text{sat}}$ for *n*-channel devices is 0.6 V and $|V_{DS\text{sat}}|$ for *p*-channel devices is 1 V. Assume $\lambda_n = |\lambda_p| = 0.1\text{ V}^{-1}$. What is the minimum required value for $(W/L)_5$ and for $(W/L)_6$? Comment on this value relative to that found in Example 15.1. (*Hint*: Refer to Eq. 13.100.)

D 15.5 Repeat part (a) of the problem in Example 15.1 for the case of inverters that do not use matched Q_N and Q_P. Rather, assume that each of the inverters uses $(W/L)_p = (W/L)_n = 0.27\ \mu\text{m}/0.18\ \mu\text{m}$. Find the threshold voltage of each inverter. Then determine the value required for the *W*/*L* of each of Q_5 to Q_8 so that the flip-flop switches.

D 15.6 The CMOS SR flip-flop in Fig. 15.4 is fabricated in a 0.13-µm process for which $\mu_n C_{ox} = 4\mu_p C_{ox} = 430\ \mu\text{A/V}^2$, $V_{tn} = |V_{tp}| = 0.4$ V, and $V_{DD} = 1.2$ V. The inverters have $(W/L)_n = 0.2\ \mu\text{m}/0.13\ \mu\text{m}$ and $(W/L)_p = 0.8\ \mu\text{m}/0.13\ \mu\text{m}$. The four NMOS transistors in the set–reset circuit have equal *W*/*L* ratios.

(a) Determine the minimum value required for this ratio to ensure that the flip-flop will switch.
(b) If a ratio twice the minimum is selected, determine the minimum required width of the set and reset pulses to ensure switching. Assume that the total capacitance between each of the *Q* and $\overline{Q}$ nodes and ground is 20 fF.

D 15.7 Consider another possibility for the circuit in Fig. 15.7: Relabel the *R* input as $\overline{S}$ and the *S* input as $\overline{R}$. Let $\overline{S}$ and $\overline{R}$ normally rest at V_{DD}. Let the flip-flop be storing a 0; thus $V_Q = 0$ V and $V_{\overline{Q}} = V_{DD}$. To set the flip-flop, the $\overline{S}$ terminal is lowered to 0 V and the clock ϕ is raised to V_{DD}. The relevant part of the circuit is then transistors Q_5 and Q_2. For the flip-flop to switch, the voltage at $\overline{Q}$ must be lowered to $V_{DD}/2$. What is the minimum required *W*/*L* for Q_5 in terms of $(W/L)_2$ and (μ_n/μ_p)?

D 15.8 The clocked SR flip-flop in Fig. 15.4 is not a fully complementary CMOS circuit. Sketch the fully complementary version by augmenting the circuit with the PUN corresponding to the PDN comprising Q_5, Q_6, Q_7, and Q_8. Note that the fully complementary circuit utilizes 12 transistors. Although the circuit is more complex, it switches faster.

****15.9** Consider the latch of Fig. 15.1 as implemented in CMOS technology. Let $\mu_n C_{ox} = 2\mu_p C_{ox} = 20\ \mu\text{A/V}^2$, $W_p = 2W_n = 24\ \mu\text{m}$, $L_p = L_n = 6\ \mu\text{m}$, $|V_t| = 1$ V, and $V_{DD} = 5$ V.

(a) Plot the transfer characteristic of each inverter—that is, v_X versus v_W, and v_Z versus v_Y. Determine the output of each inverter at input voltages of 1, 1.5, 2, 2.25, 2.5, 2.75, 3, 3.5, 4, and 5 volts.
(b) Use the characteristics in (a) to determine the loop voltage-transfer curve of the latch—that is, v_Z versus v_W. Find the coordinates of points A, B, and C as defined in Fig. 15.1(c).
(c) If the finite output resistance of the saturated MOSFET is taken into account, with $|V_A| = 100$ V, find the slope of the loop transfer characteristic at point B. What is the approximate width of the transition region?

15.10 Two CMOS inverters operating from a 5-V supply have V_{IH} and V_{IL} of 2.42 and 2.00 V and corresponding outputs of 0.4 V and 4.6 V, respectively, and are connected as a latch. The MOSFETs have $|V_t| = 1$ V. Approximating the corresponding transfer characteristic of each gate by straight lines between the break points, sketch the latch open-loop transfer characteristic. What are the coordinates of point B? What is the loop gain at B?

***15.11** Figure P15.11 shows a commonly used circuit of a D flip-flop that is triggered by the negative-going edge of the clock ϕ.

(a) For ϕ high, what are the values of $\overline{Q}$ and *Q* in terms of *D*? Which transistors are conducting?
(b) If *D* is high and ϕ goes low, which transistors conduct and what signals appear at $\overline{Q}$ and at *Q*? Describe the circuit operation.

(c) Repeat (b) for D low with the clock ϕ going low.
(d) Does the operation of this circuit rely on charge storage?

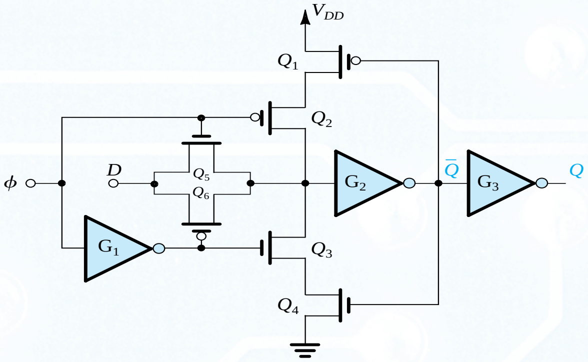


Figure P15.11

Section 15.2: Semiconductor Memories: Types and Architectures

15.12 A particular 1 M-bit square memory array has its peripheral circuits reorganized to allow for the readout of a 16-bit word. How many address bits will the new design need?

15.13 For the memory chip described in Problem 15.12, how many word lines must be supplied by the row decoder? How many sense amplifiers/drivers would a straightforward implementation require? If the chip power dissipation is 500 mW with a 5-V supply for continuous operation with a 200-ns cycle time, and all the power loss is dynamic, estimate the total capacitance of all logic activated in any one cycle. If we assume that 90% of this power loss occurs in array access, and that the major capacitance contributor will be the bit line itself, calculate the capacitance per bit line and per bit for this design. (Recall from problem 15.12 that 16 bit lines are selected simultaneously.) If closer manufacturing control allows the memory array to operate at 3 V, how much larger a memory array can be designed in the same technology at about the same power level?

15.14 An experimental 1.5-V, 1-Gbit dynamic RAM (called DRAM) by Hitachi uses a 0.16-μm process with a cell size of $0.38 \times 0.76\ \mu\text{m}^2$ in a $19 \times 38\ \text{mm}^2$ chip. What fraction of the chip is occupied by the I/O connections, peripheral circuits, and interconnect?

Section 15.3: Random-Access Memory (RAM) Cells

15.15 Consider the read operation of the 6T SRAM cell of Fig. 15.12 when it is storing a 0, that is, $V_Q = 0$ V, and $V_{\overline{Q}} = V_{DD}$. Assume that the bit lines are precharged to V_{DD} before the word-line voltage is raised to V_{DD}. Sketch the relevant part of the circuit and describe the operation. Show that the analysis parallels that presented in the text for the read-1 operation.

D 15.16 Consider a 6T SRAM cell fabricated in a 0.18-μm CMOS process for which $V_{tn} = |V_{tp}| = 0.5$ V and $V_{DD} = 1.8$ V. If during a read-1 operation it is required that $V_{\overline{Q}}$ does not exceed 0.2 V, use the graph in Fig. 15.14 to determine the maximum allowable value of the ratio $(W/L)_5/(W/L)_1$. For $L_1 = L_5 = 0.18\ \mu\text{m}$, select values for W_1 and W_5 that minimize the combined areas of Q_1 and Q_5. Assume that the minimum width allowed is 0.18 μm.

15.17 Repeat Exercise 15.4 for an SRAM fabricated in a 0.25-μm CMOS process for which $V_{DD} = 2.5$ V and $V_t = 0.5$ V.

15.18 Repeat Exercise 15.4 for an SRAM fabricated in a 0.13-μm CMOS process for which $V_{DD} = 1.2$ V and $V_t = 0.4$ V.

15.19 Locate on the graph of Fig. 15.14 the points A, B, and C that correspond to the following three process technologies:

(a) 0.25-μm: $V_{DD} = 2.5$ V and $V_t = 0.5$ V
(b) 0.18-μm: $V_{DD} = 1.8$ V and $V_t = 0.5$ V
(c) 0.13-μm: $V_{DD} = 1.2$ V and $V_t = 0.4$ V

In each case, impose the condition that in a read-1 operation $V_{\overline{Q}} = V_t$.

***15.20** Refer to the circuit in Fig. 15.13 and find the maximum ratio $(W/L)_5/(W/L)_1$ for $V_{\overline{Q}} \le V_t$, this time taking into account the velocity-saturation effect (Section 13.5, Eq. 13.100). The SRAM is fabricated in a 0.18-μm CMOS process for which $V_{DD} = 1.8$ V, $V_t = 0.5$ V, and for the n-channel devices $V_{DS\text{sat}} = 0.6$ V. Compare to the value obtained without accounting for velocity saturation. (*Hint*: Convince yourself that for this situation only Q_5 will be operating in velocity saturation.)

D *15.21 For the 6T SRAM of Fig. 15.12, fabricated in a 0.18-μm CMOS process for which $V_{DD} = 1.8$ V, $V_{t0} = 0.5$ V, $2\phi_f = 0.8$ V, and $\gamma = 0.3\ \text{V}^{1/2}$, find the maximum ratio $(W/L)_5/(W/L)_1$ for which $V_{\overline{Q}} \le V_{t0}$ during a read-1 operation (Fig. 15.13). Take into account the body effect in Q_5. Compare to the value obtained without accounting for the body effect.

D 15.22 A 6T SRAM cell is fabricated in a 0.13-μm CMOS process for which $V_{DD} = 1.2$ V, $V_t = 0.4$ V, and $\mu_n C_{ox} = 430\ \mu\text{A/V}^2$. The inverters utilize $(W/L)_n = 1$. Each of the bit lines has a 2-pF capacitance to ground. The sense amplifier requires a minimum of 0.2-V input for reliable and fast operation.

(a) Find the upper bound on W/L for each of the access transistors so that V_Q and $V_{\bar{Q}}$ do not change by more than V_t volts during the read operation.
(b) Find the delay time Δt encountered in the read operation if the cell design utilizes minimum-size access transistors.
(c) Find the delay time Δt if the design utilizes the maximum allowable size for the access transistors.

15.23 Consider the operation of writing a 1 into a 6T SRAM cell that is originally storing a 0. Sketch the relevant part of the circuit and explain the operation. Without doing detailed analysis, show that the analysis would lead to results identical to those obtained in the text for the write-0 operation.

D 15.24 For a 6T SRAM cell fabricated in a 0.13-μm CMOS process, find the maximum permitted value of $(W/L)_p$ in terms of $(W/L)_a$ of the access transistors. Assume $V_{DD} = 1.2$ V, $V_{tn} = |V_{tp}| = 0.4$ V, and $\mu_n = 4\mu_p$.

D 15.25 For a 6T SRAM cell fabricated in a 0.25-μm CMOS process, find the maximum permitted value of $(W/L)_p$ in terms of $(W/L)_a$ of the access transistors. Assume $V_{DD} = 2.5$ V, $V_{tn} = |V_{tp}| = 0.5$ V, and $\mu_n \simeq 4\mu_p$.

15.26 Locate on the graph in Fig. 15.17 the points A, B, and C corresponding to the following three CMOS fabrication processes:

(a) 0.25-μm: $V_{DD} = 2.5$ V, $V_{tn} = |V_{tp}| = 0.5$ V

(b) 0.18-μm: $V_{DD} = 1.8$ V, $V_{tn} = |V_{tp}| = 0.5$ V

(c) 0.13-μm: $V_{DD} = 1.2$ V, $V_{tn} = |V_{tp}| = 0.4$ V

For all three, $\mu_n \simeq 4\mu_p$. In each case, V_Q is to be limited to a maximum value of V_{tn}.

D 15.27 Design a minimum-size 6T SRAM cell in a 0.13-μm process for which $V_{DD} = 1.2$ V and $V_{tn} = |V_{tp}| = 0.4$ V. All transistors are to have equal $L = 0.13$ μm. Assume that the minimum width allowed is 0.13 μm. Verify that your minimum-size cell meets the constraints in Eqs. (15.5) and (15.11).

15.28 For a particular DRAM design, the cell capacitance $C_S = 30$ fF and $V_{DD} = 1.8$ V. Each cell represents a capacitive load on the bit line of 1 fF. Assume a 28-fF capacitance for the sense amplifier and other circuitry attached to the bit line. What is the maximum number of cells that can be attached to a bit line while ensuring a minimum bit-line signal of 0.05 V? How many bits of row addressing can be used? If the sense-amplifier gain is increased by a factor of 5, how many word-line address bits can be accommodated?

15.29 For a DRAM available for regular use 98% of the time, having a row-to-column ratio of 2 to 1, a cycle time of 20 ns, and a refresh cycle of 8 ms, estimate the total memory capacity.

15.30 In a particular dynamic memory chip, $C_S = 25$ fF, the bit-line capacitance per cell is 0.5 fF, and bit-line control circuitry involves 12 fF. For a 1-Mbit-square array, what bit-line signals result when a stored 1 is read? When a stored 0 is read? Assume that $V_{DD} = 1.8$ V.

15.31 For a DRAM cell utilizing a capacitance of 20 fF, refresh is required within 10 ms. If a signal loss on the capacitor of 0.2 V can be tolerated, what is the largest acceptable leakage current present at the cell?

Section 15.4: Sense Amplifiers and Address Decoders

D 15.32 Consider the operation of the differential sense amplifier of Fig. 15.20 following the rise of the sense control signal ϕ_s. Assume that a balanced differential signal of 0.1 V is established between the bit lines, each of which has a 1 pF capacitance. For $V_{DD} = 3$ V, what is the value of G_m of each of the inverters in the amplifier required to cause the outputs to reach $0.1V_{DD}$ and $0.9V_{DD}$ [from initial values of $0.5V_{DD} + (0.1/2)$ and $0.5V_{DD} - (0.1/2)$ volts, respectively] in 2 ns? If for the matched inverters, $|V_t| = 0.8$ V and $k_n' = 3k_p' = 75$ μA/V^2, what are the device widths required? If the input signal is 0.2 V, what does the amplifier response time become?

15.33 A particular version of the regenerative sense amplifier of Fig. 15.20 in a 0.5-μm technology, uses transistors for which $|V_t| = 0.8$ V, $k_n' = 2.5k_p' = 100$ μA/V^2, $V_{DD} = 3.3$ V, with $(W/L)_n = 6$ μm/1.5 μm and $(W/L)_p = 15$ μm/1.5 μm. For each inverter, find the value of G_m. For a bit-line capacitance of 0.8 pF, and a delay until an output of $0.9V_{DD}$ is reached of 2 ns, find the initial difference-voltage required between the two bit lines. If the time can be relaxed by 1 ns, what input signal can be handled? With the increased delay time and with the input signal at the original level, by what percentage can the bit-line capacitance, and correspondingly the bit-line length, be increased? If the delay time required for the bit-line capacitances to charge by the constant current available from the storage cell, and thus develop the difference-voltage signal needed by the sense amplifier, was 5 ns, what does it increase to when longer lines are used?

D 15.34 (a) For the sense amplifier of Fig. 15.20, show that the time required for the bit lines to reach $0.9V_{DD}$ and $0.1V_{DD}$ is given by $t_d = (C_B/G_m)\ln(0.8V_{DD}/\Delta V)$ where

ΔV is the initial difference-voltage between the two bit lines. (Refer to Fig. 15.21.)

(b) If the response time of the sense amplifier is to be reduced to one-half the value of an original design, by what factor must the width of all transistors be increased?

(c) If for a particular design, $V_{DD} = 1.8$ V and $\Delta V = 0.2$ V, find the factor by which the width of all transistors must be increased so that ΔV is reduced by a factor of 4 while keeping t_d unchanged?

D 15.35 It is required to design a sense amplifier of the type shown in Fig. 15.20 to operate with a DRAM using the dummy-cell technique illustrated in Fig. 15.22. The DRAM cell provides readout voltages of −100 mV when a 0 is stored and +40 mV when a 1 is stored. The sense amplifier is required to provide a differential output voltage of 2 V in at most 5 ns. Find the W/L ratios of the transistors in the amplifier inverters, assuming that the processing technology is characterized by $k_n' = 2.5k_p' = 100\ \mu\text{A/V}^2$, $|V_t| = 1$ V, and $V_{DD} = 5$ V. The capacitance of each half bit line is 1 pF. What will be the amplifier response time when a 0 is read? When a 1 is read?

D 15.36 It is required to design the sense amplifier of Fig. 15.24 to detect an input signal of 100 mV and provide a full output in 0.3 ns. If $C = 60$ fF and $V_{DD} = 1.2$ V, find the required current I and the power dissipation.

D 15.37 Consider the sense amplifier in Fig. 15.24 in the equilibrium condition shown in part (b) of the figure. Let $V_{DD} = 1.8$ V and $V_t = 0.5$ V.

(a) If Q_1 and Q_2 are to operate at the edge of saturation, what is the dc voltage at the drain of Q_1?

(b) If the switching voltage ΔV is to be about 140 mV, at what overdrive voltage V_{OV} should Q_1 and Q_2 be operated in equilibrium? What dc voltage should appear at the common-source terminals of Q_1 and Q_2?

(c) If the delay component Δt given by Eq. (15.18) is to be 0.5 ns, what current I is needed if $C = 55$ fF?

(d) Find the W/L required for each of Q_1 to Q_4 for $\mu_n C_{ox} = 4\mu_p C_{ox} = 300\ \mu\text{A/V}^2$.

(e) If Q_5 is to operate at the same overdrive voltage as Q_1 and Q_2, find its required W/L and the value of the reference voltage V_R,

15.38 Consider a 512-row NOR decoder. To how many address bits does this correspond? How many output lines does it have? How many input lines does the NOR array require? How many NMOS and PMOS transistors does such a design need?

15.39 For the column decoder shown in Fig. 15.26, how many column-address bits are needed in a 256-Kbit square array? How many NMOS pass transistors are needed in the multiplexer? How many NMOS transistors are needed in the NOR decoder? How many PMOS transistors? What is the total number of NMOS and PMOS transistors needed?

15.40 Consider the use of the tree column decoder shown in Fig. 15.27 for application with a square 256-Kbit array. How many address bits are involved? How many levels of pass gates are used? How many pass transistors are there in total?

15.41 Consider a ring oscillator consisting of five inverters, each having $t_{PLH} = 6$ ns and $t_{PHL} = 4$ ns. Sketch one of the output waveforms, and specify its frequency and the percentage of the cycle during which the output is high.

15.42 A ring-of-eleven oscillator is found to operate at 20 MHz. Find the propagation delay of the inverter.

D 15.43 Design the one-shot circuit of Fig. 15.29 to provide an output pulse of 10-ns width. If the inverters available have $t_P = 2.5$ ns delay, how many inverters do you need for the delay circuit?

Section 15.5: Read-Only Memory (ROM)

15.44 Give the eight words stored in the ROM of Fig. 15.30.

D 15.45 Design the bit pattern to be stored in a (16 × 4) ROM that provides the 4-bit product of two 2-bit variables. Give a circuit implementation of the ROM array using a form similar to that of Fig. 15.30.

15.46 Consider a dynamic version of the ROM in Fig. 15.30 in which the gates of the PMOS devices are connected to a precharge control signal ϕ. Let all the NMOS devices have $W/L = 3\ \mu\text{m}/1.2\ \mu\text{m}$ and all the PMOS devices have $W/L = 12\ \mu\text{m}/1.2\ \mu\text{m}$. Assume $k_n' = 3k_p' = 90\ \mu\text{A/V}^2$, $V_{tn} = -V_{tp} = 1$ V, and $V_{DD} = 5$ V.

(a) During the precharge interval, ϕ is lowered to 0 V. Estimate the time required to charge a bit line from 0 to 5 V. Use as an average charging current the current supplied by a PMOS transistor at a bit-line voltage halfway through the 0 to 5 V excursion (i.e., 2.5 V). The bit-line capacitance is 1 pF. Note that all NMOS transistors are cut off at this time.

(b) After the precharge interval is completed and ϕ returns to V_{DD}, the row decoder raises the voltage of the selected word line. Because of the finite resistance and capacitance of the word line, the voltage rises exponentially toward V_{DD}. If the resistance of each of the polysilicon word lines is 5 kΩ and the capacitance between the word line and ground is 2 pF, what is the (10% to 90%) rise time of the word-line voltage? What is the voltage reached at the end of one time-constant?

(c) If we approximate the exponential rise of the word-line voltage by a step equal to the voltage reached in one time constant, find the interval Δt required for an NMOS transistor to discharge the bit line and lower its voltage by 1 V.