Chapter 9
Integrated Circuits

Just as the transistor revolutionized electronics by offering more flexibility, convenience, and reliability than the vacuum tube, the integrated circuit enables new applications for electronics that were not possible with discrete devices. Integration allows complex circuits consisting of many thousands of transistors, diodes, resistors, and capacitors to be included in a chip of semiconductor. This means that sophisticated circuitry can be miniaturized for use in space vehicles, in large-scale computers, and in other applications where a large collection of discrete components would be impractical. In addition to offering the advantages of miniaturization, the simultaneous fabrication of many ICs on a single Si wafer greatly reduces the cost and increases the reliability of each of the finished circuits. Certainly discrete components have played an important role in the development of electronic circuits; however, most circuits are now fabricated on the Si chip rather than with a collection of individual components. Therefore, the traditional distinctions between the roles of circuit and system designers do not apply to IC development.

In this chapter we shall discuss various types of ICs and the fabrication steps used in their production. We shall investigate techniques for building large numbers of transistors, capacitors, and resistors on a single chip of Si, as well as the interconnection, contacting, and packaging of these circuits in usable form. All the processing techniques discussed here are very basic and general. There would be no purpose in attempting a comprehensive review of all the subtleties of device fabrication in a book of this type. In fact, the only way to keep up with such an expanding field is to study the current literature. Many good reviews are suggested in the reading list at the end of this chapter; more important, current issues of those periodicals cited can be consulted for up-to-date information regarding IC technology. Having the background of this chapter, one should be able to read the current literature and thereby keep abreast of the present trends in this very important field of electronics.

9.1 BACKGROUND

In this section we provide an overview of the nature of integrated circuits and the motivation for using them. It is important to realize the reasons, both technical and economic, for the dramatic rise of ICs to their present role in electronics. We shall discuss several main types of ICs and point out some of the applications of each. More specific fabrication techniques will be presented in later sections.

9.1.1 Advantages of Integration

It might appear that building complicated circuits, involving many interconnected components on a single Si substrate, would be risky both technically and economically. In fact, however, modern techniques allow this to be done reliably and relatively inexpensively; in most cases an entire circuit on a Si chip can be produced much more inexpensively and with greater reliability than a similar circuit built up from individual components. The basic reason is that many identical circuits can be built simultaneously on a single Si wafer (Fig. 9–1); this process is called *batch fabrication*. Although the processing steps for the wafer are complex and expensive, the large number of resulting integrated circuits makes the ultimate cost of each fairly low. Furthermore, the processing steps are essentially the same for a circuit containing millions of transistors as for a simpler circuit. This drives the IC industry to build increasingly complex circuits and systems on each chip, and use larger Si wafers (e.g., 8-inch diameter). As a result, the number of components in each circuit increases without a proportional increase in the ultimate cost of the system. The implications of this principle are tremendous for circuit designers; it greatly increases the flexibility of design criteria. Unlike circuits with individual transistors and other components wired together or placed on a circuit board, ICs allow many "extra" components to be included with-

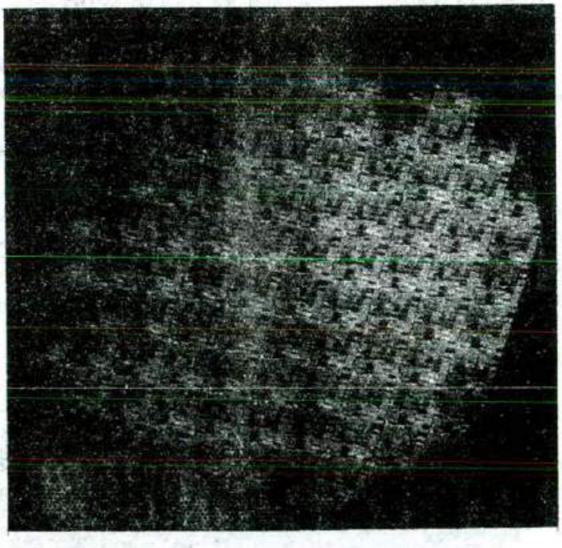


Figure 9–1
A 200 mm diameter (about 8 inch) wafer of integrated circuits. The circuits are tested on the wafer, and then sawed apart into individual chips for mounting into packages. (Photograph courtesy of IBM Corp.)

out greatly raising the cost of the final product. Reliability is also improved since all devices and interconnections are made on a single rigid substrate, greatly minimizing failures due to the soldered interconnections of discrete component circuits.

The advantages of ICs in terms of miniaturization are obvious. Since many circuit functions can be packed into a small space, complex electronic equipment can be employed in many applications where weight and space are critical, such as in aircraft or space vehicles. In large-scale computers it is now possible not only to reduce the size of the overall unit but also to facilitate maintenance by allowing for the replacement of entire circuits quickly and easily. Applications of ICs are pervasive in such consumer products as watches, calculators, automobiles, telephones, television, and appliances. Miniaturization and the cost reduction provided by ICs mean that we all have increasingly more sophisticated electronics at our disposal.

Some of the most important advantages of miniaturization pertain to response time and the speed of signal transfer between circuits. For example, in high-frequency circuits it is necessary to keep the separation of various components small to reduce time delay of signals. Similarly, in very high speed computers it is important that the various logic and information storage circuits be placed close together. Since electrical signals are ultimately limited by the speed of light (about 1 ft/ns), physical separation of the circuits can be an important limitation. As we shall see in Section 9.5, *large-scale integration* of many circuits on a Si chip has led to major reductions in computer size, thereby tremendously increasing speed and function density. In addition to decreasing the signal transfer time, integration can reduce parasitic capacitance and inductance between circuits. Reduction of these parasitics can provide significant improvement in the operating speed of the system.

We have discussed several advantages of reducing the size of each unit in the batch fabrication process, such as miniaturization, high-frequency and switching speed improvements, and cost reduction due to the large number of circuits fabricated on a single wafer. Another important advantage has to do with the percentage of usable devices (often called the *yield*) which results from batch fabrication. Faulty devices usually occur because of some defect in the Si wafer or in the fabrication steps. Defects in the Si can occur because of lattice imperfections and strains introduced in the crystal growth, cutting, and handling of the wafers. Usually such defects are extremely small, but their presence can ruin devices built on or around them. Reducing the size of each device greatly increases the chance for a given device to be free of such defects. The same is true for fabrication defects, such as the presence of a dust particle on a photolithographic mask. For example, a lattice defect or dust particle $\frac{1}{2}\mu m$ in diameter can easily ruin a circuit which includes the damaged area. If a fairly large circuit is built around the defect it will be faulty; however, if the device size is reduced so that four circuits occupy the same area on the wafer, chances are good that only the one containing the defect will be faulty and the other three will be good. Therefore, the percentage yield of usable circuits increases over a certain range of decreasing chip area. There is an optimum area for each circuit, above which defects are needlessly included and below which the elements are spaced too closely for reliable fabrication.

9.1.2 Types of Integrated Circuits

There are several ways of categorizing ICs as to their use and method of fabrication. The most common categories are *linear* or *digital* according to application, and *monolithic* or *hybrid* according to fabrication.

A linear IC is one that performs amplification or other essentially linear operations on signals. Examples of linear circuits are simple amplifiers, operational amplifiers, and analog communications circuits. Digital circuits involve logic and memory, for applications in computers, calculators, microprocessors, and the like. By far the greatest volume of ICs has been in the digital field, since large numbers of such circuits are required. Since digital circuits generally require only "on-off" operation of transistors, the design requirements for integrated digital circuits are often less stringent than for linear circuits. Although transistors can be fabricated as easily in integrated form as in discrete form, passive elements (resistors and capacitors) are usually more difficult to produce to close tolerances in ICs.

9.1.3 Monolithic and Hybrid Circuits

Integrated circuits that are included entirely on a single chip of semiconductor (usually Si) are called *monolithic* circuits (Fig. 9–1). The word monolithic literally means "one stone" and implies that the entire circuit is contained in a single piece of semiconductor. Any additions to the semiconductor sample, such as insulating layers and metallization patterns, are intimately bonded to the surface of the chip. A *hybrid* circuit may contain one or more monolithic circuits or individual transistors bonded to an insulating substrate with resistors, capacitors, or other circuit elements, with appropriate interconnections (Fig. 9–2). Monolithic circuits have the advantage that all components are contained in a single rigid structure which can be batch fabricated; that is, hundreds of identical circuits can be built simultaneously on a Si wafer. On the other hand, hybrid circuits offer excellent isolation between components and allow the use of more precise resistors and capacitors. Furthermore, hybrid circuits are often less expensive to build in small numbers.

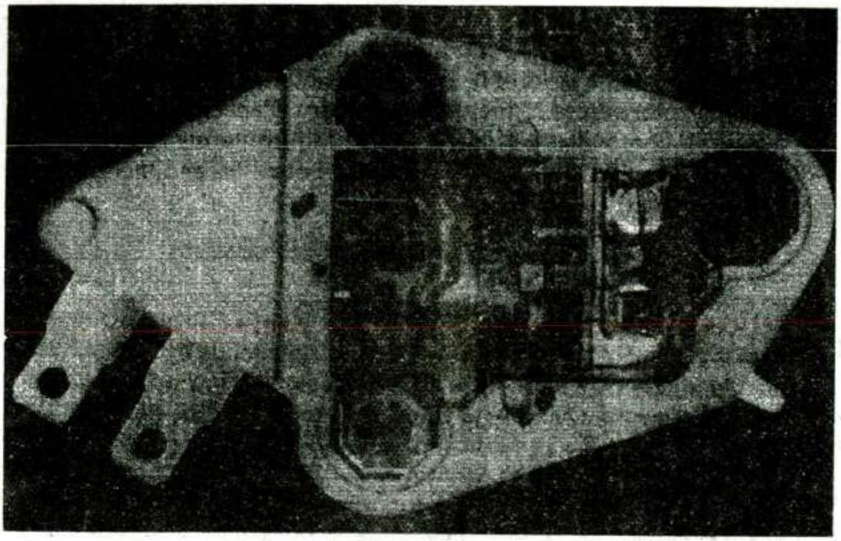


Figure 9–2
A hybrid circuit employing thick-film printing. This automobile voltage regulator circuit contains thick-film resistors and thermistors, a ceramic chip capacitor, and monolithic Si chips mounted on an insulating substrate. (Photograph courtesy of Delco Electronics Corp.)

When resistors and capacitors are made external to the monolithic Si chip, basically two types of technology are used; the passive elements are fabricated and interconnected by *thick-film* or *thin-film* processes. Although the dividing line between thin and thick films is not precise, they are fairly well separated in application to ICs: "thin" films are typically 0.1 to 0.5 μm, and "thick" films are about 25 μm.

The processing steps for the two hybrid techniques are quite different. In thick-film circuits the resistors and interconnection patterns are "printed" on a ceramic substrate (Fig. 9–2) by silk-screen or similar process. Conductive and resistive pastes consisting of metal powders in organic binders are printed on the substrate and cured in an oven. One advantage of this process is that resistors can be made below the rated values and then trimmed by abrasion, or by selective evaporation using a pulsed laser. These corrections can be made quickly with automated procedures while the resistance values are under test. Small ceramic chip capacitors can be bonded into place in the interconnection pattern, along with monolithic circuits or individual transistors.

Thin-film technology allows for greater precision and miniaturization, and is generally preferred when space is an important limitation. Thin-film interconnection patterns and resistors can be vacuum deposited on a glass or glazed ceramic substrate. The resistive films are usually made of tantalum or other resistive metal, and the conductors are often aluminum or gold. In general, the resistive materials must be deposited by sputtering. Pattern definition for the resistors and conductor paths can be achieved by depositing the films through metal shields which contain appropriate apertures. Better definition is obtained by metallizing the entire substrate, or large parts of it, and using photolithographic methods to remove the metal except in the desired pattern. Capacitors can be fabricated by thin-film techniques by depositing an insulating layer between two metal films or by oxidizing the surface of one film and then depositing a second film on top.

9.2 EVOLUTION OF INTEGRATED CIRCUITS

The IC was invented in February 1959 by Jack Kilby of Texas Instruments. The planar version of the IC was developed independently by Robert Noyce at Fairchild in July 1959. Since then the evolution of this technology has been extremely fast-paced. One way to gauge the progress of the field is to look at the complexity of ICs as a function of time. Figure 9–3a shows the number of transistors used in MOS *microprocessor* IC chips as a function of time. It is amazing that on this semi-log plot, where we have plotted log of the component count as a function of time, we get a straight line, indicating that there has been an exponential growth in the complexity of chips over three decades. The component count has roughly doubled every 18 months, as was noted early by Gordon Moore of Intel corporation. This regular doubling has become known as *Moore's law*.

The history of ICs can be described in terms of different eras, depending on the component count. Small Scale Integration (SSI) refers to the integration of 1–10^2 devices, Medium Scale Integration (MSI) between 10^2–10^3 devices, Large Scale Integration (LSI) between 10^3–10^5 devices, Very Large Scale Integration (VLSI) between 10^5–10^6 devices, and now Ultra Large Scale Integration (ULSI), where component count is between 10^6–10^9. Of course, these boundaries are somewhat fuzzy. The next generation has been dubbed as Giga-Scale Integration. Wags have suggested that after that we will have RLSI or Ridiculously Large Scale Integration.

The main factor that has enabled this increase of complexity is the ability to shrink or scale devices. Typical dimensions or feature sizes (generally the MOSFET channel lengths) of state-of-the-art *dynamic random access memories (DRAMs)* at different times are also shown as a semi-log plot in Fig. 9–3b. Once again, we see a straight line, reflecting an exponential decrease of the typical feature sizes with time over three decades. Clearly, one can pack a larger number of components with greater functionality on an IC if they are smaller. As discussed in Section 6.5.9, scaling also has other advantages in terms of faster ICs which consume less power.

While scaling represents an opportunity, it also presents tremendous technological challenges. The most notable among these challenges lie in lithography and etching, as discussed in Section 5.1. However, since scaling of horizontal dimensions also requires scaling of vertical geometries as discussed in Section 6.5.9, there are also tremendous challenges in terms of doping, gate dielectrics and metallization. Small features and large chips also require device fabrication in extremely clean environments. Particles which may not have caused yield problems in a 1 μm IC technology can have catastrophic effects for a 0.25 μm process. This requires purer chemicals, cleaner equipment, and more stringent clean rooms. In fact, the levels of cleanliness required bypassed the best surgical operating rooms early in the evolution shown in Fig. 9–3. The cleanliness of these facilities is designated by the *Class* of the clean room. For instance, a Class 1 clean room, which is state-of-the-art in 2000, has less than 1 particle of size 0.2 μm or larger per cubic foot. There are more of the smaller particles, and fewer of the larger ones. Obviously, the lower the Class of a clean room, the better it is. A Class 1 clean room is much cleaner than a Class 100 fabrication facility or "fab." As one might expect, such high levels of cleanliness come with a hefty price tag. A state-of-the-art fab in 2000 comes equipped with a price tag of about two billion dollars.

In spite of the costs, the economic payoff for ULSI is tremendous. Just for calibration, let us examine some economic statistics at the dawn of the third millennium. The total annual economic output of all the countries in the world, or the so-called Gross World Product (GWP), is about 35 trillion US dollars. The US Gross National Product is about 8 trillion dollars, or about a quarter of the GWP. The worldwide IC industry is about 150 billion dollars, and the entire worldwide electronics industry in which these ICs participate is about 1 trillion dollars. As a single industry, electronics is one of the biggest in terms of the dollar amount. It has surpassed, for example, automobiles (worldwide sales of about 50 million cars annually) and petrochemicals. About 100–200 million personal computers are sold annually worldwide.

Perhaps even more dramatic than these raw economic numbers is the growth rate of these markets. If one were to plot IC sales as a function of time, one again finds a more or less exponential increase of sales with time over three decades. Of great importance to the consumer, the cost per electronic function has dropped dramatically over the same period of time. For example, the cost per bit of semiconductor memory (DRAM) has dropped from about 1 cent/bit in 1970 to about 10^{-4} cents per bit today, a cost improvement of four orders of magnitude in 30 years. There are no parallels in any other industry for this consistent improvement in functionality with such lowered cost.

While ICs started with bipolar processes in the 1960s, they were gradually supplanted by MOS and then CMOS devices, for reasons discussed in Chapters 6 and 7. Currently, about 88 percent of the IC market is MOS-based, and about 8 percent BJT-based. Optoelectronic devices based on compound semiconductors are still a relatively small component of the semiconductor market (about 4 percent), but are expected to grow in the future. Of the MOS ICs, the bulk are digital ICs. Of the entire semiconductor industry, only about 14 percent are analog ICs. Semiconductor memories such as DRAMs, SRAMs and non-volatile flash memories make up approximately 25 percent of the market, microprocessors about 25 percent, and other application-specific ICs (ASICs) about 20 percent.

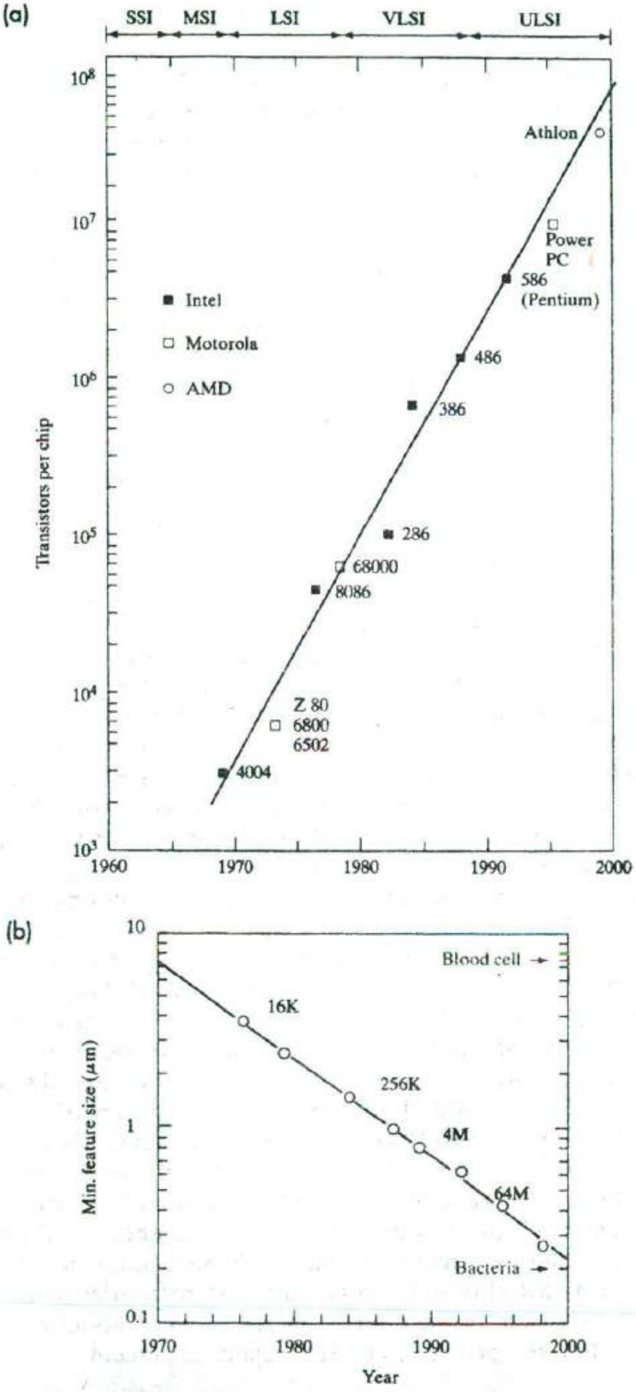


Figure 9–3 Moore's law for integrated circuits: (a) Exponential increase of transistor count as a function of time for different generations of microprocessors; (b) exponential decrease of minimum transistor gate lengths with time, for different generations of dynamic random access memories (16 kb to 256 Mb DRAMs). For reference, sizes of blood cells and bacteria are shown on the μm scale.

9.3 MONOLITHIC DEVICE ELEMENTS

Now we shall consider the various elements that make up an integrated circuit, and some of the steps in their fabrication. The basic elements are fairly easy to name—transistors, resistors, capacitors, and some form of interconnection. There are some elements in integrated circuits, however, which do not have simple counterparts in discrete devices. We shall consider one of these, charge transfer devices, in Section 9.4. Discussion of fabrication technology is difficult in a book of this type, since device fabrication engineers seem to make changes faster than typesetters do! Since this important and fascinating field is changing so rapidly, the reader should obtain a basic understanding of device design and processing from this discussion and then search out new innovations in the current literature.

9.3.1 CMOS Process Integration

A particularly useful device for digital applications is a combination of n-channel and p-channel MOS transistors on adjacent regions of the chip. This *complementary MOS* (commonly called *CMOS*) combination is illustrated in the basic inverter circuit of Fig. 9–4a. In this circuit the drains of the two transistors are connected together and form the output, while the input terminal is the common connection to the transistor gates. The p-channel device has a negative threshold voltage, and the n-channel transistor has a positive threshold voltage. Therefore, a zero voltage input ($V_{in} = 0$) gives

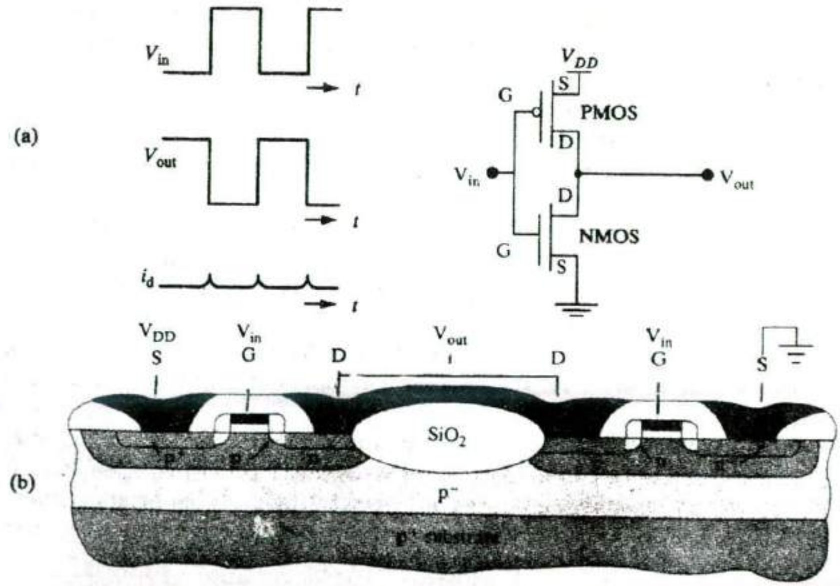


Figure 9–4
Complementary MOS structure: (a) CMOS inverter; (b) formation of p-channel and n-channel devices together.

zero gate voltage for the n-channel device, but the voltage between the gate and source of the p-channel device is $-V$. Thus the p-channel device is on, the n-channel device is off, and the full voltage V is measured at V_{out} (i.e., V appears across the nonconducting n-channel transistor). Alternatively, a positive value of V_{in} turns the n-channel transistor on, and the p-channel off. The output voltage measured across the "on" n-channel device is essentially zero. Thus, the circuit operates as an inverter—with a binary "1" at the input, the output is in the "0" state, whereas a "0" input produces a "1" output. The beauty of this circuit is that one of the devices is turned off for either condition. Since the devices are connected in series, no drain current flows, except for a small charging current during the switching process from one state to the other. Since the CMOS inverter uses ultra little power, it is particularly useful in applications such as electronic watch circuits which depend on very low power consumption. CMOS is also advantageous in ultra large scale integrated circuits (Section 9.5), since even small power dissipation in each transistor becomes a problem when millions of them are integrated on a chip.

The device technology for achieving CMOS circuits consists mainly in arranging for both n- and p-channel devices with similar threshold voltages on the same chip. To achieve this goal, a diffusion or implantation must be performed in certain areas to obtain n and p regions for the fabrication of each type of device. These regions are called *tubs, tanks,* or *wells* (Fig. 9–5). The critical parameter of the tub is its net doping concentration, which must be closely controlled by ion implantation. With the tub in place, source and drain implants are performed to make the n-channel and p-channel transistors. Matching of the two transistors is achieved by control of the surface doping in the tub and by threshold adjustment of both transistors by ion implantation.

Including bipolar transistors in the basic CMOS technology allows flexibility in circuit design, particularly for providing drive currents. The combination of bipolar and CMOS (called BiCMOS) provides circuits with increased speed.

Attention must be paid in CMOS designs to the fact that combining n-channel and p-channel devices in close proximity can lead to inadvertent (*parasitic*) bipolar structures. In fact, a p-n-p-n structure can be found in Fig. 9–4b, which can serve as an inefficient but troublesome *thyristor* (see Ch. 11). Under certain biasing conditions the p-n-p part of the structure can supply base current to the n-p-n structure, causing a large current to flow. This process, called *latchup*, can be a serious problem in CMOS circuits. Several methods have been used to eliminate the latchup problem, including using both n-type and p-type tubs, separated by trench isolation (Fig. 9–6). The use of two separate tubs (wells) also allows independent control of threshold voltages in both types of transistor.

We can illustrate most of the common fabrication steps for MOS integrated circuits by studying the flow of a twin well Self-Aligned siLICIDE (SALICIDE) CMOS process. This process is particularly important because most high-performance digital ICs, including microprocessors, memories and application specific ICs (ASICs), are made basically in this way. In order to make enhancement-mode n-channel devices, we need a p-type substrate, and vice versa. Since CMOS requires both, we must start either with an n-type or a p-type wafer and then make selected regions of the substrate have opposite doping by forming wells. For example, Fig. 9–5a shows a lightly doped p-epitaxial layer on a p^+-substrate. We can make n-channel devices in this layer.

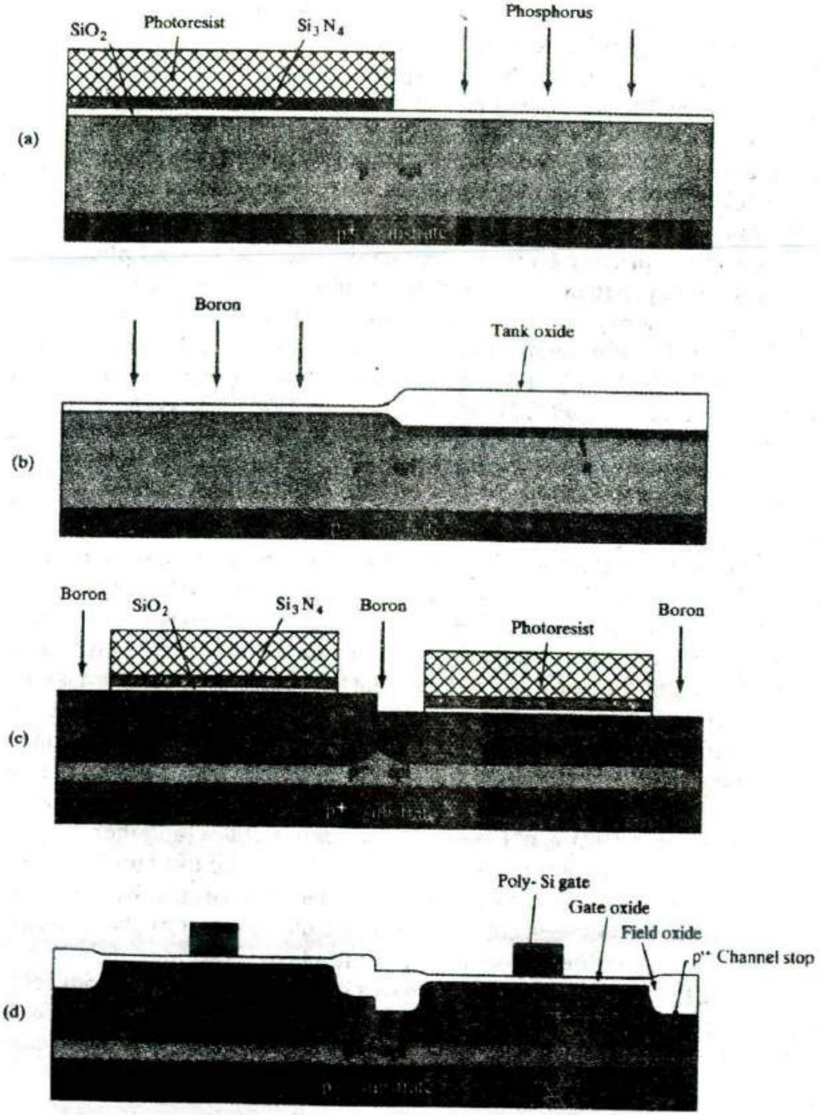


Figure 9–5
Self-aligned twin well process: (a) n-well formation using P donor implant and a photoresist mask; (b) p-well formation using B acceptor implant. A thick (~200 nm) "tank" oxide layer is grown wherever the silicon nitride–oxide stack is etched off, and the tank oxide is used to block the B implant in the n-wells in a self-aligned manner; (c) isolation pattern for field transistors showing B channel stop implant using photoresist mask; (d) local oxidation of silicon wherever nitride mask is removed, leading to thick LOCOS field oxide.

By implanting n-wells wherever needed, we can make p-channel devices also. This is an n-well CMOS process. Alternatively, if we start with an n-substrate and make p-wells in certain regions, we have a p-well CMOS process. For optimal device performance, however, it is usually desirable to separately implant both the n- and the p-well regions, which is called *twin-well* CMOS. The rationale for this can be appreciated if we keep in mind that for a state-of-the-art IC, typical doping levels are ~10^{18} cm^{-3} and junction depths are ~1 μm in these wells. The doping levels have to be high enough to prevent punchthrough breakdown due to drain-induced barrier lowering (DIBL) in the MOSFETs, but low enough to keep the threshold voltages acceptably small. If we choose to use the p-substrate for n-channel devices as in Fig. 9–5a, the p-type epitaxial layer must be doped to 10^{18} cm^{-3}, and the implanted n-type layer must be achieved by counter-doping at a level of ~2×10^{18} cm^{-3}, resulting in a net n-type doping of ~1×10^{18} cm^{-3}, but a total doping in this region of ~3×10^{18} cm^{-3}. Such high levels of total doping are detrimental to carrier transport because they cause excessive ionized impurity scattering. Hence, for high performance ICs, the starting epitaxial doping level is generally very low (~10^{16} cm^{-3}). This layer is grown on a heavily doped substrate (~10^{19} cm^{-3}), to provide a highly conducting electrical ground plane. This helps with noise problems in ICs and helps minimize the problem of *latchup* by bypassing majority carriers (in this case holes) to the p^+ substrate.

To form the twin wells in a self-aligned fashion, we first grow thermally a "pad" oxide (~20 nm) on the Si substrate, followed by low pressure chemical vapor deposition (LPCVD) of silicon nitride (~20 nm). As shown in Fig. 9–5a this oxide–nitride stack is covered by photoresist, and a window is opened for the n-well. Reactive ion etching (RIE) is then used to etch the oxide–nitride stack. Using the photoresist as an implant mask, we then do an n-type implant using phosphorus. Phosphorus is preferred to As for this purpose because P is lighter and has a higher projected range; also, P diffuses faster. This fast diffusion is needed to drive the dopants fairly deep into the substrate to form the n-well. After the implant, the photoresist is removed, and the patterned wafer is subjected to wet oxidation to grow a "tank" oxide (~200 nm). It may be noted in Fig. 9–5b that the tank oxidation process consumes Si from the substrate, and the resulting oxide swells up. In fact, for every micron of thermally-grown oxide, the oxidation consumes 0.44 μm of Si, resulting in a 2.2× volume expansion. The oxide does not grow in the regions that are protected by silicon nitride because nitride has the property that it blocks the diffusion of oxygen and water molecules (and thereby prevents oxidation of the Si substrate). The pad oxide that is used under the nitride has two roles: it minimizes the thermal-expansion mismatch and concomitant stress between silicon nitride and the substrate; it also prevents chemical bonding of the silicon nitride to the silicon substrate.

Using the tank oxide as a *self-aligned* implant mask (i.e., without actually having to do a separate photolithographic step), one does a p-type well implant using boron (Fig. 9–5b). The tank oxide must thus be much thicker than the projected range of the B. The concept of self-alignment is very important, and is a recurring theme in IC processing. It is simpler and cheaper to use self-alignment than a separate lithographic step. It also allows tighter packing density of the twin wells, because it is not required to account for lithographic misalignment during layout. The P and the B are then diffused into the substrate to a well depth of typically a micron by a drive-in diffusion at very high tempatures (~1000 °C) for several hours. After this diffusion, the silicon nitride–oxide stack and the tank oxide are etched away. Since the tank oxidation consumes Si from the substrate, etching it off leads to a step in the Si substrate delineating the n-well and p-well regions. This step is important in terms of alignment of subsequent reticles, and is shown in an exaggerated fashion in Fig. 9–5c.

Next, we form the isolation regions or the field transistors which guarantee that there will be no electrical cross-talk between adjacent transistors, unless they are interconnected intentionally (Fig. 9–5c). This is achieved by ensuring that the threshold voltage of any parasitic transistor that may form in the isolation regions is much higher than the power supply voltage on the chip, so that the parasitic channel can never turn on under operating conditions. From the threshold voltage expression (Eq. 6-38), we notice that V_T can be raised by increasing substrate doping and increasing gate oxide thickness. However, a problem with that approach is the subthreshold slope S (Eq. 6-66), which degrades with increasing substrate doping and gate oxide thickness. One needs to optimize both V_T and S such that the off-state leakage current in the field between transistors is sufficiently low at zero gate bias.

A stack of silicon dioxide–silicon nitride is photolithographically patterned as in Fig. 9–5c and subjected to RIE. A boron *"channel stop"* implant between the twin wells increases the acceptor doping and thus increases the threshold voltage in the p-well between the n-channel transistors (the *field* threshold). However, B will compensate the donor doping on the n-well side, and thus reduce the threshold in the n-well between p-channel devices. The B channel stop dose must thus be optimized to have acceptably high field thresholds in both types of wells.

After the channel stop implant, the photoresist is removed and the wafer with the patterned nitride–oxide stack shown in Fig. 9–5c (without photoresist) is subjected to wet oxidation to selectively grow a field oxide ~300 nm thick. The nitride layer blocks oxidation of the Si substrate in the regions where we plan to make the transistors. This procedure, where Si is oxidized to form SiO_2 in regions not protected by nitride, is called LOCal Oxidation of Silicon (LOCOS) (Sec. 6.4.1). In this case, LOCOS provides electrical isolation between the two transistors, as shown in Fig. 9–5d.

The volume expansion of 2.2× upon oxidation is an important issue because the selective oxidation occurs in narrow, confined regions. The compressive stress, if excessive, can cause dislocation defects in the substrate. Another issue is the lateral oxidation near the nitride mask edges, which causes the nitride mask to lift up near the edges, forming what is known as a *bird's beak* and causing a *lateral moat encroachment* of ~0.2 μm into each active region, thereby wasting precious Si real estate. There have been various modified LOCOS and other isolation schemes proposed to minimize this lateral encroachment. A notable example is *Shallow Trench Isolation* (STI) which involves using RIE to etch a shallow (~1 μm) trench or groove in the Si substrate after the isolation pattern, filling it up completely by deposition of a dielectric layer of SiO_2 and polysilicon by Low Pressure Chemical Vapor Deposition (LPCVD), and then using Chemical Mechanical Polishing (CMP) to planarize the structure (Fig. 9–6). This consumes less Si real estate compared to LOCOS, but gives superior isolation because the sharp corners at the bottom of the trench give rise to potential barriers that block leakage currents (the *corner* effect).

The pad oxide between the nitride and the Si surface minimizes the stress due to the nitride, and prevents bonding of the nitride to the Si, as mentioned above. Any residual nitride on the Si would retard subsequent gate

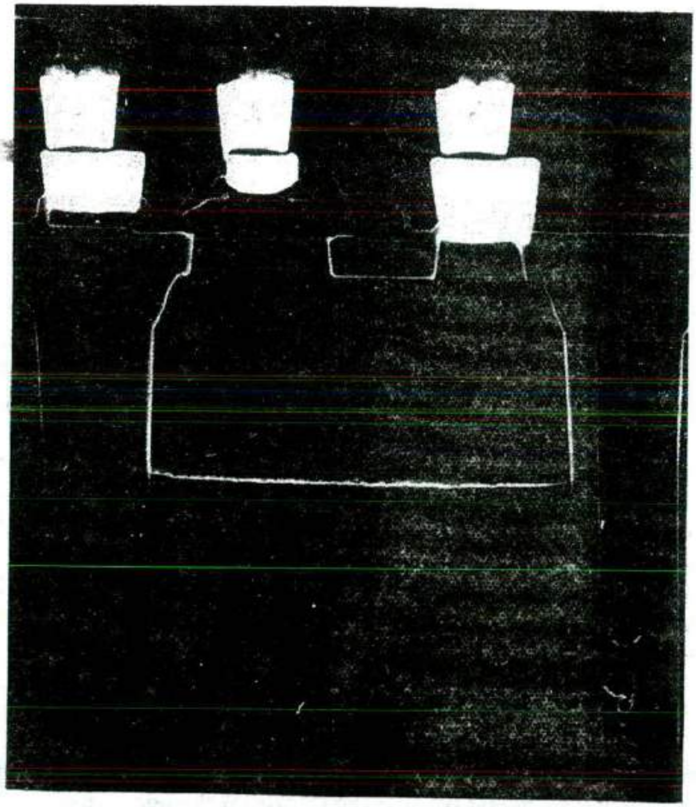


Figure 9–6
Trench isolation: A trench or groove is etched in the substrate using RIE, and refilled with oxide and polysilicon, providing superior electrical isolation compared to LOCOS, using less Si real estate.

oxide formation, leading to weak spots in the gate region of the MOSFETs. This problem is known as the *white ribbon* effect or the *Kooi* effect, after the Dutch scientist who first identified it. The pad oxide mitigates this problem, but does not solve it completely. Therefore, very often a "sacrificial" or "dummy" oxide is grown to consume a layer of Si containing any residual nitride, and this oxide is wet etched prior to the growth of the actual gate oxide.

Next, an ultra-thin (~5–10 nm) gate oxide is grown on the substrate. Since the electrical quality of this oxide and its interface with the Si substrate is of paramount importance to the operation of the MOSFETs, dry oxidation is used for this step. It is common to incorporate some nitrogen at the Si–SiO_2 interface, forming oxy-nitrides which improve the interface quality in terms of hot electron effects. After the oxidation, it is immediately covered with LPCVD polysilicon in order to minimize contamination of the gate oxide. The polysilicon gate layer is doped very heavily (typically n^+ using a phosphorus dopant source, $POCl_3$ in a diffusion furnace) all the way to the polysilicon–oxide interface in order to make it behave electrically like a metal electrode. Alternatively, the LPCVD polysilicon film may also be *in situ* doped during the deposition itself by flowing in an appropriate dopant gas such as phosphine or diborane. Heavy doping of the gate material is very important, because otherwise a depletion layer can be formed in the polysilicon gate (the *poly depletion* effect). This could result in a depletion capacitance in series with the gate oxide capacitance, thereby reducing the overall gate capacitance and, therefore, the drive current (see Eq. 6-53). The high doping ($\sim 10^{20}$ cm^{-3}) in the polysilicon gate is also important for reducing the resistance of the gate and its *RC* time constant. The uniformly high doping in the polysilicon layer is facilitated by the presence of the grain boundary defects in the film, because diffusivity of dopants along grain boundaries is many orders of magnitude higher than in single crystal Si.

The doped polysilicon layer is then patterned to form the gates, and etched anisotropically by RIE to achieve vertical sidewalls. That is extremely important because this etched polysilicon gate is used as a self-aligned implant mask for the source/drain implants. As mentioned above, self-aligned processes are always desirable in terms of process simplicity and packing density. It is particularly useful in this case because we thereby guarantee that there will be *some* overlap of the gate with the source/drain but minimal overlap. The overlap is determined by the lateral scattering of the ions and by the lateral diffusion of the dopants during subsequent thermal processing (such as source/drain implant anneals). If there were no overlap, the channel would have to be turned on in this region by the gate fringing fields. The resulting potential barrier in the channel would degrade the device current. On the other hand, if there is too much overlap, it leads to an overlap capacitance between the source or drain and the gate. This is particularly bothersome near the drain end because it leads to the Miller overlap capacitance which causes undesired capacitive feedback between the output drain terminal and the input gate terminal (see Section 6.5.8).

Fabrication steps for the n-channel MOSFETs in the p-well are shown in Fig. 9–7. After the polysilicon gate is etched, we first do a self-aligned n-type source–drain implant, during which the tank masking level is used to protect the PMOS devices with a layer of photoresist. The NMOS source and drain implants are done in two stages. The first implant is a lightly doped drain (LDD) implant (Fig. 9–7a). This is typically a dose of ~10^{13}–10^{14} cm^{-2}, corresponding to a concentration of 10^{18}–10^{19} cm^{-3}, and an ultra-shallow junction depth of 50–100 nm. When a MOSFET is operated in the saturation region, the drain-channel junction is reverse biased, resulting in a very high electric field in the pinch-off region. As we saw in Section 5.4 for reverse-biased p-n junctions, reducing the doping level increases the depletion width and makes the peak electric field at the junction smaller. As discussed in Section 6.5.9, electrons traveling from the source to the drain in the channel can gain kinetic energy and thereby become hot electrons, which create damage. The low doping in the LDD helps reduce hot carrier effects at the drain end. The shallow junction depths in the LDD are also important for reducing short channel effects such as DIBL and charge sharing (Sections 6.5.10 and 6.5.11). The penalty that we pay with the use of an LDD region is that the source-to-drain series resistance goes up, which degrades the drive current.

As the technology is evolving towards lower power supply voltages, hot carrier effects are becoming less important. This, along with the need to reduce series resistance, has driven the trend towards increasing doping in the LDD to levels above 10^{19} cm^{-3}. In fact, the use of the term LDD then becomes a bit of a misnomer, and is often replaced by the term *source/drain extension* or *tip*.

After the LDD regions are formed alongside the polysilicon gate, we implant deeper (~200 nm) and more heavily doped (10^{20} cm^{-3}) source and drain junctions farther away from the gate edges (Fig. 9–7d). This more conductive region allows ohmic contacts to the source and drain to be formed more easily than they could be directly to the LDD regions, and reduces the source/drain series resistance. This implant is done using a self-aligned scheme by the formation of *sidewall oxide spacers*. After removing the photoresist covering the PMOS devices, we deposit *conformal* LPCVD oxide (~100–200 nm thick) using an organic precursor called tetra-ethyl-ortho-silicate (TEOS) over the entire wafer at fairly high temperatures (~700 °C) (Fig. 9–7b). The term conformal means that the deposited film has the same thickness everywhere, and follows the topography on the wafer. This oxide layer is then subjected to RIE, which is anisotropic (i.e., it etches predominantly in the vertical direction) (Section 5.1.7). If the RIE step is timed to just etch off the deposited oxide on the flat surfaces, it leaves oxide sidewall spacers on the edges of the polysilicon gate, as shown in Fig. 9–7c. This sidewall spacer is used as a self-aligned mask to protect the LDD regions very near the gate during the heavier, deeper n$^+$ source and drain implants (Fig. 9–7d).

Next, the NMOS devices are masked by photoresist, and a p$^+$ source and drain implant is done for the PMOSFETs (Fig. 9–8a). It may be noted that an

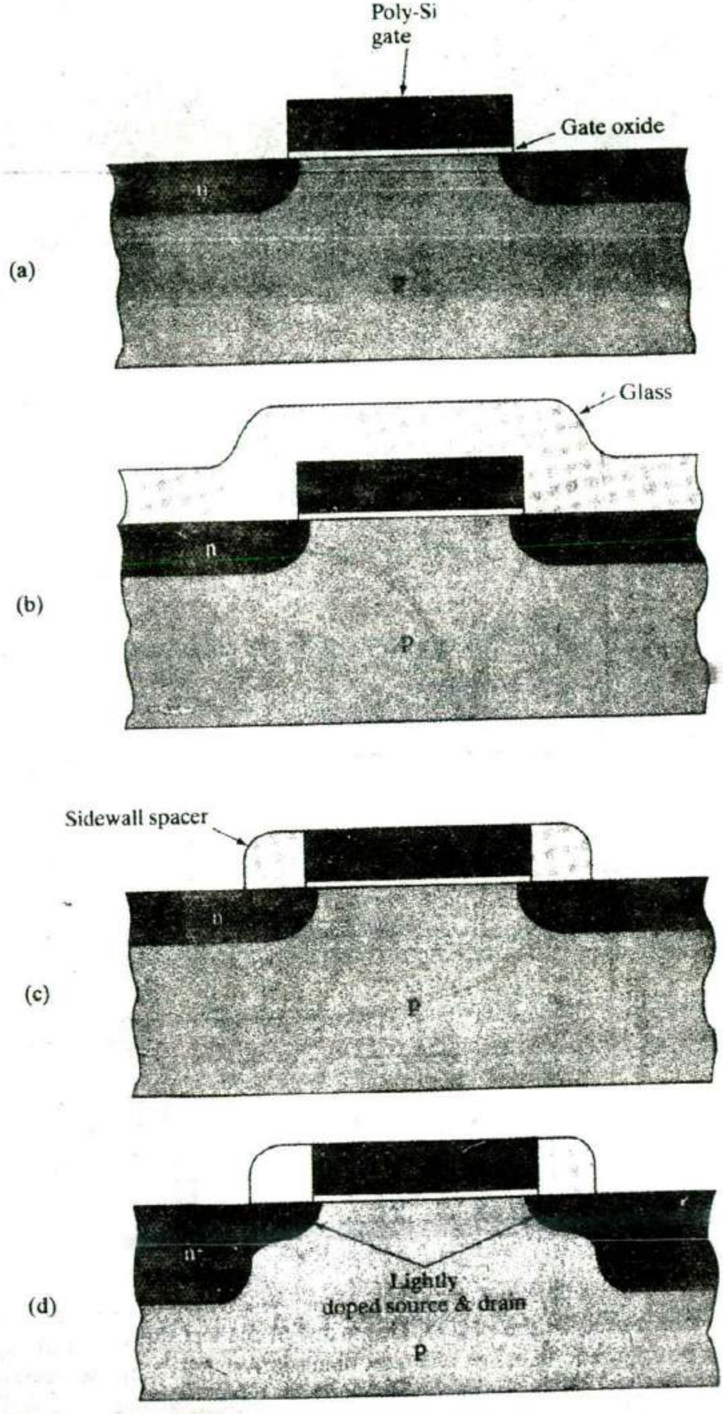


Figure 9–7 Fabrication of the lightly doped drain structure, using sidewall spacers. The polysilicon gate covers the thin gate oxide and masks the first low-dose implant (a). A thick layer is deposited by CVD (b) and is anisotropically etched away to leave only the sidewall spacers (c). These spacers serve as a mask for the second, high-dose implant. After a drive-in diffusion, the LDD structure results (d).

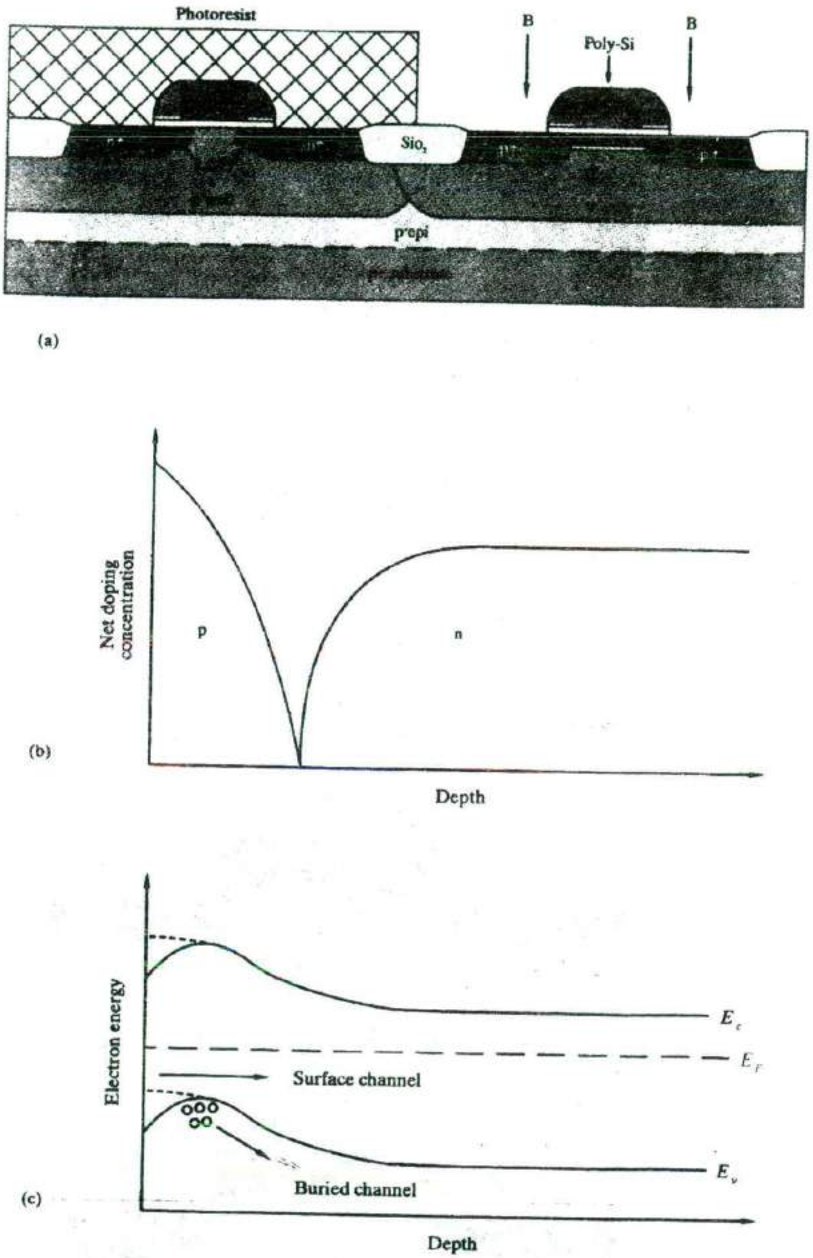


Figure 9–8 Buried channel PMOS: (a) self-aligned p⁺ source/drain implant with no LDD using photoresist to protect NMOSFETs. A p-type V_T adjust implant is shown in color in the channel; (b) doping profile as a function of depth in the middle of the channel showing the p-type V_T adjust implant near the surface; (c) electron potential energy as a function of depth in the middle of the channel, showing holes collecting in the "buried" channel. For higher gate bias the PMOS operation changes to surface channel, as indicated by the dashed line.

LDD was not used for the PMOS. This is due to the fact that hot hole effects are less problematic than hot electron degradation, partly due to the lower hole mobility and partly due to the higher Si–SiO_2 barrier in the valence band

(5 eV) than in the conduction band (3.1 eV). After the source–drain implants are done, the dopants are activated and the ion implant damage is healed by a furnace anneal, or more frequently by using a Rapid Thermal Anneal (RTA). In this anneal we use the minimum acceptable temperature and time combination (the thermal budget) because it is critically important to keep the dopant profiles as compact as possible in ultra-small MOSFETs.

We can now appreciate why most CMOS logic devices are made on p-type substrates, rather than n-type. The n-channel MOSFETs generate a lot more substrate current due to hot carrier effects than PMOSFETs. The holes, thus generated, can more easily flow to ground in a p-type substrate, than in an n-substrate. Also, it is easier to dope substrates p-type with B during Czochralski crystal growth than n-type with Sb. Antimony is the preferred donor, rather than As or P for bulk doping of the Si melt, because Sb evaporates less than the other species.

The use of n^+ polysilicon gates for both NMOS and PMOS devices raises some interesting device issues. Since the Fermi level in the n^+ gate is very close to the Si conduction band, its workfunction is well suited to achieving a low V_T for NMOS (Φ_{ms}~–1V), but not for PMOS (Φ_{ms}~0V). From the V_T expression (Eq. 6-38), we notice that the second and third terms approach zero as thin-oxide technology evolves because C_i is getting larger. For high drive current we want V_T to be in the neighborhood of ~0.3 to ~0.7V for NMOS (–0.3 to –0.7 V for PMOS). We find from Eq. (6–38) that the p-well doping can be optimized to achieve the correct V_T for the NMOS transistor, while at the same time being high enough to prevent punchthrough breakdown between source and drain. For the PMOS transistor, on the other hand, an n-well doping of the order of 10^{18} cm^{-3} prevents punchthrough, but the Fermi potential, ϕ_F, is so large and negative that the V_T is too negative. That forces us to do a separate acceptor implant to adjust V_T for the PMOS devices (Fig. 9–8b). The acceptor dose is low enough that the p-layer is fully depleted at zero gate bias, leading to enhancement mode, rather than depletion mode transistors. In CMOS we try to make the negative V_T of the PMOS device about the same value as the positive V_T of the NMOS.

Close examination of the band diagram in the channel of the PMOSFET along the vertical direction (perpendicular to the gate oxide) shows that the energy minimum for the holes in the inversion layer is slightly below (~100 nm) the oxide–silicon interface, leading to what is known as *buried channel operation* for PMOS (Fig. 9–8c). On the other hand, for NMOS, the electron energy minimum in the inversion layer occurs right at the oxide–silicon interface, leading to surface channel operation.[1] There are good and bad aspects of this

[1] We can qualitatively understand why the acceptor implant in the channel leads to such buried channel behavior. Assume for a moment that the acceptor dose was high enough that the p-layer was not depleted at zero bias. In such a depletion mode device, a positive gate bias to turn off the device would first deplete the surface region of holes, still leading to hole conduction deeper in the substrate away from the oxide–Si interface.

buried channel behavior for PMOS. Since the holes in the inversion layer of the PMOSFET travel slightly away from the oxide–silicon interface, they do not suffer as much channel mobility degradation as the electrons in the NMOSFET due to surface roughness scattering. That is good, because hole mobilities in Si are generally lower than electron mobilities, which forces us to make PMOS devices wider than the NMOSFETs to get similar drive currents. However, buried channel devices have a greater propensity towards DIBL and punchthrough breakdown. Hence, as the size of MOSFETs is reduced, the DIBL problem becomes worse, and there is a desire to have surface channel operation for both NMOS and PMOS. Thus, there is interest in so-called *dual gate* CMOS, where n^+ polysilicon gates are used for NMOSFETs and p^+ gates are used for PMOSFETs. Such dual workfunction gates can be achieved by depositing the polysilicon undoped and then using the source and drain implants themselves to also dope the gates appropriately. This approach exploits the high polysilicon grain boundary diffusivities to degenerately dope the gates, while at the same time having ultra-shallow source and drain junctions to minimize DIBL.

As a historical footnote, it may be added that MOSFETs initially used Al gates which could not withstand high temperature processing. Hence, the source and drain regions had to be formed first, either by diffusion or by implant, and then the Al gate was deposited and patterned. Such non–self-aligned processes suffered from the Miller capacitance mentioned previously. What made self-aligned source and drain regions viable was the use of polysilicon, which has a sufficiently high melting point to withstand subsequent processing. Recent research on MOSFET technology is going back full circle to metal gates, but this time using refractory metals such as tungsten (W). These metals have better conductivity than heavily doped polysilicon, and a workfunction that is better suited for CMOS. The Fermi level of W is near the middle of the Si bandgap, which makes the flatband voltage and the threshold voltage more symmetric and better matched between NMOS and PMOS, and avoids the buried channel effect.

The next step is to form a metal-silicon alloy or silicide in the source/drain and gate regions of the MOSFETs in order to reduce the series resistance (and thereby the *RC* time constants), and increase the drive current. This involves depositing a thin layer of a refractory metal such as Ti over the entire wafer by sputtering, and reacting the Ti with Si wherever they come in *direct* contact, by doing a two-step heat treatment in a N ambient (Fig. 9–9a). A 600 °C anneal results in the formation of Ti_2Si (the C49 phase according to metallurgists, which has fairly high resistivity), followed by an 800 °C anneal which converts the Ti_2Si to the C54 phase, $TiSi_2$, which has an extremely low resistivity of ~17 $\mu\Omega$-cm, much lower than that of the most heavily doped Si. On the other hand, the Ti on top of the sidewall oxide spacers does not form a silicide, and stays as unreacted Ti or forms TiN because the process is done in a N ambient. The Ti and TiN can be etched off selectively using a wet hydrogen peroxide-based etch which does not attack the titanium disilicide, thereby electrically isolating the gates from the source/drains. It may be noted that this process results in a SALICIDE

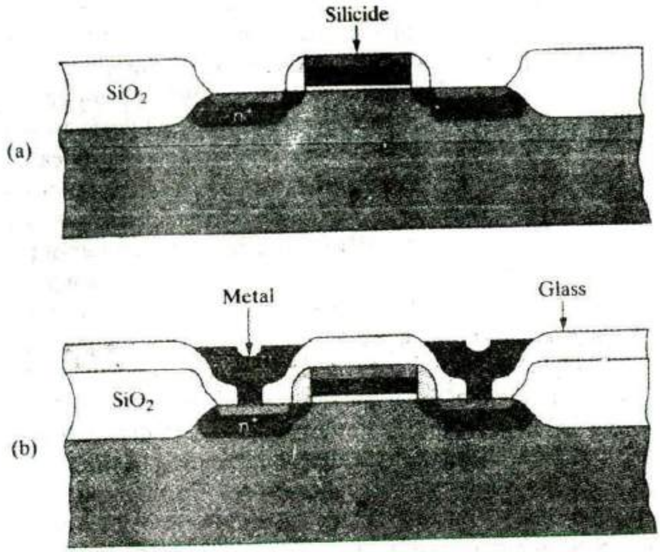


Figure 9–9
The formation of silicided source–drain and gate regions for low-resistance contacts: (a) a layer of refractory metal is deposited and reacted in regions of exposed Si to form a conducting silicide layer; (b) the unreacted metal is removed and a CVD glass is deposited and patterned for contact metallization.

without a separate masking level only on the source/drains and the polysilicon gate, where the silicide is often termed a *polycide*. This results in very high performance MOSFETs.

Finally, the MOSFETs have to be properly interconnected according to the circuit layout, using the metallization level. This involves LPCVD of an oxide dielectric layer doped with B and P, which is known as *boro-phosphosilicate glass* (BPSG) on the entire wafer, patterning it using the contact level reticle and using RIE to open up the contact holes to the substrate (Fig. 9–9b). The B and P allow the oxide layer to soften and reflow more readily upon annealing, thereby helping planarize or smooth out the topography on the wafer. This shaping of the millions of very small contact holes is critical on a ULSI chip, because otherwise metal deposited on the surface into the contact holes may not reach completely into the holes, leading to a catastrophic open circuit. In fact, sometimes a CVD tungsten layer is selectively deposited in the contact holes to form a contact *plug* before proceeding to the next step. Then a suitable metal layer such as Al (alloyed with ~1 percent Si, and ~4 percent Cu) is sputter deposited over the wafer, patterned using the metal interconnect level, and subjected to metal RIE. The Si is added to the Al to solve the junction *spiking* problem, where pure Al can incorporate the solid solubility limit of Si from the shallow source/drain regions. This would allow the Al to "spike" or short through the p-n junction. The Cu is added to enlarge the grain size in the Al interconnect films, which are polycrystalline, making it harder for the electrons moving during current flow to nudge the Al atoms along, thereby opening voids (open circuits) in the interconnect. This is an example of an *electromigration* phenomenon.

In a modern ULSI chip, the complexity of the device layout generally demands that multiple levels of metallization be used for interconnecting the devices (Fig. 9–10). Hence, after depositing the first metal, an inter-metal dielectric isolation layer such as SiO_2 is deposited by low temperature CVD. Low temperatures are very important in this *back-end* part of the processing because by now all the active devices are in place and one cannot allow the dopants to diffuse significantly. Also, the Al metallization cannot withstand temperatures higher than ~500 °C. The dielectric isolation layer must be suitably planarized prior to the deposition of the next layer of metal, and this is generally done by CMP. Planarization is important because if metal is deposited on a surface with rough topography and subjected to RIE, there can be residual metal sidewall filaments or "stringers" at the steps for the same reason one gets sidewall oxide spacers on either side of the MOSFET gate in Fig. 9–7. These metal stringers can cause short circuits between adjacent metal lines. Planarization is also important in maintaining good depth of focus during photolithography. After planarization of the isolation layer, one uses photolithography to open up a new set of contact holes called *vias*, followed by deposition, patterning and RIE of the next layer of metal, and so on for multi-level metallization. As mentioned previously, W metal plugs are sometimes selectively deposited to fill up the via holes prior to the metal deposition, and reduce the likelihood of an open circuit.

Finally, a protective overcoat is deposited on the IC to prevent contamination and failure of the devices due to the ambient (Fig. 9–10). This generally involves plasma CVD of silicon nitride, which has the nice attribute that it blocks the diffusion of water vapor and Na through it. Sodium, as mentioned in Section 6.4.3, causes a mobile ion problem in the gate dielectric of MOS devices. Sometimes, the protective overcoat is a BPSG layer. After the

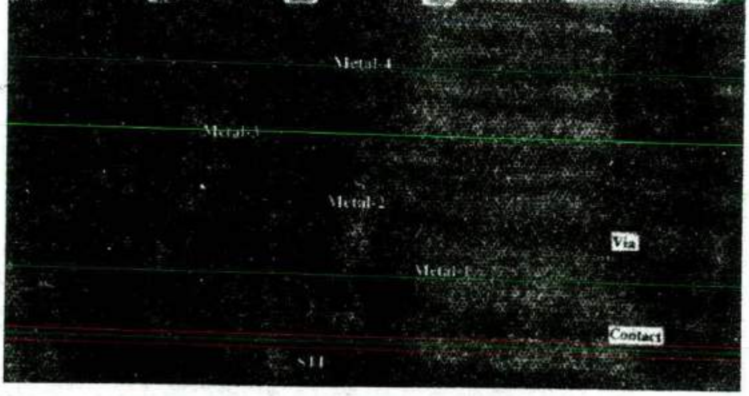


Figure 9–10 Multi-level interconnect: Cross section of IC showing 5 levels of A1 interconections with suitably planarized intermetal dielectrics. The transistors are at the very bottom, and are electrically isolated by shallow trench isolation (STI). (Photograph courtesy of Motorola.)

overcoat is deposited, openings are etched for the metal bond pads. After the chips are tested in an automated tester, the *known good dies* are packaged and wire bonded, as discussed in Section 9.6.

9.3.2 Silicon-on-Insulator (SOI)

An interesting and useful extension of the Si MOS process can be achieved by growing very thin films of single crystal Si on insulating substrates (Fig. 9–11). Two such substrates which have the appropriate thermal expansion match to Si are sapphire and spinel ($MgO–Al_2O_3$). Epitaxial Si films can be grown on these substrates by chemical vapor deposition (e.g., the pyrolysis of silane), with typical film thickness of about 1 μm. The film can be etched by standard photolithographic techniques into islands for each transistor. Implantation of p^+ and n^+ areas into these islands for source and drain regions result in the MOS devices. Since the film is so thin, the source and drain regions can be made to extend entirely through the film to the insulating substrate. As a result, the junction capacitance is reduced to the very small capacitance associated with the sidewalls between the source/drain and the channel region. In addition, since interconnections between devices pass over the insulating substrate, the usual interconnection–Si substrate capacitance is eliminated (along with the possibility of parasitic induced channels in the field between devices). These capacitance reductions improve considerably the high-frequency operation of circuits using such devices.

Other insulators can be used for SOI devices, including SiO_2. Since oxide can easily be grown on Si substrates, it serves as an attractive insulator for subsequent growth of thin-film Si. Since polycrystalline Si can be deposited directly over SiO_2, devices can be made in thin poly-Si films. However, to avoid grain boundaries and other defects typical of polycrystalline material, a variety of techniques have been developed to grow single crystal Si on oxide. For example, the oxide layer can be formed beneath the surface of a Si wafer by high-dose oxygen implantation. The thin Si layer remaining on the surface above the implanted oxide is usually about 0.1 μm thick, and can

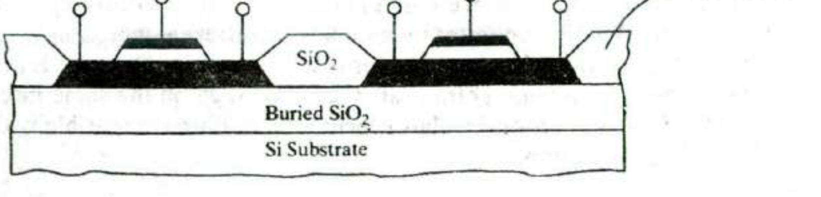


Figure 9–11 Silicon on insulator. Both n-channel and p-channel enhancement transistors are made in islands of Si film on the insulating substrate. These devices can be interconnected for CMOS applications.

be used as the thin film for CMOS or other device fabrication. This process is called *separation by implantation of oxygen (SIMOX)*. In some cases a thicker Si film is grown epitaxially on the SIMOX wafer, using the thin Si crystalline layer over the oxide as a seed for the epitaxy.

Another approach for making SOI is to place two oxidized Si wafers face-to-face, and thermally bond the oxide layers by high temperature annealing in a furnace. One of the wafers is then chemically etched back almost completely, leaving about a micron of single crystal Si material on top of the SiO_2. This approach is known as *Bond-and-Etch-back* SOI (BE-SOI). Since it is challenging to etch off about 600 μm of Si and stop controllably so as to leave about a micron of Si, an initial p^+ implanted layer is often used to act as an etch stop near the end of the chemical etch. The chemical recipe that is used for etching Si has a much lower etch rate in a p^+ layer than in lightly doped Si.

Another recent approach is a modification of BE-SOI using very high dose H implantation into one of the oxidized Si wafers, such that the peak of the H profile is about a micron below the Si surface. During the high temperature annealing step, the H atoms coalesce to form tiny H bubbles which cause a thin layer of Si to cleave off from the implanted wafer, leaving the Si layer bonded to the oxidized Si surface. This way, one does not have to chemically etch off one Si wafer for every BE-SOI wafer. If properly done, the cleaved-off Si wafer has a smooth surface and can be reused.

Examining the two devices of Fig. 9–11 more closely, one of them is unlike the transistors we have considered thus far. The thin Si film is lightly doped p-type, and therefore the device labeled "p-channel" appears *junctionless*. Such a device is able to operate in the enhancement mode (normally off) because of the equilibrium effects of the work function difference and the interface charge. With the usual Φ_{ms} and Q_i a depletion region is formed in the central p material of each device with zero gate voltage. In fact, for a Si film of about 0.1 μm or less, this depletion region can extend all the way through the Si to the insulator. Such a device is called *fully depleted* and no drain-to-source current flows. In the n-channel device a positive gate voltage greater than V_T induces an inverted region at the surface, as usual for an n-channel enhancement device. For the p-channel case a small negative voltage V_G removes the depletion and causes hole accumulation beneath the gate. The result is the formation of a conducting channel by a small negative gate voltage, as is the case for a conventional p-channel enhancement device. Although the fully depleted type of p-channel device operates by a somewhat different mechanism, its current–voltage characteristics are similar to the conventional device. Since both p-channel and n-channel transistors can be made on the same insulating surface, the silicon-on-insulator technique is quite compatible with CMOS circuit fabrication.

The SOI approach does not suffer from the latchup problem of bulk CMOS because there is no p-n-p-n thyristor from the power supply to ground. For circuits requiring high speeds, low standby power (due to the elimination of junction leakage to the substrate), and radiation tolerance (due to the elimination of the Si substrate), the extra expense of preparing sapphire substrates or growing crystalline Si films on oxide is compensated by increased performance.

9.3.3 Integration of Other Circuit Elements

One of the most revolutionary developments of integrated circuit technology is the fact that integrated transistors are cheaper to make than are more mundane elements such as resistors and capacitors. There are, however, numerous applications calling for diodes, resistors, capacitors and inductors in integrated form. In this section we discuss briefly how these circuit elements can be implemented on the chip. We will also discuss a very important circuit element—the interconnection pattern which ties all of the integrated devices together in a working system.

Diodes. It is simple to build p-n junction diodes in a monolithic circuit. It is also common practice to use transistors to perform diode functions. Since many transistors are included in a monolithic circuit, no special diffusion step is required to fabricate the diode element. There are a number of ways in which a transistor can be connected as a diode. Perhaps the most common method is to use the emitter junction as the diode, with the collector and base shorted. This configuration is essentially the narrow base diode structure, which has high switching speed with little charge storage (Prob. 5.35). Since all the transistors can be made simultaneously, the proper connections can be included in the metallization pattern to convert some of the transistors into diodes.

Resistors. Diffused or implanted resistors can be obtained in monolithic circuits by using the shallow junctions used in forming the transistor regions (Fig. 9–12a). For example, during the base implant, a resistor can be implanted which is made up of a thin p-type layer within one of the n-type islands. Alternatively a p region can be made during the base implant, and an n-type

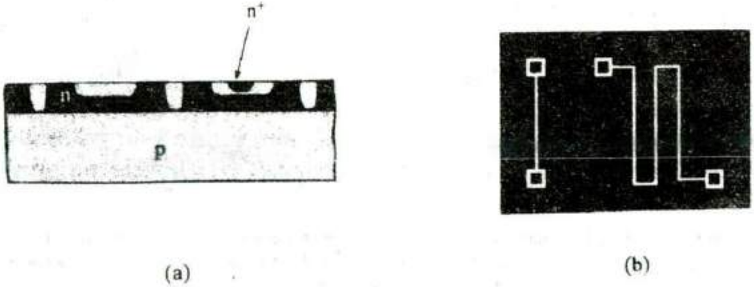


Figure 9–12 Monolithic resistors: (a) cross section showing use of base and emitter diffusions for resistors; (b) top view of two resistor patterns.

resistor channel can be included within the resulting p region during the emitter implant step. In either case, the resistance channel can be isolated from the rest of the circuit by proper biasing of the surrounding material. For example, if the resistor is a p-type channel obtained during the base implant, the surrounding n material can be connected to the most positive potential in the circuit to provide reverse-bias junction isolation. The resistance of the channel depends on its length, width, depth of the implant, and resistivity of the implanted material. Since the depth and resistivity are determined by the requirements of the base or emitter implant, the variable parameters are the length and width. Two typical resistor geometries are shown in Fig. 9–12b. In each case the resistor is long compared with its width, and a provision is made on each end for making contact to the metallization pattern.

Design of diffused resistors begins with a quantity called the *sheet resistance* of the diffused layer. If the average resistivity of a diffused region is ρ, the resistance of a given length L is $R = \rho L/w\mathrm{t}$, where w is the width and t is the thickness of the layer. Now if we consider one square of the material, such that $L = w$, we have the sheet resistance $R_s \equiv \rho/\mathrm{t}$ in units of ohms per square. We notice that R_s measured for a given layer is numerically the same for any size square. This quantity is simple to measure for a thin diffused layer by a four-point probe technique.[2] Therefore, for a given diffusion, the sheet resistance is generally known with good accuracy. The resistance then can be calculated from the known value of R_s and the ratio L/w (the *aspect ratio*) for the resistor. We can make the width w as small as possible within the requirements of heat dissipation and photolithographic limitations and then calculate the required length from w and R_s. Design criteria for diffused resistors include geometrical factors, such as the presence of high current density at the inside corner of a sharp turn. In some cases it is necessary to round corners slightly in a folded or zigzag resistor (Fig. 9–12b) to reduce this problem.

To reduce the amount of space used for resistors or to obtain larger resistance values, it is often necessary to obtain surface layers having larger sheet resistance than is available during the standard base or emitter implants. We can use a different implant, such as the V_T adjust implant to form shallow regions having very high sheet resistance ($\sim 10^5$ Ω/square). This procedure can provide a considerable saving of space on the chip. In integrated FET circuits it is common to replace load resistors with depletion-mode transistors, as mentioned in Section 6.5.5.

Capacitors. One of the most important elements of an integrated circuit is the capacitor. This is particularly true in the case of memory circuits, where charge is stored in a capacitor for each bit of information. Figure 9–13 illustrates a one-transistor DRAM cell, in which the n-channel MOS transistor provides

[2] This is a very useful method, in which current is introduced into a wafer at one probe, collected at another probe, and the voltage is measured by two probes in between. Special formulas are required to calculate resistivity or sheet resistance from these measurements.

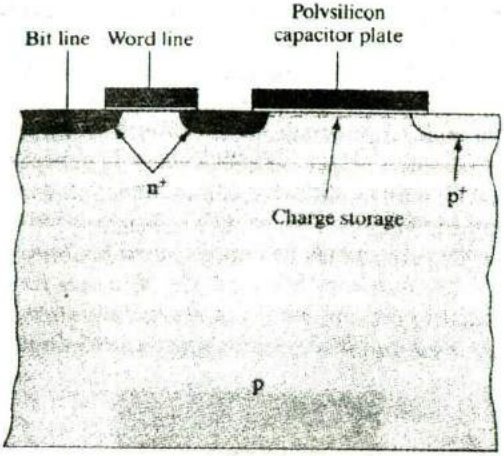


Figure 9–13 Integrated capacitor for DRAM cells. A one-transistor memory cell in which the transistor stores and accesses charge in an adjacent planar MOS capacitor.

access to the adjacent MOS capacitor. The top plate of the capacitor is polysilicon, and the bottom plate is an inversion charge contacted by an n^+ region of the transistor. The terms *bit line* and *word line* refer to the row and column organization of the memory (Section 9.5.2). One can also make use of the capacitance associated with p-n junctions, as discussed in Section 5.5.5.

Inductors. Inductors have not been incorporated into ICs in the past, because it is much harder to integrate inductors than the other circuit elements. Also, there has not been a great need for integrating inductors. Recently, that has changed because of the growing need for rf analog ICs for portable communication electronics. Inductors are very important for such applications, and can be made with reasonable Q factors using spiral wound thin metal films on an IC. Such spiral patterns can be defined by photolithography and etching techniques compatible with IC processing, or they can be incorporated in a hybrid IC.

Contacts and Interconnections. During the metallization step, the various regions of each circuit element are contacted and proper interconnection of the circuit elements is made. Aluminum is commonly used for the top metallization, since it adheres well to Si and to SiO_2 if the temperature is raised briefly to about 550°C after deposition. Gold is used on GaAs devices, but the adherence properties of Au to Si and SiO_2 are poor. Gold also creates deep traps in Si.

As mentioned throughout this chapter, silicide contacts and doped polysilicon conductors are commonly used in integrated circuits. By opening windows through the oxide layers to these conductors, Al metallization can be used to contact them and connect them to other parts of the circuit. In cases where Al is used to contact the Si surface, it is usually necessary to use Al containing about 1 percent Si to prevent the metal from incorporating Si from the layer being contacted, thereby causing "spikes" in the surface. Thin diffusion barriers are also used between the Al and Si layers, to prevent migration between the two. The refractory silicides mentioned in Section 9.3.1 serve this purpose.

Increased complexity and packing density in integrated circuits inevitably leads to a need for multilayer metallization. Multiple levels of Cu metallization can be incorporated with interspersing dielectrics. In general, the metals may all be Al, Cu or they may be different conductors such as polysilicon or refractory metals (depending on the heat each is subjected to in subsequent processing). Also, the dielectrics may be deposited oxides, boro-phospho-silicate glass for reflow planarization, nitrides, etc. The planarization of the surface is extremely important to prevent breaks in the metallization, which can occur in traversing a step on the surface. Various approaches using reflow glass, poly-

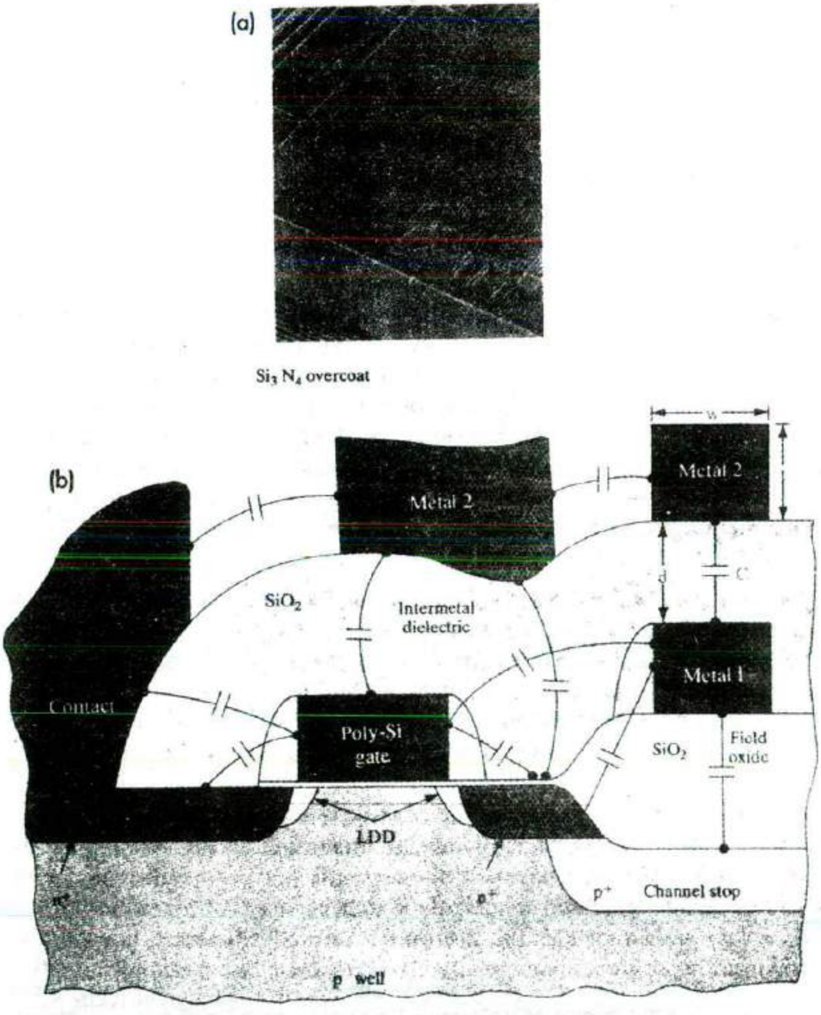


Figure 9–14 Multi-level interconnects: (a) Micrograph of multi-level interconnects in an IC. The inter-level dielectrics have been etched off to reveal the copper interconnect lines; (b) equivalent circuit illustrating the various parasitic capacitive elements associated with a multi-level interconnect. On the top right hand corner of the figure, we focus on the parallel plate capacitor model referred to in Eqs. (9–1) and (9–2). (Photograph courtesy of IBM.)

imide, and other materials to achieve planarization have been used, along with chemical mechanical polishing.

The most important challenge in designing interconnects is the *RC* time constant, which affects the speed and active power dissipation of the chip. A very simplistic model of two layers of interconnects with an inter-metal dielectric (Fig. 9–14b) shows that it can be regarded as a parallel plate capacitor. Regarding the interconnect as a rectangular resistor, its resistance is given by

$$R = \frac{\rho L}{tw} = R_s \frac{L}{w} \tag{9–1a}$$

where R_s is the sheet resistance, and the other symbols are defined in Fig. 9–14b. The capacitance is given by

$$C = \frac{\epsilon L w}{d} \tag{9–1b}$$

The RC time constant is then

$$\left(\frac{\rho L}{wt}\right)\left(\frac{\epsilon L w}{d}\right) = \frac{\rho \epsilon L^2}{td} = \frac{R_s \epsilon L^2}{d} \tag{9–2}$$

Interestingly, for this simple one-dimensional model, the width of the interconnect *w* cancels out. Therefore, it does not make sense to use wider conductors for high speed operation. It is also impractical to do so in terms of packing density. Of course, in reality we must account for the fringing electric fields, and therefore account for width dependence. From Eq. 9–2, it is clear we need as thick a metal layer (within practical limits of deposition times and etching times) and as low a resistivity as possible. Low resistivities are also important in minimizing ohmic voltage drops in metal bus lines that carry power from one end of a chip to the other. Aluminum is very good in this regard, and thus was a mainstay for Si technology for many years. Aluminum also has other nice attributes such as good ohmic contacts to both n and p-type Si, and good adhesion to oxides.

Copper has even lower resistivity (1.7 μΩ-cm) than Al (3 μΩ-cm) and is about two orders of magnitude less susceptible to electromigration. Hence, it is an excellent alternative to Al for very high speed ICs (Fig. 9–14a). The process breakthroughs that have made Cu viable for metallization include new electrodeposition and electroplating techniques because CVD or sputter deposition is not very practical for Cu. It is also very difficult to use reactive ion etching (RIE) for Cu because the etch byproducts for Cu are not very volatile. Hence, instead of RIE, Cu patterning is based on the so-called *Damascene* process where grooves are first etched in a dielectric layer, Cu is deposited on it, and the metal layer is chemically-mechanically polished down, leaving inlaid metal lines in the oxide grooves. In this method, the metal does not have to be etched directly using RIE, which is always a difficult process. The name "damascene" is derived from a metallurgical technique in ancient Turkey where metal artwork was inlaid into swords and other artifacts using this type of process. Copper can create traps deep in the bandgap of Si; hence, a suitable barrier layer such as Ti is needed between the Cu layer and the Si substrate.

Other parameters in Eq. (9–2) that can minimize the *RC* time constant are clearly the use of a thick inter-metal dielectric layer (once again within the limits of practicality in terms of deposition times), and as low a dielectric constant material as possible. Silicon dioxide has a relative dielectric constant of 3.9. There is active research in low dielectric constant materials (sometimes referred to as low-*K* materials). These include organic materials such as polyimides, or xerogels/aerogels which have air pockets or porosity purposely built in to minimize the dielectric constant.

In designing the layout of elements for a monolithic circuit, topological problems must be solved to provide efficient interconnection without *crossovers*—points at which one conductor crosses another conductor. If crossovers must be made on the Si surface, they can be accomplished easily at a resistor. Since the implanted or diffused resistor is covered by SiO_2, a conductor can be deposited crossing the insulated resistor. In cases requiring crossovers where no resistor is available, a low-value implanted resistor can be inserted in one of the conductor paths. For example, a short n^+ region can be implanted during the source/drain step and contacted at each end by one of the conductors. The other conductor can then cross over the oxide layer above the n^+ region. Usually, this can be accomplished without appreciable increase in resistance, since the n^+ region is heavily doped and its length can be made small.

During the metallization step, appropriate points in the circuit are connected to relatively large *pads* to provide for external contacts. These metal pads are visible in photographs of monolithic circuits as rectangular areas spaced around the periphery of the device. In the mounting and packaging process, these pads are contacted by small Au or Al wires or by special techniques such as those discussed in Section 9.6.

9.4 CHARGE TRANSFER DEVICES

One of the most interesting and broadly useful integrated devices is the *charge-coupled device (CCD)*. The CCD is part of a broader class of structures known generally as *charge transfer devices*. These are dynamic devices that move charge along a predetermined path under the control of clock pulses. These devices find applications in memories, various logic functions, signal processing, and imaging. In this section we lay the groundwork for understanding these devices, but their present forms and variety of applications must be found in the current literature.

9.4.1 Dynamic Effects in MOS Capacitors

The basis of the CCD is the dynamic storage and withdrawal of charge in a series of MOS capacitors. Thus we must begin by extending the MOS discussion of Chapter 6 to include the basics of dynamic effects. Figure 9–15

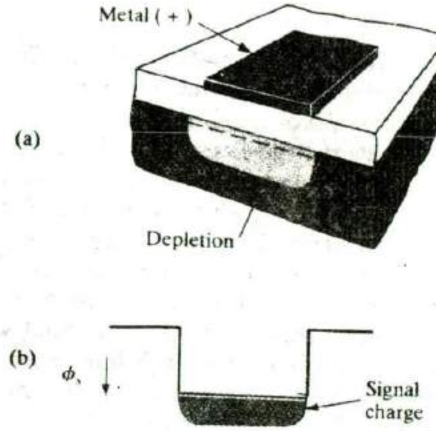


Figure 9–15 An MOS capacitor with a positive gate pulse: (a) depletion region and surface charge; (b) potential well at the interface, partially filled with electrons corresponding to the surface charge shown in (a).

shows an MOS capacitor on a p-type substrate with a large positive gate pulse applied. A depletion region exists under the gate, and the surface potential increases considerably under the gate electrode. In effect, the surface potential forms a *potential well*, which can be exploited for the storage of charge.

If the positive gate bias has been applied for a sufficiently long time, electrons accumulate at the surface and the steady state inversion condition is established. The source of these carriers is the thermal generation of electrons at or near the surface. In effect, the inversion charge tells us the capacity of the well for storage charge. The time required to fill the well thermally is called the *thermal relaxation time*, and it depends on the quality of the semiconductor material and its interface with the insulator. For good materials the thermal relaxation time can be much longer than the charge storage times involved in CCD operation.

If instead of a steady state bias we apply a large positive pulse to the MOS gate electrode, a deep potential well is first created. Before inversion has occurred by thermal generation, the depletion width is greater than it would be at equilibrium ($W > W_m$). This transient condition is sometimes called *deep depletion*. If we can inject electrons into this potential well electrically or optically, they will be stored there.[3] The storage is temporary, however, because we must move the electrons out to another storage location before thermal generation becomes appreciable.

What is needed is a simple method for allowing charge to flow from one potential well to an adjacent one quickly and without losing much charge in the process. If this is accomplished, we can inject, move, and collect packets of charge dynamically to do a variety of electronic functions.

[3] The potential well should not be confused with the depletion region, which extends into the bulk of the semiconductor. The "depth" of the well is measured in electrostatic potential, not distance. Electrons stored in the potential well are in fact located very near the semiconductor surface.

9.4.2 The Basic CCD

The original CCD structure proposed in 1969 by Boyle and Smith of Bell Laboratories consisted of a series of metal electrodes forming an array of MOS capacitors as shown in Fig. 9–16. Voltage pulses are supplied in three lines (L_1, L_2, L_3), each connected to every third electrode in the row (G_1, G_2, G_3). These voltages are clocked to provide potential wells, which vary with time as in Fig. 9–16. At time t_1 a potential well exists under each G_1 electrode, and we assume this well contains a packet of electrons from a previous operation. At time t_2 a potential is applied also to the adjacent electrode G_2, and the charge equalizes across the common G_1–G_2 well. It is easy to visualize this process by thinking of the mobile charge in analogy with a fluid, which flows to equalize its level in the expanding container. This fluid model continues at t_3 when V_1 is reduced, thus decreasing the potential well under G_1. Now the charge flows into the G_2 well, and this process is completed at t_4 when V_1 is zero. By this process the packet of charge has been moved from under G_1 to G_2. As the procedure is continued, the charge is next passed to the G_3 position, and continues down the line as time proceeds. In this way charge can be

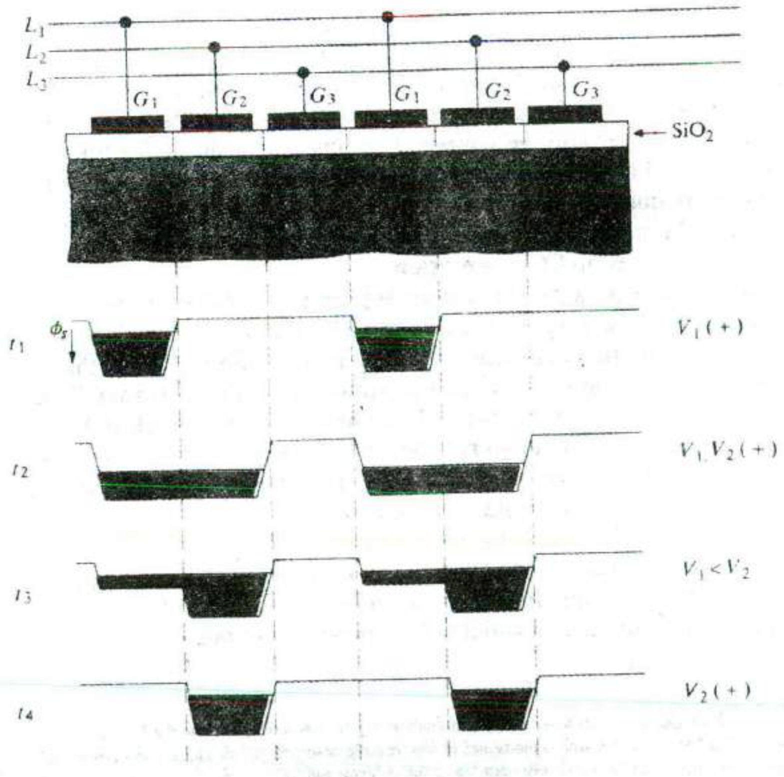


Figure 9–16 The basic CCD, composed of a linear array of MOS capacitors. At time t_1, the G_1 electrodes are positive, and the charge packet is stored in the G_1 potential well. At t_2 both G_1 and G_2 are positive, and the charge is distributed between the two wells. At t_3 the potential on G_1 is reduced, and the charge flows to the second well. At t_4 the transfer of charge to the G_2 well is completed.

injected using an input diode, transported down the line, and detected at the other end.

9.4.3 Improvements on the Basic Structure

Several problems arise in the implementation of the CCD structure of Fig. 9–16. For example, the separation between electrodes must be very small to allow coupling between the wells. An improvement can be made by using an overlapping gate structure such as that shown in Fig. 9–17. This can be done, for example, with poly-Si electrodes separated by SiO_2 or with alternating poly-Si and metal electrodes.

One of the problems inherent to the charge transfer process is that some charge is inevitably lost during the many transfers along the CCD. If the charges are stored at the Si–SiO_2 interface, surface states trap a certain amount of charge. Thus if the "0" logic condition is an empty well, the leading edge of a train of pulses is degraded by the loss of charge required in filling the traps which were empty in the "0" condition. One way of improving this situation is to provide enough bias in the "0" state to accommodate the interface and bulk traps. This procedure is colorfully referred to as using a *fat zero*. Even with the use of fat zeros, the signal is degraded after a number of transfers, by inherent inefficiencies in the transfer process.

Transfer efficiency can be improved by moving the charge transfer layer below the semiconductor–insulator interface. This can be accomplished by using ion implantation or epitaxial growth to create a layer of opposite type than the substrate. This shifts the maximum potential under each electrode into the semiconductor bulk, thus avoiding the semiconductor–insulator interface. This type of device is referred to as a *buried channel* CCD.

The three-phase CCD shown in Fig. 9–16 is only one example of a variety of CCD structures. Figure 9–18 illustrates one method for achieving a two-phase system, in which voltages are sequentially applied to alternating gate electrodes from two lines. A two-level poly-Si gate structure is used, in which the gate electrodes overlap, and a donor implant near the Si surface creates a built-in well under half of each electrode. When both gates are turned off (b), potential wells exist only under the implanted regions, and charge can be stored in any of these wells. With electrode G_2 pulsed positively, the charge

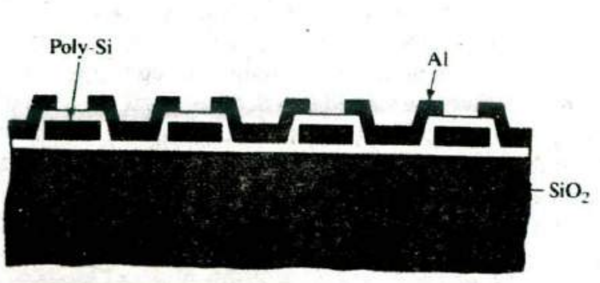


Figure 9–17
An overlapping gate CCD structure. One set of electrodes is polycrystalline Si, and the overlapping gates are Al in this case. SiO_2 separates the adjacent electrodes.

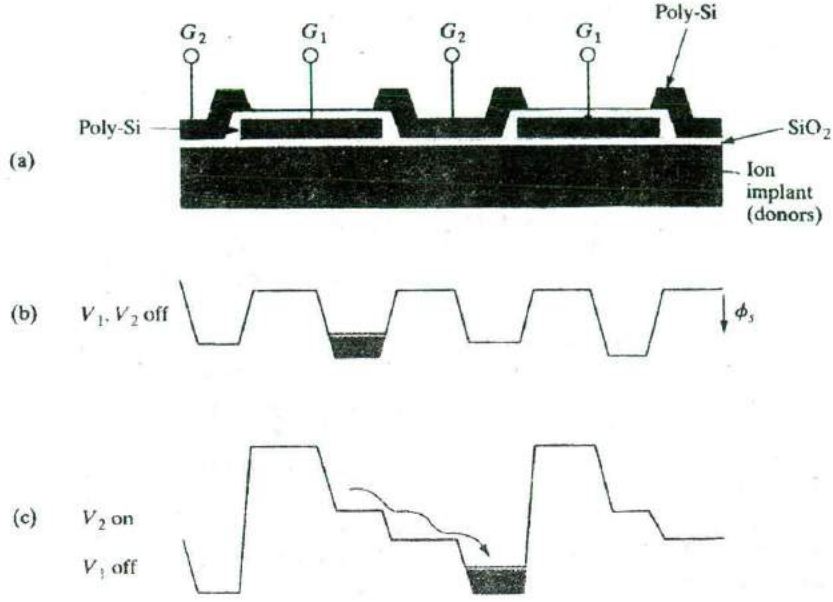


Figure 9–18 A two-phase CCD with an extra potential well built in under the right half of each electrode by donor implantation.

packet shown in (b) is transferred to the deepest well under G_2, which is its implanted region as (c) indicates. Then with both gates off, the wells appear as in (b) again, except that the charge is now under the G_2 electrode. The next step in the transfer process is obviously to pulse G_1 positively, so that the charge moves to the implanted region under the G_1 electrode to the right.

Other improvements to the basic structure are important in various applications. These include channel stops or other methods for achieving lateral confinement for the stored charge. Regeneration points must be included in the array to refresh the signal after it has been degraded.

9.4.4 Applications of CCDs

CCDs are used in a number of ways, including signal processing functions such as delay, filtering, and multiplexing several signals. Another interesting application of CCDs is in imaging for astronomy or in solid state TV cameras, in which an array of photosensors is used to form charge packets proportional to light intensity, and these packets are shifted to a detector point for readout. There are numerous ways of accomplishing this in CCDs, including the linear array line scanner, in which the second dimension is obtained by moving the scanner relative to the image. Alternatively, an area image sensor can be made which scans the image electronically in both dimensions. The latter device can be used as an alternative to the electron beam-addressed television imaging tube (Fig. 9–19).

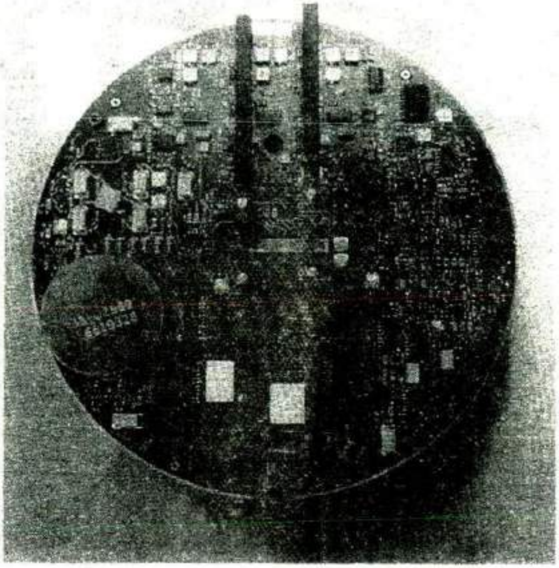


Figure 9–19 A charge-coupled device image sensor shown as large white rectangular areas, with peripheral signal processing circuitry. (Photograph courtesy of Texas Instruments.)

9.5 ULTRA LARGE-SCALE INTEGRATION (ULSI)

In the early development of integrated circuits it was felt that the inevitable defects that occur in processing would prevent the fabrication of devices containing more than a few dozen logic gates. One approach to integration on a larger scale tried in the late 1960s involved fabricating many identical logic gates on a wafer, testing them, and interconnecting the good ones (a process called *discretionary wiring*). While this approach was being developed, however, radical improvements were made in device processing which increased the yield of good chips on a wafer dramatically. By the early 1970s it was possible to build circuits with many hundreds of components per chip, with reasonable yield. These improvements made discretionary wiring obsolete almost as soon as it was developed. By reducing the number of processing defects, improving the packing density of components, and increasing the wafer size, it is now possible to place millions of device elements on a single chip of silicon and to obtain many perfect chips per wafer.

A major factor in the development of integrated circuits has been the continual reduction in size of the individual elements (transistors, capacitors) within each circuit. Through improved design and better lithography, there has been a dramatic shrinking of the minimum feature size (e.g., a transistor gate) used in these devices. The results of shrinking the elements in a 16-Mb-DRAM are shown in Fig. 9–20. By reducing the minimum feature size in successive steps from 0.43 to 0.3 μm, the die area was reduced from about 135 mm^2 in the first-generation design to less than 42 mm^2 in the fifth-generation device. Obviously, more of the smaller chips can be made by batch fabrication on the wafer, and the effort in shrinking the design is rewarded in a more profitable device.

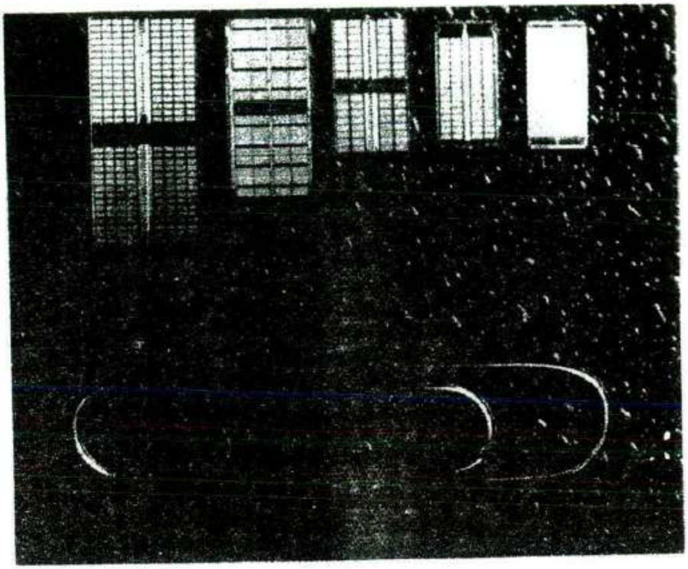


Figure 9–20 Size reduction of a 16 Mb DRAM die as the minimum feature size is reduced from 0.43 μm for the first generation design (die on the left) to 0.3 μm for the die on the right. (Photograph courtesy of Micron Technology, Inc.)

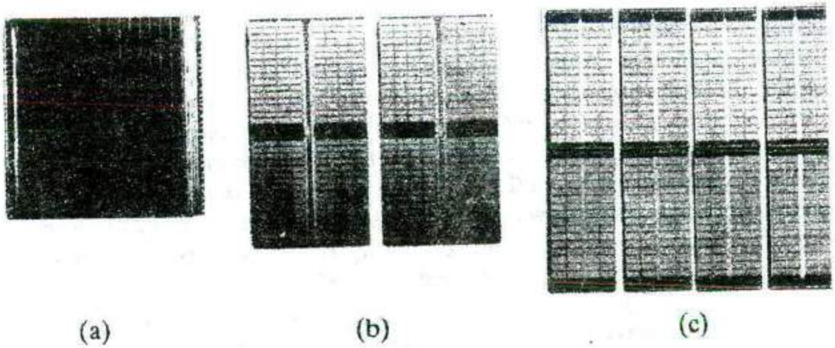


Figure 9–21 Three ways of obtaining 128 million bits of dynamic random access memory: (a) one 128-Mb chip; (b) two 64-Mb chips; (c) eight 16-Mb chips. The 128-Mb die measures almost 1 by 1 inch. (Photograph courtesy of Micron Technology, Inc.)

Successive designs using reduced feature sizes have made dramatically increased circuit complexity possible. DRAM design has set the pace over the past two decades, in which successive 1-Mb, 4-Mb, and 16-Mb memories led to similar powers of two increase to the 128-Mb range. Figure 9–21 illustrates the size comparison of a 128-Mb memory with an equivalent amount of memory in the form of two 64-Mb and eight 16-Mb chips. These are examples of ultra large-scale integration (ULSI).

Although the achievement of many powers of two in memory is impressive and important, other ULSI chips are important for the integration of many different system functions. A *microprocessor* includes functions for a computer central processing unit (CPU), along with memory, control, timing, and interface circuits required to perform very complex computing func-

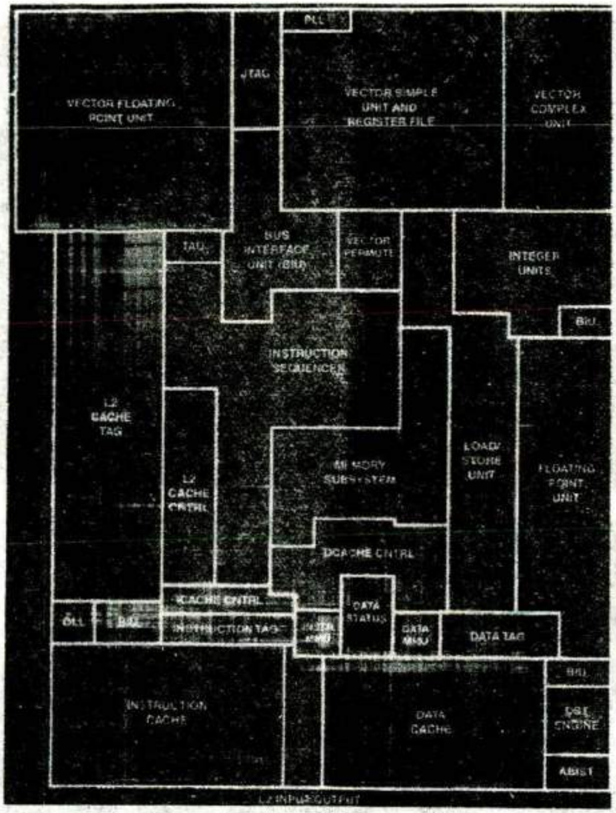


Figure 9–22
An example of ULSI, the Motorola Power PC microprocessor includes 10 million transistors employing CMOS technology with 0.35 μm minimum feature size. (Photograph courtesy of Motorola, Inc.)

tions. The complexity of such devices is shown in Fig. 9–22, which illustrates a microprocessor chip with various areas outlined by function.

Before leaving this section it might be useful to provide some calibration regarding the dimensions we have been discussing. Figure 9–23 compares the size of 64Mb DRAM circuit interconnect elements with a human hair, on the same scale. We can see that the densely packed 0.18 μm lines on this ULSI memory chip are dwarfed by the scanning electron micrograph of a human hair which has the diameter of about 50 microns. This makes it dramatically clear why ULSI chips must be fabricated in ultra-clean environments.

Although the focus of this book is devices and not circuits, it is important to look at some typical applications of MOS capacitors and FETs in semiconductor logic and memory ULSI, which constitute about 90 percent of all ICs. This should give the reader a better feel for why we have studied the physics of MOS devices in Chapter 6. This is clearly not a comprehensive discussion, because the design and analysis of circuits is a large subject covered in other books and courses. We will first look at some digital logic applications, followed by some typical memory devices.

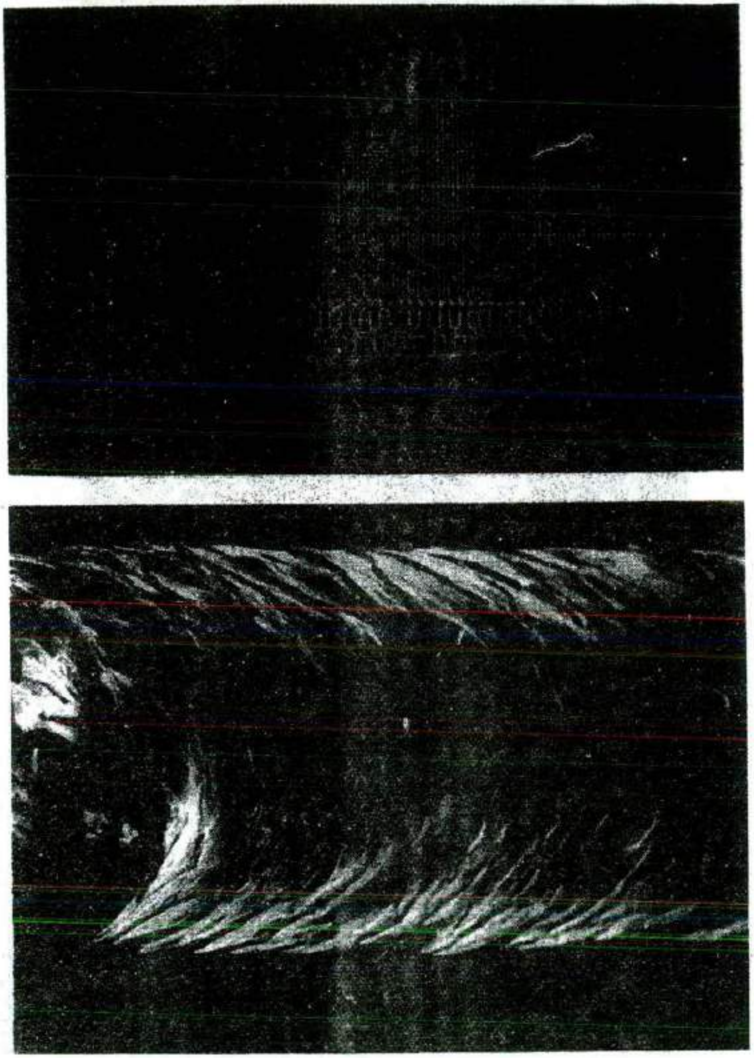
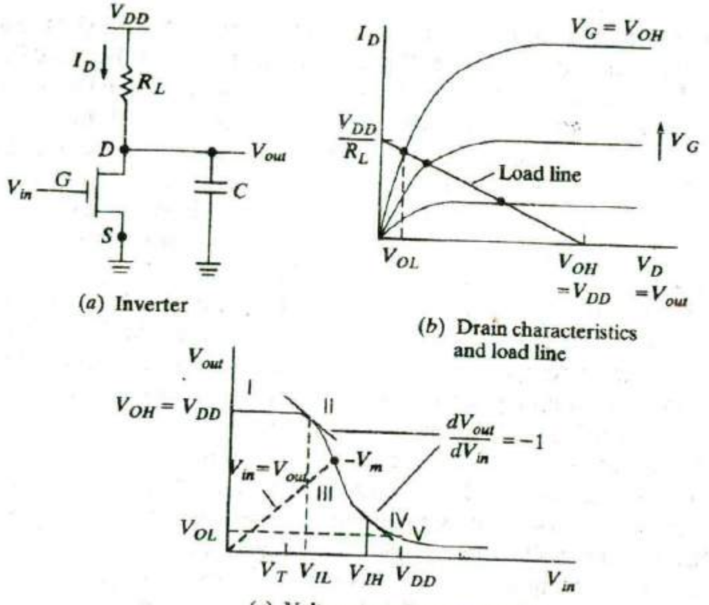


Figure 9–23 Size comparison of ULSI circuit elements with human hair: (a) densely-packed 0.18 μm interconnect lines in the array of a 64Mb DRAM chip; (b) scanning electron micrograph of a human hair. Both micrograqphs are shown to the same scale. (Photograph coutesy of Micron Technology, Inc.)

9.5.1 Logic Devices

A very simple and basic circuit element is the inverter, which serves to flip the logic state. When its input voltage is high (corresponding to logic "1"), its output voltage is low (logic "0"), and vice versa. Let us start the analysis with a resistor–loaded n-channel MOSFET inverter to illustrate the principles in the simplest possible manner (Fig. 9–24a). Then, we will extend the treatment to the slightly more complicated CMOS inverters which are much more useful and more common today.



(*a*) Inverter

(*b*) Drain characteristics and load line

(*c*) Voltage transfer characteristic

Figure 9–24 Resistor load inverter voltage transfer characteristics (VTC): (a) NMOSFET with load resistor, R_L, and load parasitic capacitance, C; (b) determination of VTC by superimposing load line (linear I–V ohmic characteristics of resistor) on NMOSFET output characteristics; (c) VTC showing output voltage as a function of input voltage. The five key points on the VTC are logic high (V_{OH}), logic low (V_{OL}), unity gain points (V_{IL} and V_{IH}), and logic threshold where input equals output (V_m).

A key concept for inverters is the *voltage transfer characteristic* (VTC), which is a plot of the output voltage as a function of the input bias (Fig. 9–24c). The VTC gives us information, for example, about how much noise the digital circuit can handle, and the speed of switching of the logic gates. There are five key operating points (marked I through V) on the VTC. They include V_{OH}, corresponding to the logic high or "1", V_{OL}, corresponding to the logic low or "0", and V_m corresponding to the intersection of a line with unity slope (where $V_{out} = V_{in}$) with the VTC. V_m, known as the logic threshold (not to be confused with the V_T of the MOSFETs), is important when two inverters are cross-coupled in a flip-flop circuit because the output of one is fed to the input of the other, and vice versa. Two other key points are the unity gain points, V_{IL} and V_{IH}. The significance of these points is that if the input voltage is between them, the change of the input is amplified and we get a larger change of the output voltage. Outside of this operating range, the change of the input voltage is attenuated. Clearly, any noise voltage which puts the input voltage between V_{IL} and V_{IH} would be amplified, and lead to a potential problem with the circuit operation.

Let us see how to go about determining the VTC. From the circuit in Figure 9–24a, we see that in the output loop from the power supply to ground, the current through the resistor load is the same as the drain current of the MOSFET. The power supply voltage is equal to the voltage drop across the resistor plus the drain-to-source voltage. To determine the VTC, we superimpose the load line of the load element (in this case a straight line for an ohmic resistor) on the output characteristics of the MOSFET (Figure 9–24b). This is similar to our load line discussion in Section 6.1.1. The load line goes through V_{DD} on the voltage axis because when the current in the output loop is zero, there is no voltage drop across the resistor and all the voltage appears across the MOSFET. On the current axis, the load line goes through V_{DD}/R_L because when the voltage across the MOSFET is zero, the voltage across the resistor must be V_{DD}. As we change the input bias, V_{in}, we change the gate bias on the MOSFET, and thus in Figure 9–24b, we go from one constant V_G curve to the next. At each input bias (and a corresponding constant V_G curve) the intersection of the load line with that curve tells us what the drain bias V_D is, which is the same as the output voltage. This is because at the point of intersection, we satisfy the condition that for the d-c case where the capacitor does not play any role, the current through the resistor is the same as the MOSFET current. (Later on, we shall see that in the a-c case when the logic gates are switched, we need to worry about the displacement current through the capacitor when it is charged or discharged.) It can be clearly seen from Fig.9–24c that as the input voltage (or V_G) changes from low to high, the output voltage decreases from a high of V_{DD} to a low of V_{OL}. We can solve for any point on this VTC curve analytically simply by recognizing whether the MOSFET is in the linear region or in saturation, using the corresponding drain current expression [Eq. (6–49) or (6–53)] and setting it equal to the resistor current. As an illustration, suppose we want to determine the logic "0" level, V_{OL}. This occurs when the input V_G is high and the output V_D is low, putting the transistor in the linear region. Using equation (6–49), we can write

$$I_D = k\left[V_G - V_T - \frac{V_D}{2}\right]V_D = k\left[V_{DD} - V_T - \frac{V_{OL}}{2}\right]V_{OL} \qquad (9\text{–}3a)$$

Since in the d-c case the current through the MOSFET is the same as that through the resistor,

$$I_D = I_L = \frac{V_{DD} - V_{OL}}{R_L} \qquad (9\text{–}3b)$$

We can solve for V_{OL} if we know R_L and the MOSFET parameters. Alternatively, we can design the value of R_L to achieve a certain V_{OL}. What might dictate the choice of R_L? We shall see later in this section that for many applications we use two of these inverters in a cross-coupled manner to form a bistable flip-flop. The output of one flip-flop is fed back to the input of the other, and vice versa. Clearly, the V_{OL} must be designed to be significantly less than the V_T of the MOSFET. Otherwise, neither MOSFET will be fully turned off, and the flip-flop will not function properly. Similarly, all the other points on the VTC can be determined analytically by using the appropriate MOSFET drain current expression, and setting it equal to the current through the resistor.

We can make some general observations from this analysis. We want the transition region of the VTC (between V_{IL} and V_{IH}) to be as steep (i.e., high gain) as possible, and the transition should be around $V_{DD}/2$. High gain guarantees a high-speed transition from one logic state to the other. It is necessary to increase the load resistance to increase this gain in the transition region.

The transition around $V_{DD}/2$ guarantees high *noise immunity* or *margin* for both logic "1" and logic "0" levels. To appreciate the importance of noise immunity, we must recognize that in combinatorial or sequential digital circuits, the output of one inverter or logic gate is often fed into the input of the next stage. Noise immunity is a measure of how much noise voltage the circuit can tolerate at the input, and still have the digital outputs be at the correct logic level in the subsequent stages. For example in Fig. 9–24c, if the input is nominally at zero, the output should be high (logic "1"). If this is fed into another inverter stage, its output should be low, and so on. If a noise spike causes the input of the first stage to go above V_m, the output voltage decreases sufficiently to potentially create errors in the digital levels in subsequent stages. Having a symmetric transition of the VTC around $V_{DD}/2$ ensures that the noise margin is high for both logic levels.

One problem with the resistor load inverter is that the V_{OL} is low, but not zero. This, coupled with the fact that the load element is a passive resistor that cannot be turned off, causes high standby power dissipation in this circuit. These problems are addressed by the CMOS structure described next.

We can determine the VTC for the CMOS case exactly as for the resistor load, although the math is somewhat more messy (Fig. 9–25). As mentioned previously, for an input voltage V_{in}, the V_G of the NMOSFET is V_{in}, but that of the PMOSFET is $V_{\text{in}}-V_{DD}$. Similarly, if the output voltage is V_{out}, the V_D of the NMOSFET is V_{out}, but that of the PMOS is $V_{\text{out}}-V_{DD}$. The load element now is not a simple resistor with a linear current–voltage relationship, but instead is the PMOSFET device whose "load line" is a set of I_D–V_D output characteristics (Fig. 9–25b). The V_{out} can be determined as a function of the V_{in} by recognizing whether the NMOSFET and the PMOSFET are in the linear or saturation region of their characteristics, and using the appropriate current expressions. At each point, we would set the NMOSFET I_D equal to the PMOSFET I_D.

As in the case of the resistive load, there are five key points on the VTC (Fig. 9–25c). They are logic "1" equal to V_{DD}, logic "0" equal to 0, logic threshold V_m where $V_{\text{in}} = V_{\text{out}}$, and the two unity gain points, V_{IH} and V_{IL}. In region I in Fig. 9–25c, the NMOSFET is OFF, and $V_{\text{out}} = V_{DD}$. Similarly, in region V, the PMOSFET is OFF, and $V_{\text{out}} = 0$. We can illustrate the calculation in region II, where the NMOSFET is in saturation and the PMOSFET is in the linear region. In this case, we must use Eq. (6–53) for the saturation drain current of the NMOSFET.

$$I_{DN} = \frac{k_N}{2}(V_{in} - V_{TN})^2 \qquad (9\text{–}4a)$$

On the other hand, we must use Eq. (6–49) for the PMOSFET in the linear region.

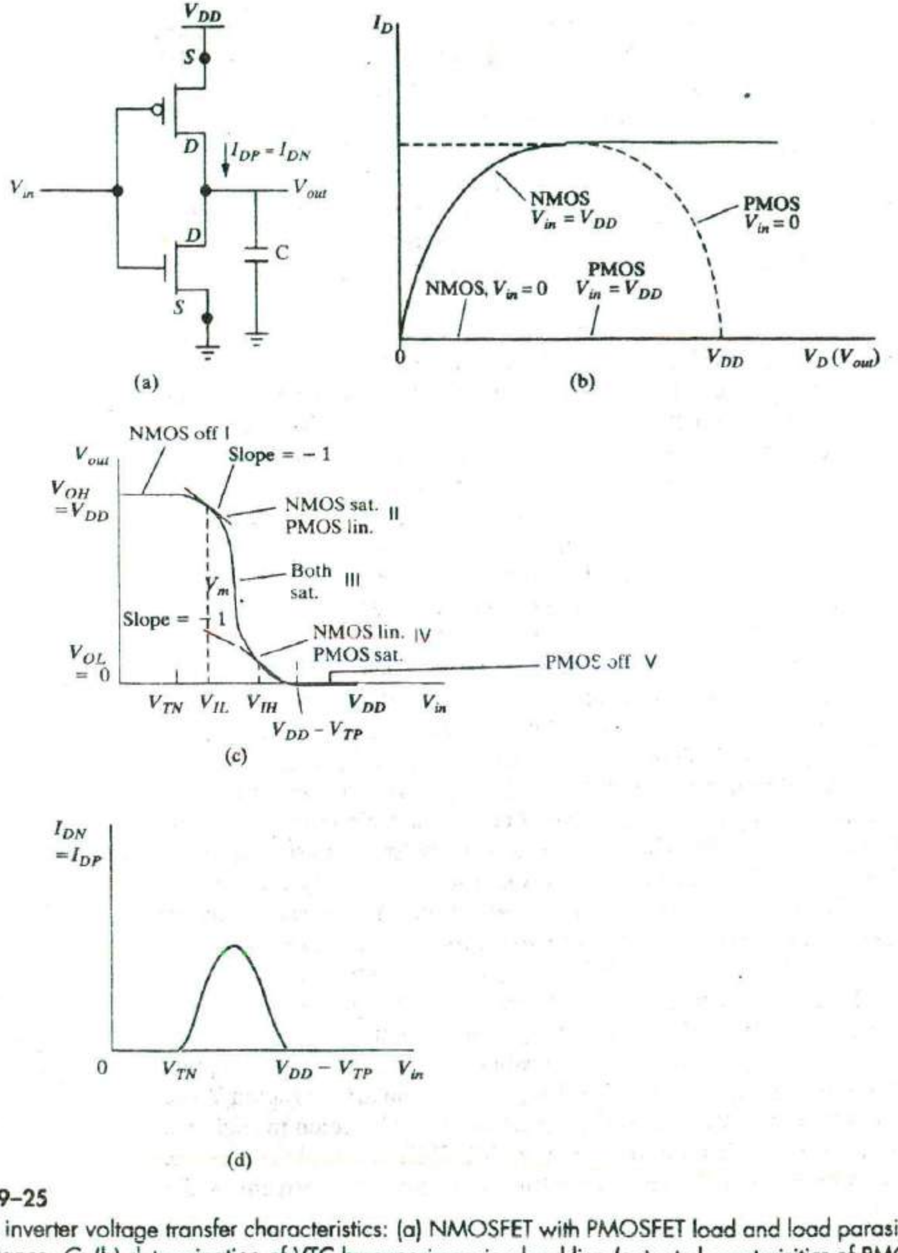


Figure 9–25

CMOS inverter voltage transfer characteristics: (a) NMOSFET with PMOSFET load and load parasitic capacitance, *C*; (b) determination of VTC by superimposing load line (output characterisitics of PMOSFET shown as dotted line) on NMOSFET output characteristics; (c) VTC showing output voltage as a function of input voltage. The five key points on the VTC are logic high (V_{OH}), logic low (V_{OL}), unity gain points (V_{IL} and V_{IH}), and logic threshold where input equals output (V_m); (d) switching current from V_{DD} to ground when the input voltage is in a range where both the NMOSFET and the PMOSFET are on.

$$I_{DP} = k_P\left[(V_{DD} - V_{in}) + V_{TP} - \frac{(V_{DD} - V_{out})}{2}\right](V_{DD} - V_{out}) \qquad (9\text{–}4b)$$

Here V_{TN} and V_{TP} are the n- and p-channel threshold voltages. In the d-c case, since the output load capacitor does not play a role, the drain current through the PMOSFET device must be equal in magnitude to that through the NMOSFET. (However, for the a-c case, we need to consider the displacement current through the capacitor.)

$$I_{DN} = I_{DP} \qquad (9\text{–}5a)$$

Using Eq. (6–53) for the NMOSFET in saturation, and Eq. (6–49) for the PMOSFET in the linear region.

$$\frac{k_N}{2k_p}(V_{in} - V_{TN})^2 = \left[V_{DD} - V_{in} + V_{TP} - \frac{V_{DD} - V_{out}}{2}\right](V_{DD} - V_{out})$$

$$= \left[\frac{V_{DD}}{2} - V_{in} + V_{TP} + \frac{V_{out}}{2}\right](V_{DD} - V_{out}) \qquad (9\text{–}5b)$$

From Eq. (9–5b), we can get an analytical relation between the input and output voltages valid in Region II. We can get similar relationships in the other regions of the VTC.

Region IV is very similar to region II in Fig. 9–25c, except that now the NMOS is in the linear regime, while the PMOSFET is in saturation. In region III, both the NMOSFET and the PMOSFET are in saturation. Since the output impedance of a MOSFET is very high, this is tantamount to a semi-infinite load resistor, thereby resulting in a very steep transition region. That is why a CMOS inverter switches faster than the resistor load case. The CMOS inverter is also preferable because in either logic state (regions I or V), either the NMOSFET or the PMOSFET is OFF, and the standby power dissipation is very low. In fact, the current in either logic state corresponds to the (very low) source/drain diode leakage.

We want the transition region (region III) to be at $V_{DD}/2$ from the point of view of symmetry and noise immunity. Once again, by setting the NMOSFET I_D equal to that of the PMOSFET, it can be shown that the transition occurs at

$$V_{in} = (V_{DD} + \chi V_{TN} + V_{TP})/(1 + \chi) \qquad (9\text{–}6a)$$

where

$$\chi = \left(\frac{k_N}{k_p}\right)^{\frac{1}{2}} = \frac{\left[\bar{\mu}_n C_i \left(\frac{Z}{L}\right)_N\right]^{\frac{1}{2}}}{\left[\bar{\mu}_p C_i \left(\frac{Z}{L}\right)_P\right]^{\frac{1}{2}}} \qquad (9\text{–}6b)$$

We can design V_{in} to be at $V_{DD}/2$ by choosing $V_{TN} = -V_{TP}$ and $\chi = 1$. Since the effective electron mobility in the channel of a Si MOSFET is roughly twice that of the hole mobility, we must design CMOS circuits to have a $(Z/L)_P = 2(Z/L)_N$ to achieve the condition $\chi = 1$.

We can combine such CMOS inverters to form other logic gates for combinatorial circuits such as NOR gates and NAND gates (Fig. 9–26). The truth tables for these gates are shown in Fig. 9–27. By applying combinations of logic "high" or logic "low" to inputs A and B, we get the output states corresponding to the truth tables. The synthesis of logic circuits corresponding to these truth tables can be done using Boolean algebra and De Morgan's laws. The upshot of these laws is that any logic circuit can be made using inverters in conjunction with either NAND gates or NOR gates. Which would be preferable from a device physics point of view? We see from Fig. 9–26, that in the NOR gate the PMOSFET devices T_3 and T_4 are in series, while for the

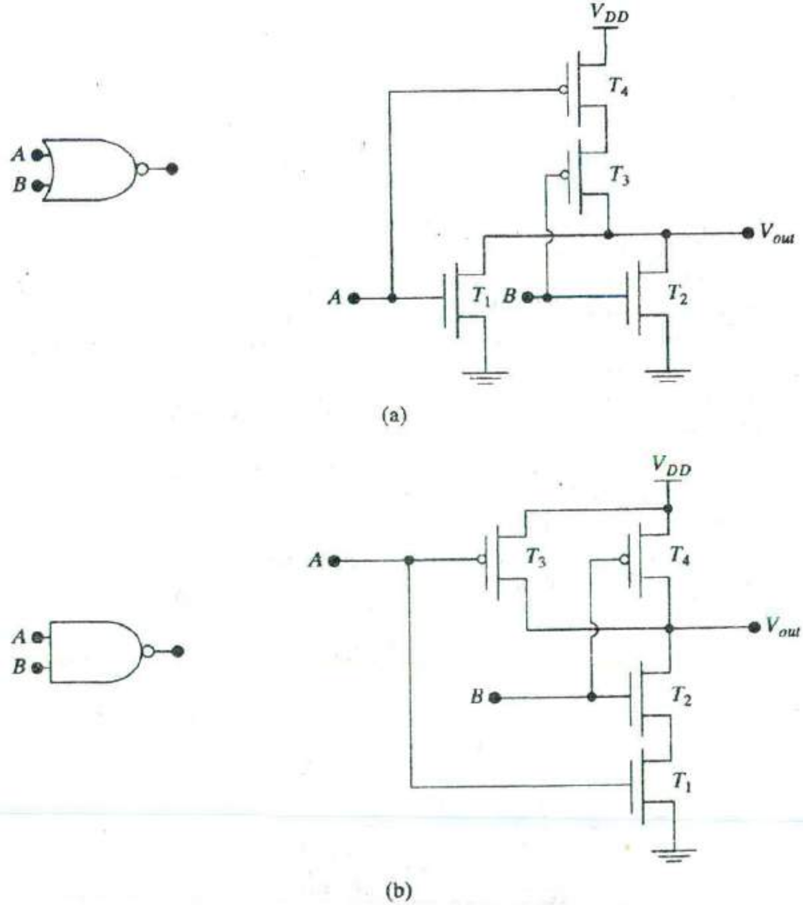
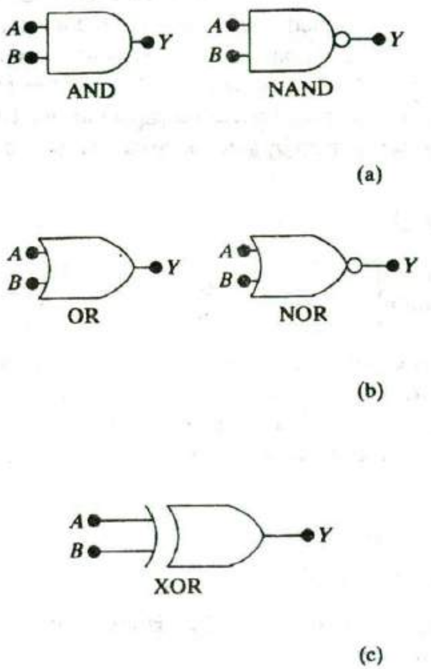


Figure 9–26 Logic gates and CMOS implementation of (a) NOR gate (b) NAND gate.



Input		Output Y	
A	B	AND	NAND
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Input		Output Y	
A	B	OR	NOR
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

Input		Output Y
A	B	XOR
0	0	0
0	1	1
1	0	1
1	1	0

Figure 9–27 (a) AND/NAND logic symbols and truth table. (b) OR/NOR logic symbols and truth table. (c) XOR logic symbol and truth table.

NAND it is the NMOSFETs (T_1 and T_2). Since the electron channel mobilities are twice hole mobilities, we would obviously prefer NMOSFETs. Therefore, the preferred choice is NAND, along with inverters.

We can also estimate the power dissipation in the inverter circuit. We already know that the standby power dissipation is very small, being governed by the OFF state leakage current of either the NMOSFET or the PMOSFET, depending on the logic state. This leakage current depends on the source and drain diode leakage currents, or if the V_T is low, on the subthreshold leakage of the MOSFET that is turned OFF (see Section 6.5.7).

While the inverter is switching, there is also a transient current from the power supply to ground when both the transistors are ON (see Fig. 9–25d). This is known as the *switching current* or the *commutator current*. The magnitude of this current will clearly depend on the values of V_{TN} and V_{TP}. The higher the magnitudes of the thresholds, the less is the input voltage swing for which both the PMOSFET and the NMOSFET will be ON while the input voltage is being changed. The commutator current is then less during switching, which is desirable from a reduced power dissipation point of view. However, this reduction of power dissipation by increasing threshold voltages is obtained at the expense of reduced drive current and, therefore, overall speed of the circuit.

The speed penalty due to reduction of drive current is because in a digital circuit, while switching between logic states, the MOSFET drive currents must charge and discharge the parasitic capacitors that are inevitably associated with the output node (Fig. 9–25a). There is also some power dissipation involved in charging and discharging load capacitors attached to the output of the inverter. This load capacitance depends mostly on the input gate oxide capacitance of the MOSFETs of the next inverter stage (or logic gate) that this inverter (or logic gate) may be driving, along with some small parasitic capacitances. The input load capacitance of a single inverter is given by gate oxide capacitance per unit area C_i times the device areas.

$$C_{\text{inv}} = C_i\{(ZL)_N + (ZL)_P\} \tag{9–7}$$

The total load capacitance is then multiplied by a factor that depends on the *fan-out* of the circuit, which is the number of gates that are being driven in parallel by the inverter (or logic gate). It is necessary to add up the load capacitances for all the inverters or logic gates that are being driven by this inverter stage. The energy expended in charging up the equivalent load capacitor, C, is the integral of the product of the time-dependent voltage times the time-dependent displacement current through the capacitor during the charging cycle.

$$\begin{aligned} E_C &= \int i_p(t)[V_{DD} - v(t)]dt \\ &= V_{DD}\int i_p(t)dt - \int i_p(t)v(t)dt \end{aligned} \tag{9–8a}$$

The energy stored in C is then obtained by considering the displacement current $(i_p(t) = C\,dv/dt)$ through the capacitor:

$$E_c = V_{DD}\int C\frac{dv}{dt}dt - \int Cv\frac{dv}{dt}dt = CV_{DD}\int_0^{V_{DD}} dv - C\int_0^{V_{DD}} v\,dv = CV_{DD}^2 - \frac{1}{2}CV_{DD}^2 \tag{9–8b}$$

Similarly, during one discharging cycle we get

$$E_d = \int i_n(t)v(t)dt = -\int_{V_{DD}}^{0} Cv\,dv = \frac{1}{2}CV_{DD}^2 \tag{9–9}$$

If the inverter (or gate) is being charged and discharged at a frequency f, we get an active power disspation

$$P = CV_{DD}^2 f \tag{9–10}$$

In addition to power dissipation, we are also concerned with the speed of logic circuits. The speed of a gate, such as the one shown in Fig. 9–25, is determined by the *propagation delay* time t_P. We define the time required for the output to go from the logic high V_{OH} to $V_{OH}/2$ as t_{PHL}. The converse (to go from logic low $V_{OL}(=0)$ to $V_{OH}/2$) is defined as t_{PLH}. We can write down approximate estimates for these times by recognizing that for the output to go from high to low (or logic "1" to "0"), the NMOSFET has to discharge the output node towards ground. During this period, the NMOSFET will be in saturation. Assuming a constant saturation current as an approximation, we obtain from Eq. (6–53)

$$t_{PHL} = \frac{\frac{1}{2}CV_{DD}}{I_{DN}} = \frac{\frac{1}{2}CV_{DD}}{\frac{k_N}{2}(V_{DD} - V_{TN})^2} \qquad (9\text{–}11a)$$

This is the decrease of charge on the capacitor divided by the discharging current. Conversely,

$$t_{PLH} = \frac{\frac{1}{2}CV_{DD}}{I_{DP}} = \frac{\frac{1}{2}CV_{DD}}{\frac{k_P}{2}(V_{DD} + V_{TP})^2} \qquad (9\text{–}11b)$$

Knowing these times helps us considerably in designing circuits that meet the speed requirements of a design. Of course, for accurate numerical estimates of these propagation time delays or of the power dissipation we need to use computers. A very popular program to do so is the Simulation Program with Integrated Circuit Emphasis (SPICE). This discussion illustrates that the device physics plays an important role in the design and analysis of such circuits.

9.5.2 Semiconductor Memories

In addition to logic devices such as microprocessors, integrated circuits depend on semiconductor memories. We can illustrate many key MOS device physics issues by looking at three of the most important types of semiconductor memory cells: the *static random access memory (SRAM)*, the *dynamic random access memory (DRAM)*, and the non-volatile *flash memory cell*. SRAMs and DRAMs are volatile in the sense that the information is lost if the power supply is removed. For flash memories, however, information is stored indefinitely. For SRAMs, the information is static, meaning that as long as the power supply is on, the information is retained. On the other hand, the information stored in the cells of a DRAM must periodically be refreshed because stored charge representing one of the logic states leaks away rapidly. The refresh time must be short compared with the time needed for stored charge to degrade.

The overall organization of all these types of memories is rather similar, and is shown in Fig. 9–28. We will not describe the memory organization in great detail here, but will instead focus on the device physics. We need to know the type of cell that is used at the intersection of the rows or wordlines, and the columns or bitlines. These memories are all random access in the sense that the cells can be addressed for write or read operations in any order, depending on the row and column addresses provided to the address pins, unlike memories such as hard disks or floppy disks on a computer which can only be addressed sequentially. Generally, the same set of pins is used for both the row and the column addresses, in order to save pin count. This forces us to use what is known as address multiplexing. First, the row addresses are provided at the address pin, and decoded using row decoders. For N row addresses, we can have 2^N rows or wordlines. The row decoders then cause the

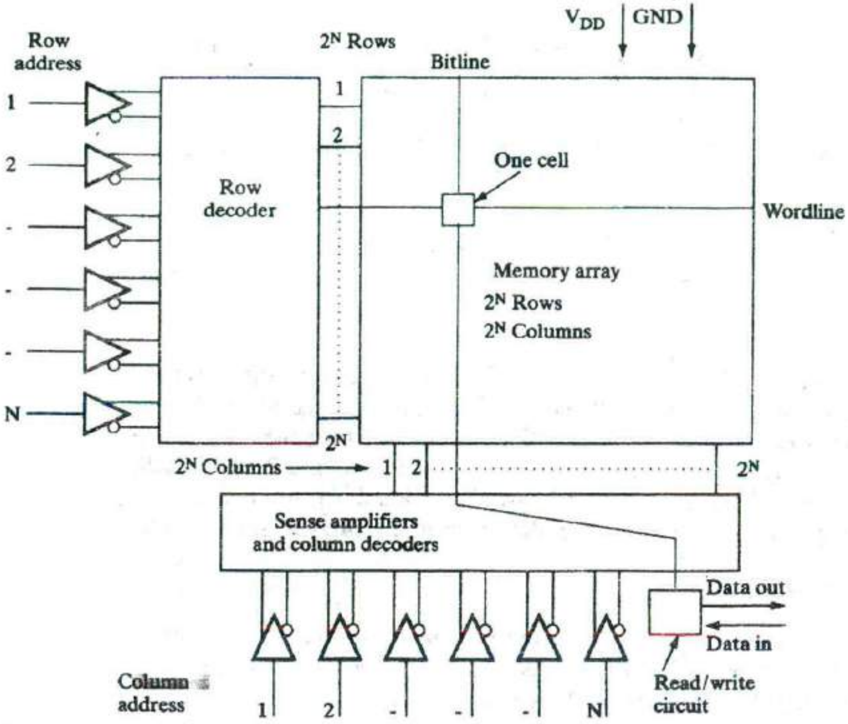


Figure 9–28

Organization of a random access memory (RAM): The memory array consists of memory cells arranged in an orthogonal array. There is one memory cell at the intersection of one row (wordline) and one column (bitline). To address a particular memory cell, the N row addresses are latched in from the N address pins, and decoded by the 2^N row decoders. All the memory cells on the selected row are read by the 2^N sense amplifiers. Of those, a cell (one bit) or group of cells (byte or word) is selected for transfer to the data output buffers depending on the column addresses that are decoded by the 2^N column decoders. Generally, to save pin count on the package, the N column addresses are provided in a multiplexed fashion to the same N address pins as the row addresses, *after* the row addresses have already been latched in.

selected wordline to go high, so that all the 2^N cells (corresponding to N column addresses) on this wordline are accessed for either read or write, through sense amplifiers at the end of the 2^N columns or bitlines. After the appropriate row has been decoded, the appropriate column addresses are provided to the same address pins, and the column decoders are used to select the bit or group of bits (known as *byte* or *word*) out of all the 2^N bits on the selected wordline. We can either write into or read from the selected bit (or group of bits) using the sense amplifiers, which are basically flip-flops used as differential amplifiers.

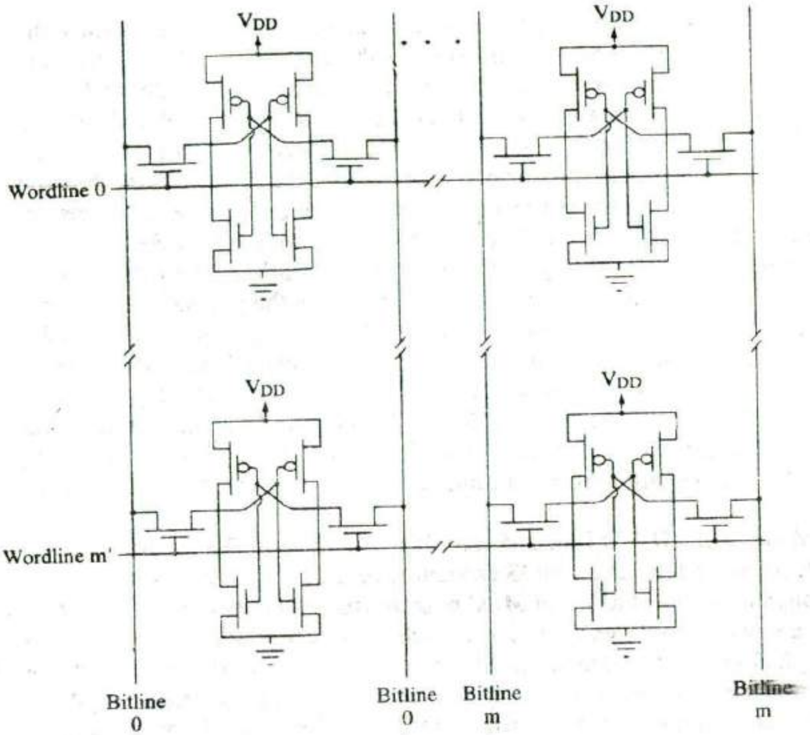


Figure 9–29 An array of 4 CMOS SRAM cells. The bitline and $\overline{\text{bitline}}$ (bitline-bar) are logical complements of each other.

SRAMs. A group of four six-transistor CMOS SRAM cells is shown in Fig. 9–29. Each cell is found in this case at the intersection of a row or wordline, and a column or bitline (along with its logical complement known as bitline-bar). The cell is a flip-flop, consisting of two cross-coupled CMOS inverters. Clearly, it is bistable: if the output of one inverter is high (corresponding to the NMOSFET being OFF, and the PMOSFET ON), that high voltage is fed to the input of the other cross-coupled inverter, and the output of the other inverter will be low. This is one logic state (say "1") of the SRAM. Conversely, the other stable state of the flip-flop can be considered to be the other logic state (say "0"). Many of the device issues are identical to those described in Section 9.5.1 in connection with the VTC of inverters. We aim for a symmetric transition from V_{OH} to V_{OL} at $V_{DD}/2$ with a high gain in the transition region, to improve noise immunity and speed of convergence of the SRAM cell. The speed of convergence determines how fast the SRAM flip-flop latches into one stable logic state or the other. The cells are accessed through two access transistors whose gates are controlled by the wordline. That is why this is called a 6-transistor cell. Other SRAM cells use load resistors in the inverters, rather than PMOSFETs, leading to a 4-transistor, 2-resistor cell. As discussed in Section 9.5.1, the CMOS cell has superior performance, but at the expense of occupying more area.

Unless the row decoders cause a particular wordline to go high, the SRAM cells on that wordline are electrically isolated. By selecting a particular wordline, the access transistors on that row are turned ON and act as logic transmission gates between the output nodes of the SRAM cell and the bitline and its complement, the bitline-bar. During a read operation, the bitline and its complement are both precharged to the same voltage. Once the access transistors are turned ON, a small voltage differential develops between bitline and bitline-bar because the output nodes of the SRAM are at different voltages (0 and V_{DD}). The voltage differential that is established is due to a charge redistribution that occurs between the parasitic capacitance associated with the output nodes of the SRAM and the bitline capacitance. This voltage difference is amplified by the sense amplifiers. As mentioned previously, the sense amplifiers are differential amplifiers, very similar in configuration to the SRAM flip-flop cell itself. The bitline and bitline-bar (complement of the bitline) are fed to the two inputs of the sense amplifier, and the voltage differential is amplified until the voltage separation is V_{DD}.

DRAMs. The DRAM cell structure is shown in Fig. 9–30. The information is stored as charge on an MOS capacitor, which is connected to the bitline through a switch which is an MOS *pass transistor*, the gate of which is controlled by the wordline. There is one such cell at each intersection of the orthogonal array of wordlines and bitlines, exactly as for SRAMs. When the wordline voltage becomes higher than the V_T of the pass transistor (MOSFET between the bitline and the storage capacitor), the channel is turned ON, and connects the bitline to the MOS storage capacitor. The gate of this capacitor (or capacitor plate) is permanently connected to the power supply voltage V_{DD}, thereby creating a potential well under it which tends to be full of inversion electrons for a p-type substrate (Fig. 9–31a). We apply either 0V to the bitline (generally corresponding to logic "0"), or V_{DD} (corresponding

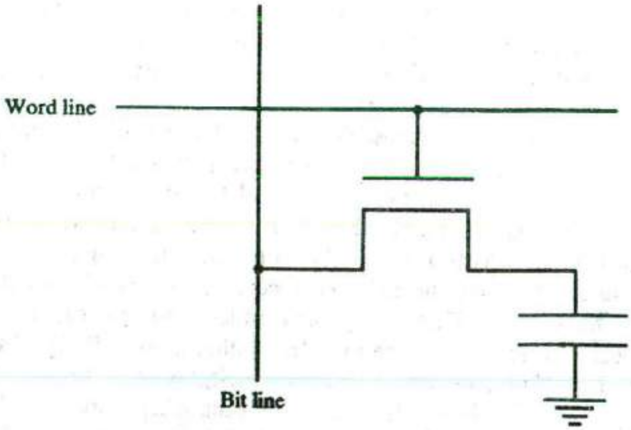


Figure 9–30 One transistor, one capacitor DRAM cell equivalent circuit: the storage MOS capacitor is connected to the bitline through the pass transistor (MOSFET switch) whose gate is controlled by the wordline.

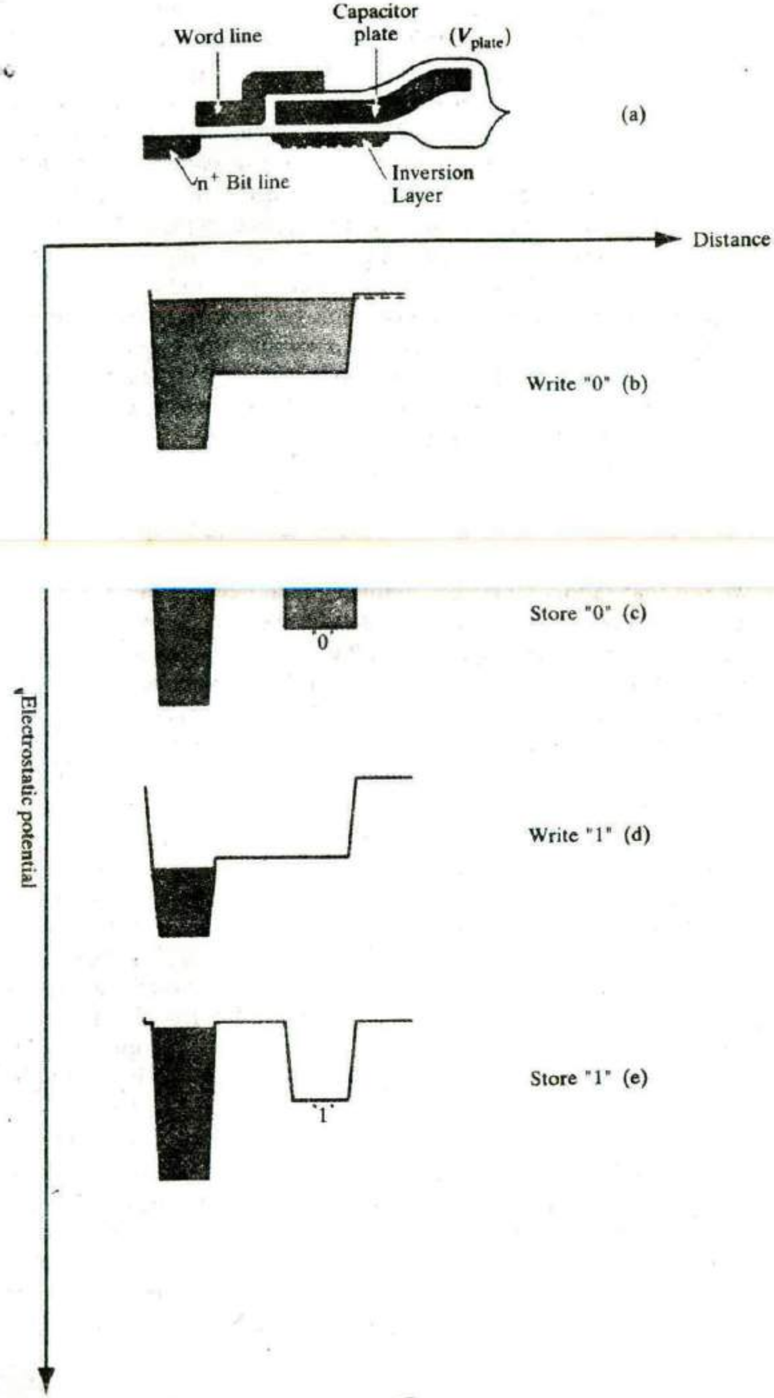


Figure 9–31 DRAM cell structure and cell operation: (a) cell structure corresponding to equivalent circuit of Fig. 9–30; (b)–(e) potentials under bitline, pass transistor channel and storage capacitor during write "0", store "0", write "1" and store "1" operations. It shows that the logic state "0" corresponds a filled potential while the logic state "1" corresponds to an empty potential well (unstable state) that is filled up over time by minority carriers generated in the substrate and leakage through the pass transistor.

to logic "1"), and the appropriate voltage appears as the substrate potential of the MOS capacitor. For a stored "0" in the cell, the potential appears as the substrate potential of the MOS capacitor. For a stored "0" in the cell, the potential well that is created under the MOS capacitor by the plate voltage is full of inversion charge (Fig. 9–31b,c). When the wordline voltage is turned low such that the MOS pass transistor is turned off, the inversion charge under the storage capacitor stays the same; this is the stable state of the capacitor. On the other hand, if a positive voltage (V_{DD}) is applied to the bitline, it draws out the inversion electrons through the pass transistor (Fig. 9–31d,e). When the pass transistor is cut off. we end up with an empty potential well under the MOS capacitor plate. Over a period of time, the potential well tends to be filled up by minority carrier electrons that are constantly created by thermal generation-recombination in the substrate and are collected under the charged MOS capacitor plate. Hence, the logic "1" degrades towards the logic "0". That is why a DRAM is considered to be "dynamic" unlike an SRAM. It is necessary to periodically restore the logic levels or "refresh" the stored information.

There are interesting device physics issues regarding the pass transistor. This is like the access transistor in the SRAM, or a logic transmission gate. We see that in this MOSFET, neither the source nor the drain is permanently grounded. In fact, which side acts as the source and which as the drain depends on the circuit operation. When we are writing a logic "1" into the cell, the bitline voltage is held high ($=V_{DD}$). As this voltage is written into the cell, it is as if the source of the pass transistor gets charged up to V_{DD}. Another way of looking at this is that with respect to the source, the substrate bias of the pass transistor is $-V_{DD}$. The body effect of the MOSFET (Section 6.5.6) causes its V_T to increase. This is very important because for the pass transistor to operate as a transmission gate it is necessary that it be in the linear regime throughout, and not get into saturation (with a concomitant voltage drop across the pinch-off region). Hence, the gate or the wordline voltage must be held at V_{DD} (which is the final voltage of the source/drains) *plus* the V_T of the MOSFET, taking body effect into account. It is also important to make sure that the leakage of the pass transistor is low enough to satisfy refresh requirements of the DRAM. Not only must the source/drain diodes be low leakage, but the V_T and the subthreshold slope must be optimized such that subthreshold leakage for the grounded wordline case is low enough.

The stored charge difference between the two logic states can be determined by looking at the capacitance–voltage (C–V) characteristics of the MOS capacitor (Fig. 9–32). For a stored "1", essentially there is a substrate bias applied to the MOS capacitor, which raises its V_T due to the body effect (Section 6.5.6). Hence, the C–V characteristics shift to the right for a stored "1". Since the MOS capacitance is not a fixed capacitance, but is voltage dependent, we saw earlier that it must be defined in a differential form (Eq. 6-34a). Alternatively, we can write down the stored charge under the capacitor as

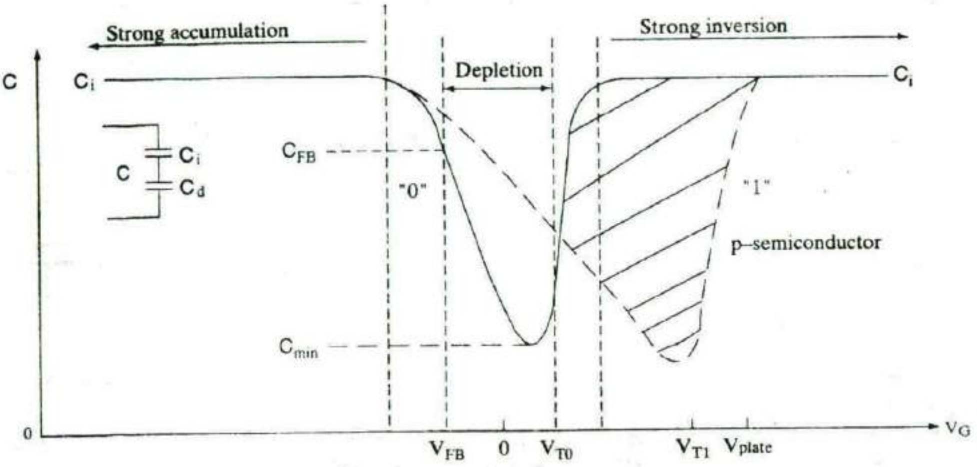


Figure 9–32
C–V characteristics of DRAM MOS capacitor in stored "0" and stored "1" states. The difference of area under the C–V curves shown by hatch-marked pattern reflects the charge differential between the two

$$Q = \int C(V)dV \tag{9–12}$$

This is simply the area under the C–V curve. The charge differential that distinguishes the logic "1" and the logic "0" is the difference of areas under the capacitance–voltage curves in the two cases (Fig. 9–32).

When reading the cell, the pass transistor is turned on, and the MOS storage capacitor charge is dumped on the bitline capacitance C_B, precharged to V_B (typically = V_{DD}). The swing of the bitline voltage will clearly depend on the voltage V_C stored in the storage cell capacitance C_C. As in the case of the SRAM, the change of the bitline voltage depends on the capacitance ratio between the bitline and the cell. To do differential sensing in the case of DRAMs, we do not use two bitlines per cell as for SRAMs. Instead, we compare the bitline voltage for the selected cell with a reference bitline voltage to which is connected a dummy cell whose MOS capacitance, C_D, is roughly half that of the actual cell capacitance, C_C. Typical values of C_B, C_C and C_D in a DRAM are 800 fF, 50 fF and 20 fF, respectively. The voltage differential that is applied to the sense amplifier then becomes (Fig. 9–33)

$$\begin{aligned}\Delta V &= \frac{C_C V_C + C_B V_B}{C_B + C_C} - \frac{C_C V_D + C_B V_B}{C_B + C_D}\\ &= \frac{(V_B - V_D)C_B C_D - (V_B - V_C)C_B C_D - (V_C - V_D)C_C C_D}{(C_B + C_C)(C_B + C_D)}\end{aligned} \tag{9–13a}$$

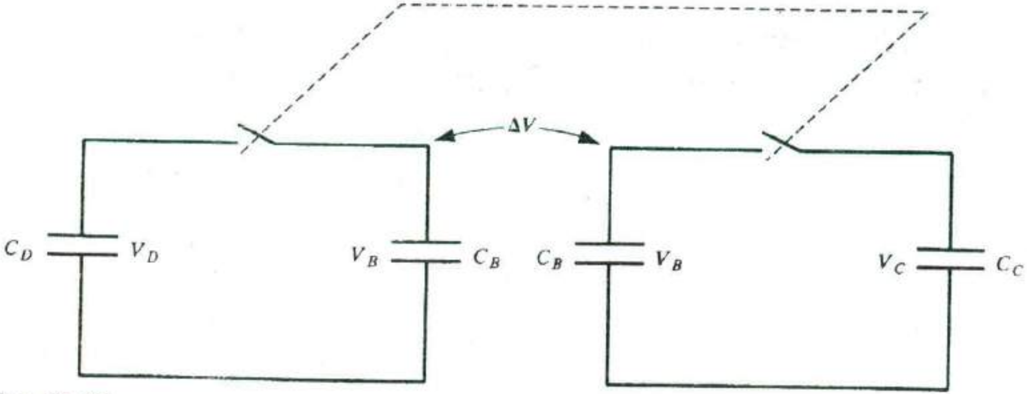


Figure 9 33
Equivalent circuit showing charge redistribution between cell capacitance (C_C) and bit line capacitance (C_B) on one side, versus dummy cell (C_D) and bit line capacitance (C_B) on other.

If V_D is set to zero, the expression simplifies to:

$$\Delta V = \frac{(V_B C_B + V_C C_C) C_D \quad - V_C) C_B C_C}{(C_B + C_C)(C_B + C_D)} \tag{9–13b}$$

Putting the cell voltage V_C equal to 0V or 5 V, and typical, acceptable bitline-to-cell capacitance ratios C_B/C_C (= 15–20) in Eq. (9–13b), we get different polarities of the differential voltage of the order of ±100 mV for logic "1" and logic "0", respectively, which can be detected by sense amplifiers. From Eq. (9–13b), it can be seen that for much higher bitline-to-cell capacitance ratios, the swing of the bitline voltage will be negligible, regardless of the cell voltage. The minimum required cell capacitance C_C is about 50 fF, governed by so-called soft errors. DRAMs, like everything else on Earth, are constantly being bombarded by cosmic rays, and high energy alpha particles can create electron–hole pairs in semiconductors. A typical collected charge due to one of these events is about 100 fC. This spurious charge can be neglected if the cell capacitance is 50 fF and 5 V is applied to the cell, for which the stored charge is roughly 250 fC. The DRAM cell then becomes immune to typical alpha particle hits.

Maintaining a cell capacitance of 50 fF as the cell dimensions are reduced from one generation of DRAM to the next is a tremendous technological challenge. One way to look at this problem is shown in Fig. 9–34. The challenge is to store more charge per unit area on the planar surface (A_s) of the Si substrate. Approximating the MOS capacitance as a fixed, voltage-independent capacitor, one can write the stored charge Q as

$$Q = CV = (\epsilon A_C/d)V \tag{9–14}$$

Time	Past	Present	Future
Approaches	Scaled dielectric	Trench/ stacked capacitor	Alternate dielectric
Problems	Tunneling & wearout	Fabrication	Material properties

$$\frac{Q}{A_s} = \frac{CV}{A_s} = \frac{V}{d} \times \frac{A_c}{A_s} \times \epsilon$$

Figure 9–34
Various approaches (past, present and future) of achieving higher DRAM cell capacitance and charge storage density without increasing cell size. A_s = Area on wafer taken by capacitor; A_c = Area of capacitor. For a planar capacitor $A_c = A_s$; however, for non-planar structures $A_c > A_s$. $C = A_c \epsilon/d$ is the capacitance; and $Q = CV$ is the total stored charge in a fixed, voltage-independent capacitor.

where ϵ is the permittivity of the dielectric, d is its thickness, and A_C is the capacitor area. As shown in Fig. 9–34, the historical way of achieving the desired capacitance has been to scale the dielectric thickness, d. But that runs into the problems discussed in Section 6.4.7. Another approach, which is being taken currently, is to use fabrication schemes to increase the area devoted to the MOS storage capacitor, A_c, even as we reduce the planar surface area on the wafer, A_s, used for making this storage capacitance. Obviously, this can be done by moving away from a purely planar structure, and exploiting the third dimension. We can go down into the Si by digging "trenches" in the substrate with RIE and forming a trench storage capacitor on the sidewalls of the trench (Figure 9–35a). Alternatively, we can go up from the substrate by stacking multiple layers of capacitor electrodes to increase the "stacked" capacitor area (Fig. 9–35b). Other tricks that have been tried are to purposely create a rough polysilicon surface on the capacitor

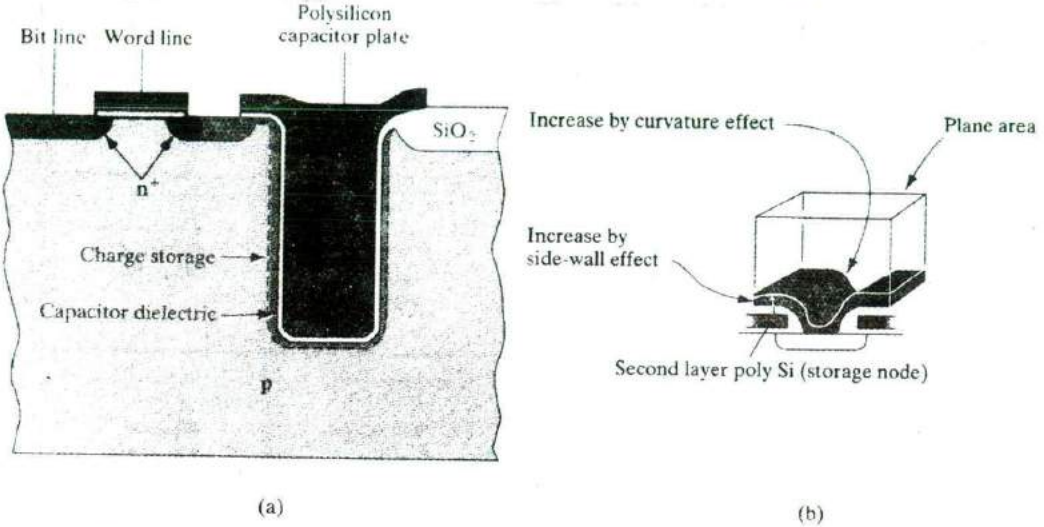

Bit line
Word line
Polysilicon capacitor plate
SiO$_2$
n$^+$
Charge storage
Capacitor dielectric
p
(a)
Increase by curvature effect
Plane area
Increase by side-wall effect
Second layer poly Si (storage node)
(b)

Figure 9–35
Increasing cell capacitance by exploiting the vertical dimension: (a) trench capacitors involve etching a trench in the substrate so that the larger area on the sidewalls can be used to increase capacitance; (b) stacked capacitors go "up" rather than "down" as in trenches, and increase capacitor area by using multiple polysilicon capacitor plates or "fins", as well as by exploiting the topgraphy of the cell surface.

plates to increase the surface area. In the future, alternative materials may be used. For example, the ferroelectrics have much higher dielectric constant than SiO_2 and offer larger capacitance without increasing area or reducing thickness. Promising materials include barium strontium titanate and zirconium oxide.

Flash Memories. Another interesting MOS device is the flash memory, which is rapidly becoming the most important type of non-volatile memory. The memory cell structure is shown in Figure 9–36. It is very simple and compact, and looks just like a MOSFET, except that it has two gate electrodes, one on top of the other. The top electrode is the one that we have direct electrical access to, and is known as the control gate. Below that we have a so-called "floating" gate that is capacitively coupled to the control gate and the underlying silicon.

The capacitive coupling of the floating gate to the various terminals is illustrated in Fig. 9–36 in terms of the various coupling capacitance components. The floating gate and the control gate are separated by a stacked oxide–nitride–oxide dielectric in typical flash devices. The capacitance between these two gates is called C_{ONO} because of the oxide–nitride–oxide makeup of the dielectric stack. The total capacitance C_{TOT} is the sum of all the parallel components shown in Fig. 9–36.

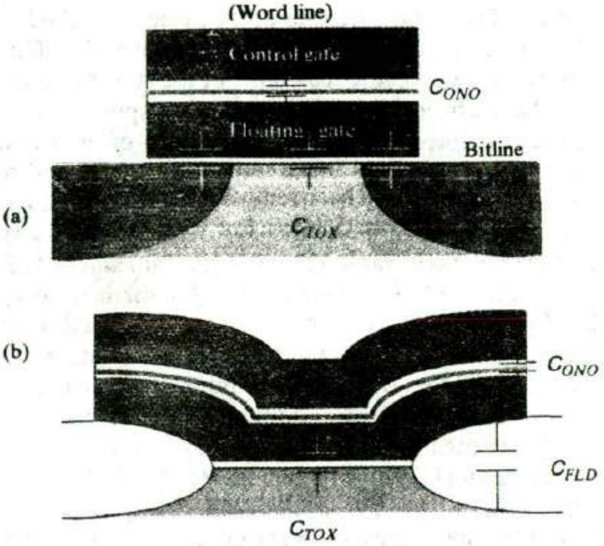


Figure 9–36 Flash memory cell structure: (a) cell structure shown along the channel length showing the control gate (wordline), floating gate below it, the source and the drain (bitline); (b) view of cell along the width of the MOSFET. The various coupling capacitors to the floating gate are shown.

$$C_{TOT} = C_{ONO} + C_{TOX} + C_{FLD} + C_{SRC} + C_{DRN} \tag{9–15}$$

where C_{TOX} is the floating gate-to-channel capacitance through the tunnel oxide, C_{FLD} is the floating gate-to-substrate capacitance in the LOCOS field oxide region, and C_{SRC} and C_{DRN} are the gate-to-source/drain overlap capacitances.

Since it is isolated by the surrounding dielectrics, the charge on the floating gate Q_{FG} is not changed by (moderate) changes of the terminal biases.

$$Q_{FG} = 0 = C_{ONO}(V_{FG} - V_G) + C_{SRC}(V_{FG} - V_S) + C_{DRN}(V_{FG} - V_D) \tag{9–16}$$

We assume that the substrate bias is fixed, and hence ignore the contributions from C_{TOX} and C_{FLD}, which couple the floating gate to the substrate. The floating gate voltage can be indirectly determined by the various terminal voltages, in terms of the gate, drain and source coupling ratios as defined in Eq. (9–17).

$$V_{FG} = V_G \cdot GCR + V_S \cdot SCR + V_D \cdot DCR \tag{9–17}$$

where

$$GCR = \frac{C_{ONO}}{C_{TOT}}$$

$$DCR = \frac{C_{DRN}}{C_{TOT}}$$

$$SCR = \frac{C_{SRC}}{C_{TOT}}$$

The basic cell operation involves putting charge on the floating gate or removing it, in order to program the MOSFET to have two different V_T's, corresponding to two logic levels. We can think of the stored charge on the floating gate to be like the fixed oxide charge in the V_T expression (equation 6-38). If many electrons are stored in the floating gate, the V_T of an NMOSFET is high; the cell is considered to have been "programmed" to exhibit the logic state "1". On the contrary, if electrons have been removed from the floating gate, the cell is considered to have been "erased" into a low V_T state or logic "0".

How do we go about transferring charges into and out of the floating gate? To program the cell, we can use channel hot carrier effects that we discussed in Section 6.5.9. We apply a high field to both the drain (bitline) and floating gate (wordline) such that the MOSFET is in saturation. It was discussed in Section 6.5.9 that the high longitudinal electric field in the pinch-off region accelerates electrons toward the drain and makes them energetic (hot). We maximize such hot carrier effects near the drain pinch-off region in a flash device by making the drain junction somewhat shallower than the source junction (Fig. 9–37a). This can be achieved by a separate higher energy source implant that is masked in the drain region. If the kinetic energy of electrons is high enough, a few can become hot enough to be scattered into the floating gate. They must surmount the 3.1 eV energy barrier that exists between the conduction band of Si and that of SiO_2, or hot electrons can tunnel through the oxide (Fig. 9–37b). Once they get into the floating gate, electrons become trapped in the 3.1 eV potential well between the floating polysilicon gate and the oxides on either side. This barrier is extremely high

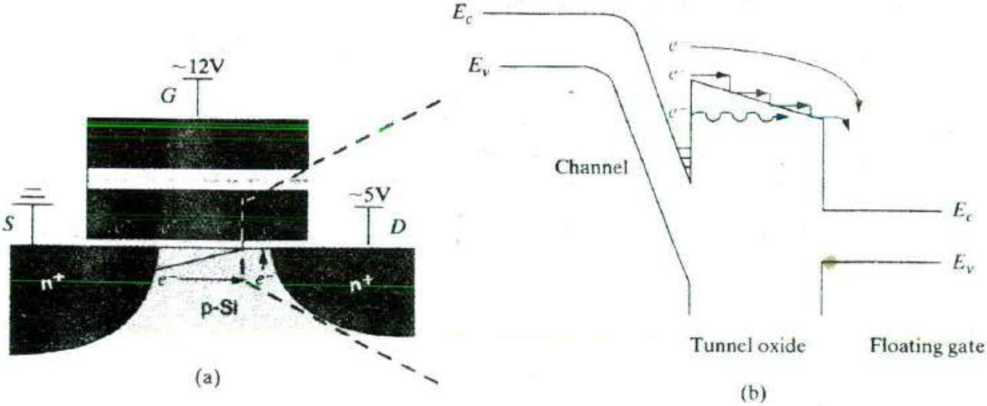


Figure 9–37
Hot carrier programming of the flash cell: (a) flash memory cell structure with typical biases required for writing into the cell. The channel of the MOSFET is pinched off in saturation; (b) band diagram along a vertical line in the middle of MOSFET channel showing hot electrons in the channel being injected across the gate oxide and getting trapped in the floating gate.

for a trapped (low kinetic energy) electron. Therefore the trapped electrons essentially stay in the floating gate forever, unless the cells is intentionally erased. That is why a flash memory is non-volatile.

To erase the cell, we use Fowler–Nordheim tunneling between the floating gate and the source in the overlap region (Fig. 9–38a). A high positive voltage (say ~12 V) is applied to the source with the control gate grounded. The polarity of the field is such that electrons tunnel from the floating gate into the source region, through the oxide barrier (Section 6.4.7). The band diagram (along a vertical line in this overlap region) during the operation is shown in Fig. 9–38b. Interestingly, in a flash device we make use of two effects that are considered to be "problems" in regular MOS devices: hot carrier effects and Fowler–Nordheim tunneling.

During the read operation, one applies a moderate voltage (~1 V) to the bitline (drain of the MOSFET), and a wordline (control gate) voltage V_{CG} that causes the capacitively coupled floating gate voltage to be between that of the high V_T and the low V_T state of the programmed flash memory cell (Fig. 9–39). There will be negligible drain current flow in the bit line (drain) for the high V_T case because the gate voltage is less than the threshold voltage. We will then interpret the selected cell as being in state "1". For the low V_T case, since the applied gate voltage is higher than the threshold voltage of the cell, there will be drain current flow in the bitline (drain), and this can be interpreted as state "0". The read operation can be understood by looking at the transfer characteristics of the MOSFET in the programmed and erased states (Fig. 9–39).

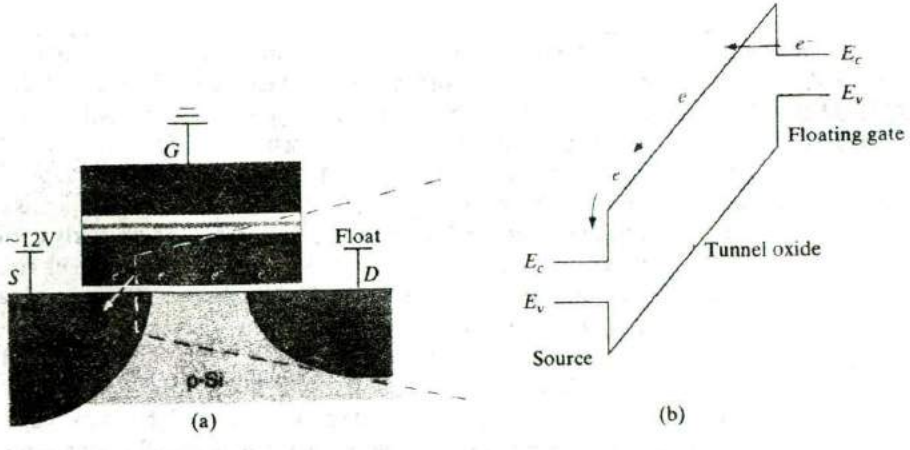


Figure 9–38
Fowler–Nordheim tunneling erasure: (a) flash memory cell structure with typical biases required for ersasing the cell; (b) band diagram as a function of depth in the gate/source overlap region of the MOSFET showing quantum mechanical tunneling of carriers from the floating gate into the oxide, and subsequent drift to the source.

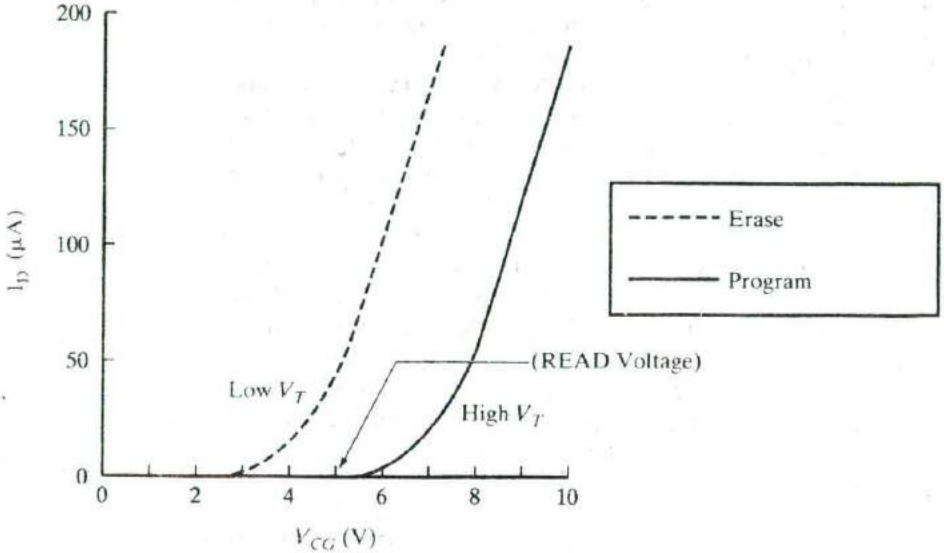


Figure 9–39
Drain (bitline) current versus control gate (wordline) voltage transfer characteristics of the MOSFET in a flash cell: if the cell is programmed to a high V_T (logic "1"), and a read voltage is applied to the wordline that is below this V_T, the MOSFET does not conduct, and there is negligible bitline current. On the other hand, if the cell had been erased to a low V_T state (logic "0") the MOSFET is turned ON, and there is significant bitline current.

9.6 TESTING, BONDING, AND PACKAGING

After the preceding discussions of rather dramatic fabrication steps in monolithic circuit technology, the processes of attaching leads and packaging the devices could seem rather mundane. Such an impression would be far from accurate, however, since the techniques discussed in this section are crucial to the overall fabrication process. In fact, the handling and packaging of individual circuits can be the most critical steps of all from the viewpoints of cost and reliability. The individual IC chip must be connected properly to outside leads and packaged in a way that is convenient for use in a larger circuit or system. Since the devices are handled individually once they are separated from the wafer, bonding and packaging are expensive processes. Considerable work has been done to reduce the steps required in bonding. We shall discuss the most straightforward technique first, which involves bonding individual leads from the contact pads on the circuit to terminals in the package. Then we shall consider two important methods for making all bonds simultaneously. Finally, we shall discuss a few typical packaging methods for ICs.

9.6.1 Testing

After the wafer of monolithic circuits has been processed and the final metallization pattern defined, it is placed in a holder under a microscope and is aligned for testing by a multiple-point probe (Fig. 9–40). The probe contacts the various pads on an individual circuit, and a series of tests are made of the electrical properties of the device. The various tests are programmed to be made

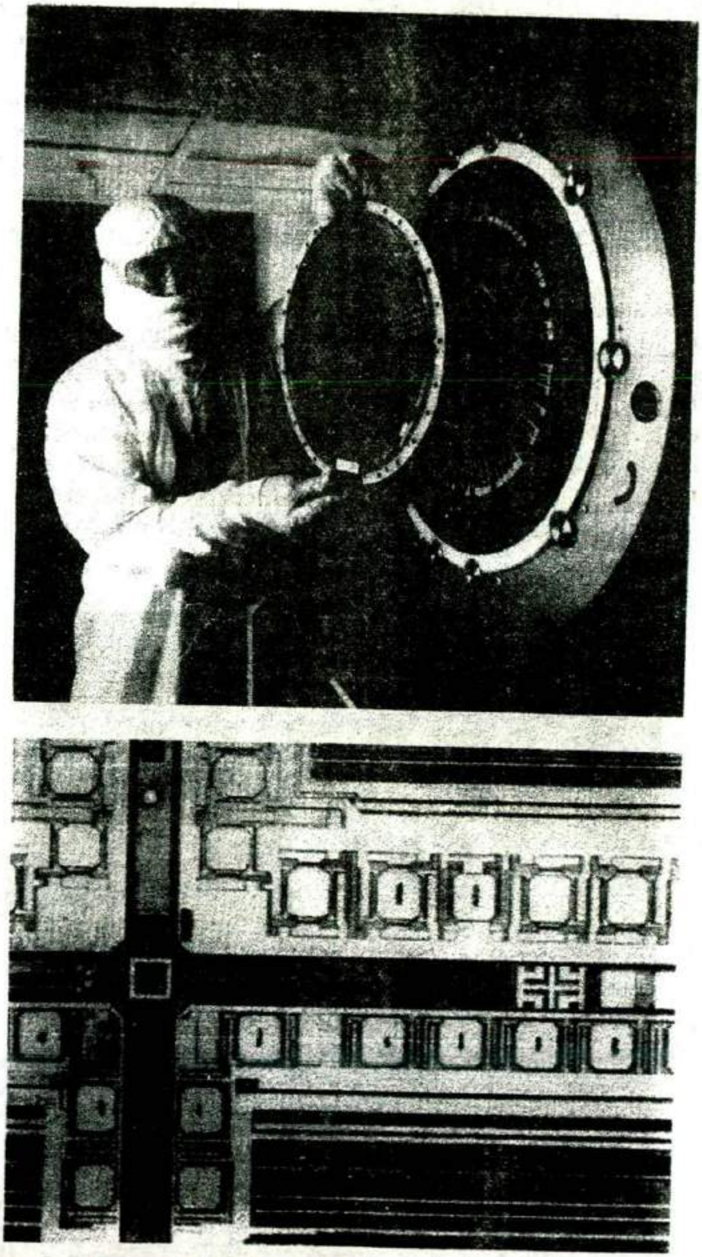


Figure 9–40 Automated probing of devices: (a) high speed testing of ICs is done using probe cards having a rigidly fixed array of probes that corresponds to the bond pad pattern on the IC to be tested. Many electrical signals are provided or measured by the automated tester at the various pins. After one chip is tested, the tester mechanically moves the wafer to the next die location; (b) array of Al bond pads near the chip periphery, some showing probe marks. The space between the arrays of bond pads is the "scribe line" along which the wafer will be sawed into individual chips after testing. (Photographs courtesy of Micron Technology.)

automatically in a very short time. These tests may take only milliseconds for a simple circuit, to several seconds for a complex ULSI chip. The information from these tests is fed into a computer, which compares the results with information stored in its memory, and a decision is made regarding the acceptability of the circuit. If there is some defect so that the circuit falls below specifications, the computer remembers that chip must be discarded. The probe automatically steps the prescribed distance to the next circuit on the wafer and repeats the process. After all of the circuits have been tested and the substandard ones noted, the wafer is removed from the testing machine, sawed between the circuits, and broken apart (Fig. 9–41). Then each die that passed the test is picked up and placed in the package. In the testing process, information from tests on each die can be stored to facilitate analysis of the rejected circuits or to evaluate the fabrication process for possible changes.

9.6.2 Wire Bonding

The earliest method used for making contacts from the monolithic chip to the package was the bonding of fine Au wires. Later techniques expanded wire bonding to include Al wires and several types of bonding processes. Here we shall outline only a few of the most important aspects of wire bonding.

If the chip is to be wire bonded, it is first mounted solidly on a metal lead frame or on a metallized region in the package. In this process a thin layer of Au (perhaps combined with Ge or other elements to improve the

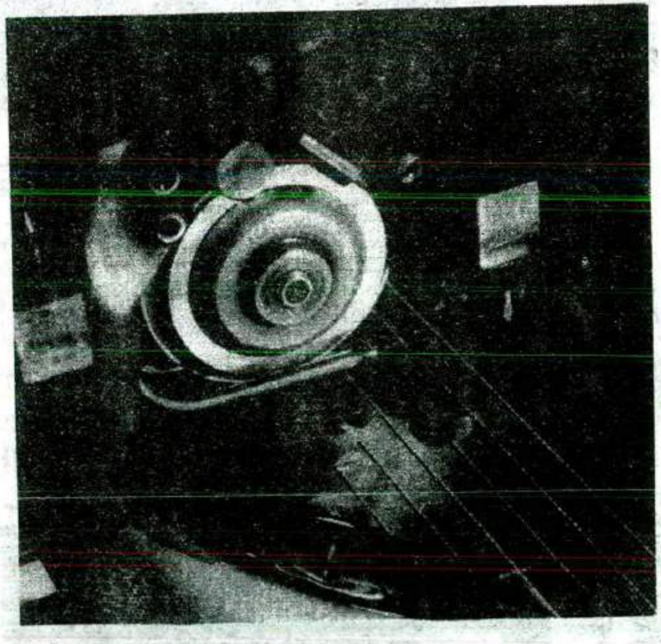


Figure 9–41 Sawing of a wafer along scribe lines: After the wafer is tested, the "known good dies" are "inked" or identified. The wafer is then sawed into individual dies for subsequent packaging. (Photograph courtesy of Micron Technology.)

metallurgy of the bond) is placed between the bottom of the chip and the substrate; heat and a slight scrubbing motion are applied, forming an alloyed bond which holds the chip firmly to the substrate. This process is called *die bonding*. Generally, die bonding is done by a robotic arm that picks up each die, orients it, and places it for bonding. Once the chip is mounted, the interconnecting wires are attached from the various contact pads to posts on the package (Fig. 9–42).

In Au wire bonding, a spool of fine Au wire (about 0.007–0.002-inch diameter) is mounted in a *lead bonder* apparatus, and the wire is fed through a glass or tungsten carbide *capillary* (Fig. 9–43a). A hydrogen gas flame jet is swept past the wire to form a ball on the end. In *thermocompression bonding* the chip (or in some cases the capillary) is heated to about 360°C, and the capillary is brought down over the contact pad. When pressure is exerted by the capillary on the ball, a bond is formed between the Au ball and the Al pad (Fig. 9–43b). Then the capillary is raised and moved to a post on the package. The capillary is brought down again, and the combination of force and temperature bonds the wire to the post. After raising the capillary again, the hydrogen flame is swept past, forming a new ball (Fig. 9–43c); then the process is repeated for the other pads on the chip.

There are many variations in this basic method. For example, the substrate heating can be eliminated by *ultrasonic bonding*. In this method a tungsten carbide capillary is held by a tool connected to an ultrasonic transducer.

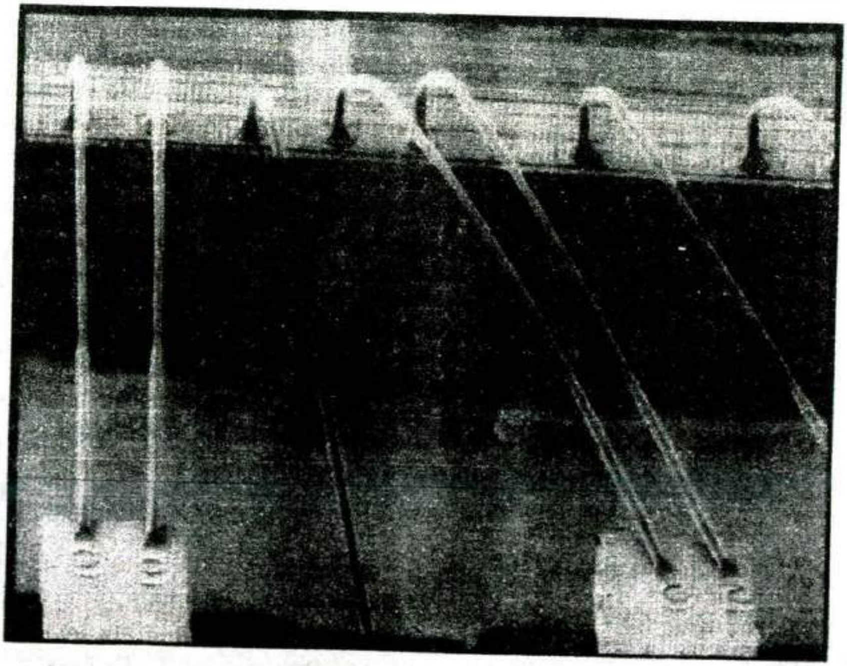


Figure 9–42
Attachment of leads from the Al pads on the periphery of the chip to posts on the package. (Photograph courtesy of Micron Technology.)

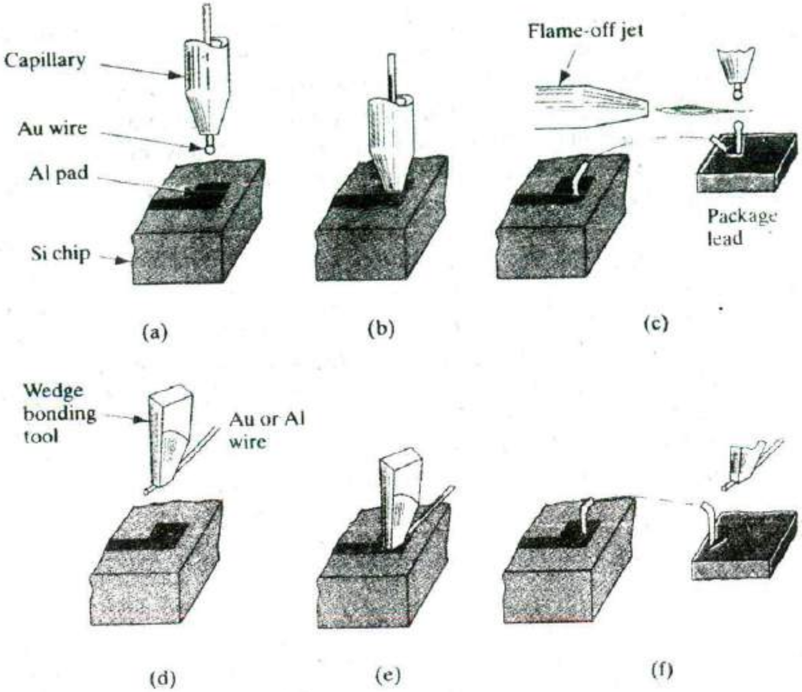


Figure 9–43 Wire bonding techniques: (a) capillary positioned over one of the contact pads for a ball (nail-head) bond; (b) pressure exerted to bond the wire to the pad; (c) post bond and flame-off; (d) wedge bonding tool; (e) pressure and ultrasonic energy applied; (f) post bond completed and wire broken or cut for next bond.

When it is in contact with a pad or a post, the wire is vibrated under pressure to form a bond. Other variations include techniques for automatically removing the "tail," which is left on the post in Fig. 9–43c. When the bond to the chip is made by exerting pressure on a ball at the end of the Au wire, it is called a *ball bond* or a *nail-head bond*, because of the shape of the deformed ball after the bond is made (Fig 9–44a).

Aluminum wire can be used in ultrasonic bonding; it has several advantages over Au, including the absence of possible metallurgical problems in bonds between Au and Al pads. When Al wire is used, the flame-off step is replaced by cutting or breaking the wire at appropriate points in the process. In forming a bond, the wire is bent under the edge of a wedge-shaped bonding tool (Fig. 9–43d). The tool then applies pressure and ultrasonic vibration, forming the bond (Fig. 9–43e and f). The resulting flat bond, formed by the bent wire wedged between the tool and the bonding surface, is called a *wedge bond.* A closeup view of ball and wedge bonds is given in Fig. 9–44.

9.6.3 Flip-chip Techniques

The time consumed in bonding wires individually to each pad on the chip can be overcome by several methods of simultaneous bonding. The *flip-chip*

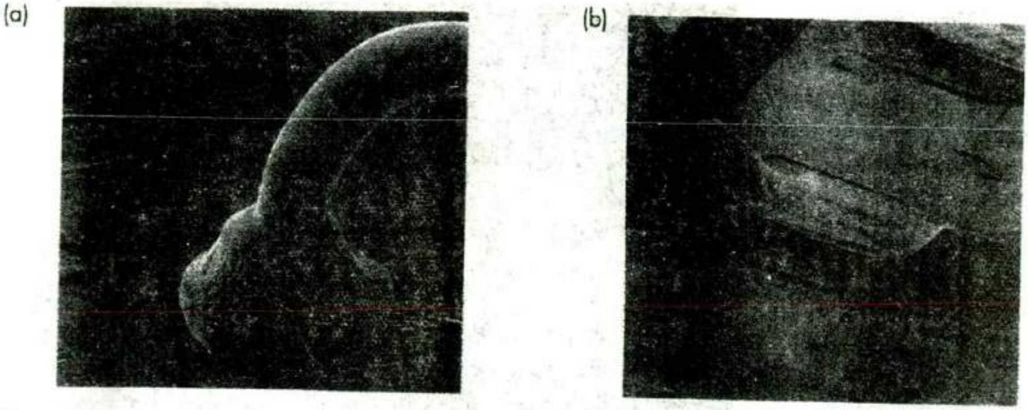


Figure 9–44
Scanning electron micrographs of a ball bond (a) and a wedge bond (b). (Photographs courtesy of Micron Technology, Inc.)

approach is typical of these methods. Relatively thick metal is deposited on the contact pads before the devices are separated from the wafer. After separation, the deposited metal is used to contact a matching metallized pattern on the package substrate.

In the flip-chip method, "bumps" of solder or special metal alloys are deposited on each contact pad. These metal bumps can be distributed over the die (Fig. 9–45). After separation from the wafer, each chip is turned upside down, and the bumps are properly aligned with the metallization pattern on the substrate. At this point, ultrasonic bonding or solder attaches each bump to its corresponding connector on the substrate. An obvious advantage of this method is that all connections are made simultaneously. Disadvantages include the fact that the bonds are made under the chip and therefore cannot be inspected visually. Furthermore, it is necessary to heat and/or exert pressure on the chip.

9.6.4 Packaging

The final step in IC fabrication is packaging the device in a suitable medium that can protect it from the environment of its intended application. In most cases this means the surface of the device must be isolated from moisture and contaminants and the bonds and other elements must be protected from corrosion and mechanical shock. The problems of surface protection are greatly minimized by modern passivation techniques, but it is still necessary to provide some protection in the packaging. In every case, the choice of package type must be made within the requirements of the application and cost considerations. There are many techniques for encapsulating devices, and the various methods are constantly refined and

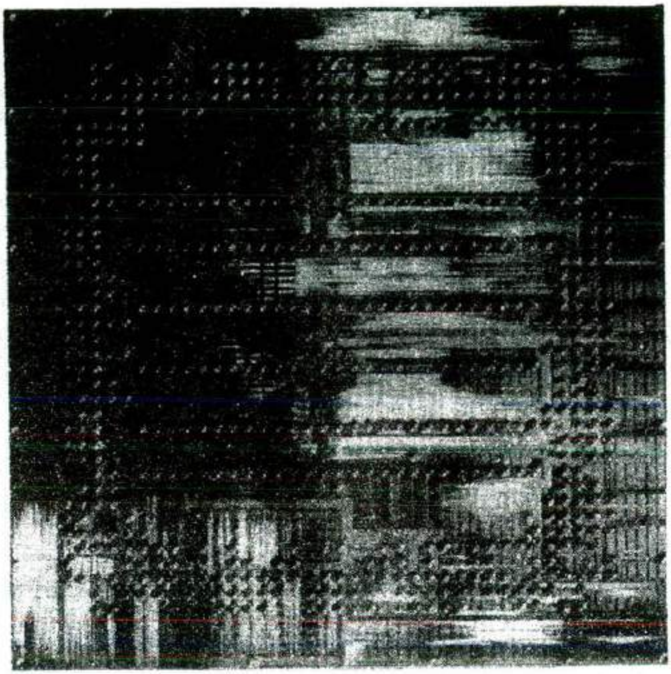


Figure 9–45 Flip-chip bonding. The Power PC chip has metallized "bumps" distributed over the surface instead of contact pads around the periphery. These bumps are aligned with the interconnection pattern on the package and bonded simultaneously. (Photograph courtesy of IBM Corp.)

changed. Here we shall consider just a few general methods for the purpose of illustration.

In the early days of IC technology, all devices were packaged in metal headers. In this method the device is alloyed to the surface of the header, wire bonds are made to the header posts, and a metal lid is welded over the device and wiring. Although this method has several drawbacks, it does provide complete sealing of the unit from the outside environment. This is often called a *hermetically sealed* device. After the chip is mounted on the header and bonds are made to the posts, the header cap can be welded shut in a controlled environment (e.g., an inert gas), which maintains the device in a prescribed atmosphere.

Integrated circuits are now mounted in packages with many output leads (Fig. 9–46). In one version the chip is mounted on a stamped metal lead frame and wire bonding is done between the chip and the leads. The package is formed by applying a ceramic or plastic case and trimming away the unwanted parts of the lead frame.

Broadly speaking, packages can be through-hole-mount that involve inserting the package pins through holes on the printed circuit board (PCB) before soldering, or surface mount type where the leads do not pass through holes in the PCB. Instead, surface-mounted package leads are aligned to electrical contacts on the PCB, and are connected simultaneously by solder

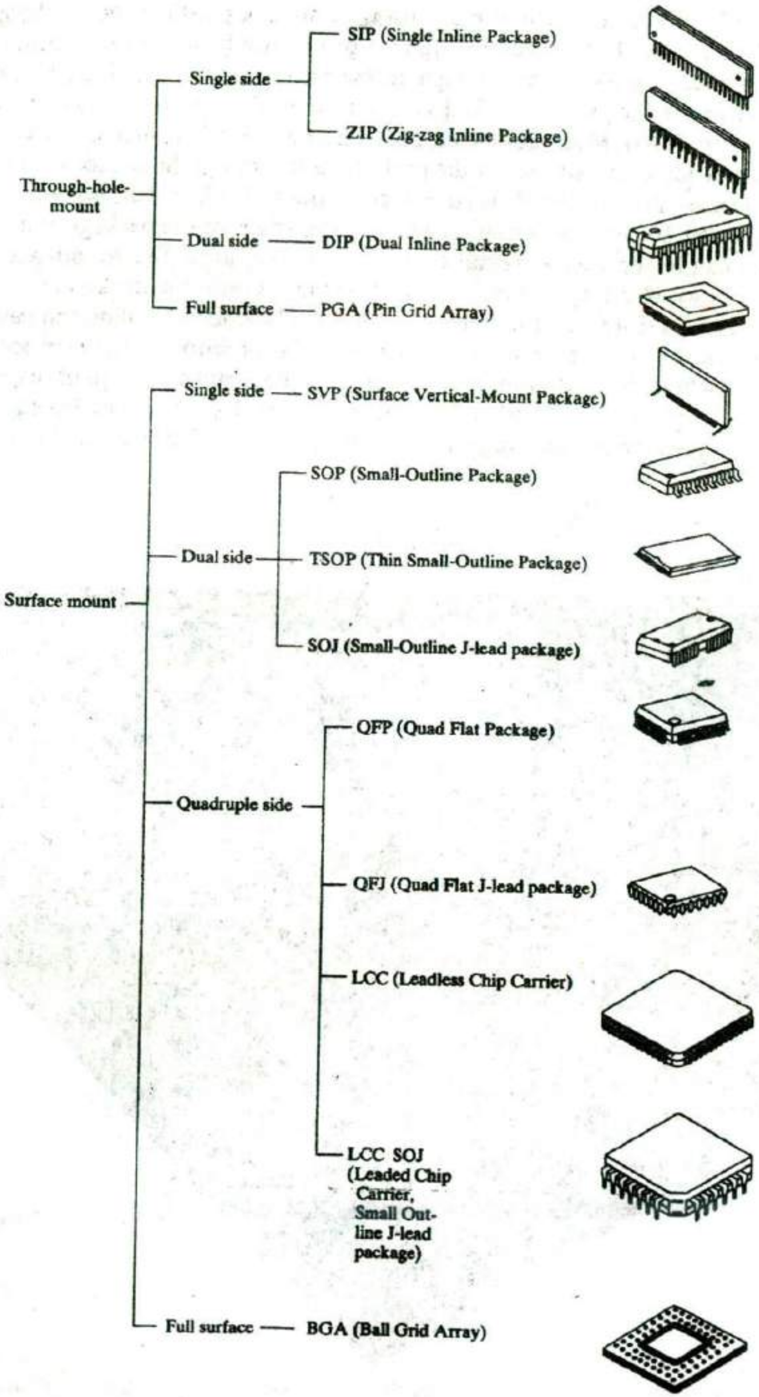

Through-hole-mount
Single side
SIP (Single Inline Package)
ZIP (Zig-zag Inline Package)
Dual side
DIP (Dual Inline Package)
Full surface
PGA (Pin Grid Array)
Surface mount
Single side
SVP (Surface Vertical-Mount Package)
Dual side
SOP (Small-Outline Package)
TSOP (Thin Small-Outline Package)
SOJ (Small-Outline J-lead package)
Quadruple side
QFP (Quad Flat Package)
QFJ (Quad Flat J-lead package)
LCC (Leadless Chip Carrier)
LCC SOJ (Leaded Chip Carrier, Small Out-line J-lead package)
Full surface
BGA (Ball Grid Array)

Figure 9–46 Various types of packaging for ICs: Packages can be through-hole-mount or surface mount type, and be made out of plastic or ceramic. The pins can be on one side (SIP), two sides (DIP) or four side (quad) of the package, or distributed over the surface of the package (PGAs or BGAs)

reflow. Most packages can be made using ceramic or plastic (which is cheaper). The ICs are hermetically sealed for protection from the environment. The pins can be on one side (single inline or zig-zag pattern of leads), two sides (dual inline package or DIP) or four sides of the package (quad package) (Fig. 9–46). More advanced packages have leads distributed over a large portion of the surface of the package as in through-hole–mounted pin grid arrays (PGAs) (Fig. 9–47) or surface-mounted ball grid arrays (BGAs) (Fig. 9–48). By not restricting the leads to the edges of the package, the pin count can be increased dramatically, which is very attractive for advanced ULSI in which a large number of electrical leads must be accessed.

Since a sizable fraction of the cost of an IC is due to bonding and packaging, there have been a number of innovations for automating the process. These include the use of film reels that contain the metal contact pattern onto which the chips can be bonded. The film can then be fed into packaging equipment, where the position registration capabilities of a film reel can be used

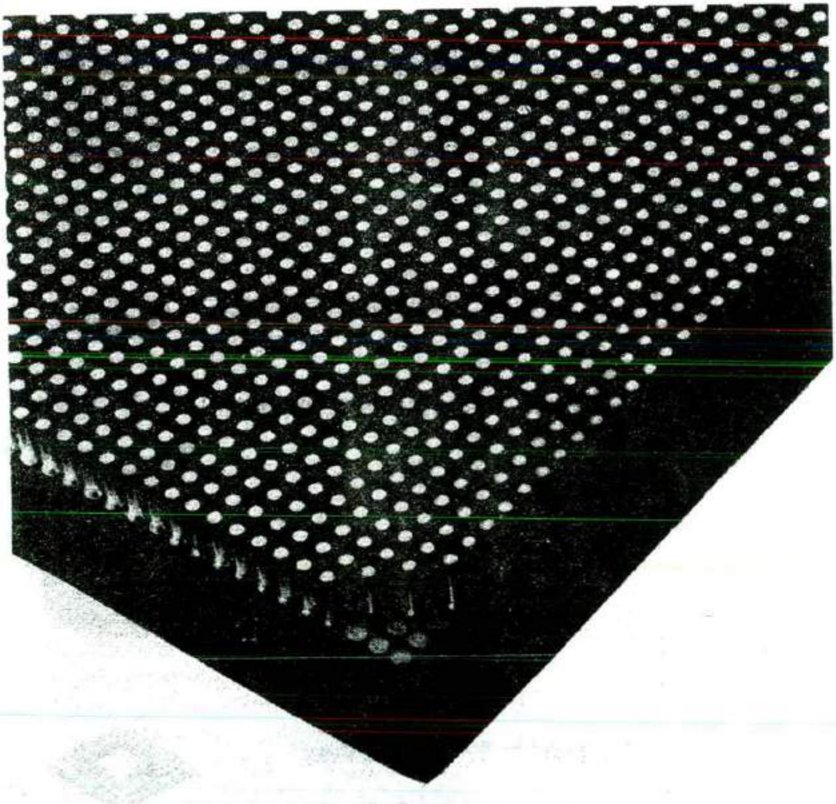


Figure 9–47 Ceramic column grid array (CCGA): This advanced ceramic package is a type of pin grid array made up of several hundred metal columns. Several ICs with metallized solder bumps on them as in Fig. 9–45 can be flip chip bonded on the back of this package, making this a multi-chip module (MCM). (Photograph courtesy of IBM.)

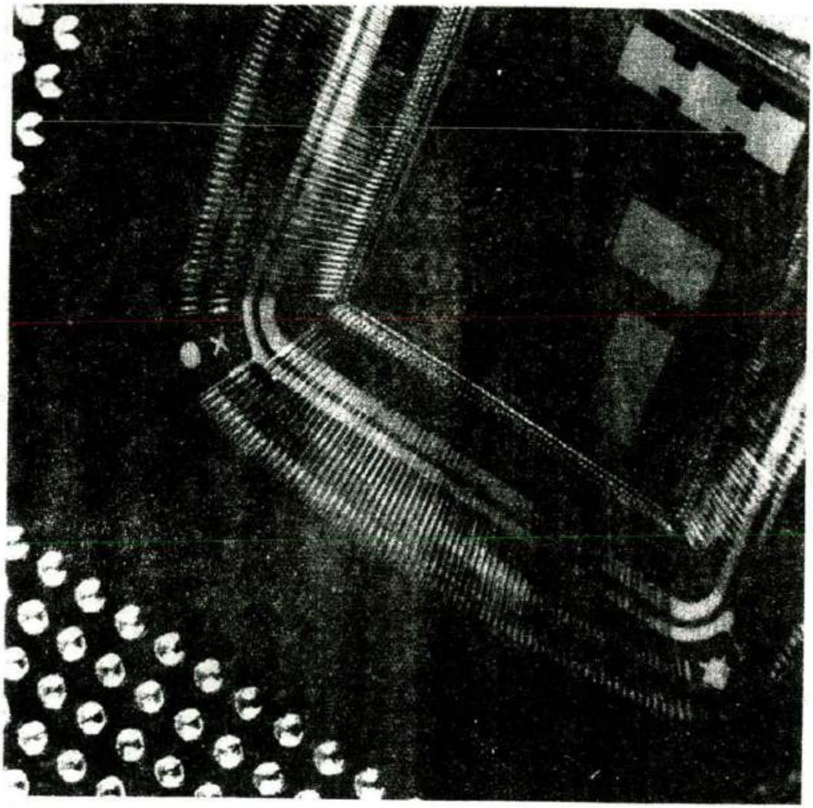


Figure 9–48 Ball grid array: In this package, the IC in the middle is wire bonded to electrical connections on the package. The package itself has an array of solder "balls" on the top, which can be properly aligned and surface-mount connected simultaneously to electrical sockets on a PCB using solder reflow. (Photograph courtesy of IBM.)

for automated handling. This process, called *tape-automated bonding (TAB)*, is particularly useful in mounting several chips on a large ceramic substrate having multilevel interconnection patterns (called a *multichip module*).

PROBLEMS

9.1 Assume that boron is diffused into a uniform n-type Si sample, resulting in a net doping profile $N_a(x)–N_d$. Set up an expression relating the sheet resistance of the diffused layer to the acceptor profile $N_a(x)$ and the junction depth x_j. Assume that $N_a(x)$ is much greater than the background doping N_d over most of the diffused layer.

9.2 A typical sheet resistance of a base diffusion layer is 200 Ω/square.

(a) What should be the aspect ratio of a 10-k Ω resistor, using this diffusion?

(b) Draw a pattern for this resistor (see Fig. 9–12b) which uses little area for a width $w = 5$ μm.

9.3 A 3-μm n-type epitaxial layer ($N_d = 10^{16}$ cm^{-3}) is grown on a p-type Si substrate. Areas of the n layer are to be junction isolated (see Fig. 9–12a) by a boron diffusion at 1200°C ($D = 2.5 \times 10^{-12}$ cm^2/s). The surface boron concentration is held constant at 10^{20} cm^{-3} (see Prob. 5.2).

(a) What time is required for this isolation diffusion?

(b) How far does an Sb-doped buried layer ($D = 2 \times 10^{-13}$ cm^2/s) diffuse into the epitaxial layer during this time, assuming the concentration at the substrate-epitaxial boundary is constant at 10^{20} cm^{-3}?

9.4 A 500 μm thick p-type Si wafer with a doping level of 1×10^{15} cm^{-3} has a certain region in which we do a constant source solid-solubility-limited P diffusion, resulting in a junction depth of 0.8 μm and a surface concentration of 6×10^{19} cm^{-3}. We do sheet resistance measurements on the two parts of the wafer. What is the measured sheet resistance of the p-type part? If we have a sheet resistance of 90 Ω/square in the n-type part, what is the *average* resistivity there? At what temperature was the P diffusion done, keeping in mind that typical diffusion temperatures are less than 1100°C? (Refer to Prob. 5.2.)

READING LIST

Blouke, M. M. "Charge-Coupled Devices Reach Maturity." *Laser Focus World* 27 (March 1991): A17– A19.

Campbell, S. A. *The Science and Engineering of Microelectronic Fabrication*. New York: Oxford, 1996.

Chang, C. Y., and S. M. Sze. *ULSI Technology*. New York: McGraw-Hill, 1996.

Ghandhi, S. K. *VLSI Fabrication Principles*, 2nd ed. New York: Wiley, 1994.

Hess, D. W., and K. F. Jensen, eds. *Microelectronics Processing: Chemical Engineering Aspects*. Washington, DC: American Chemical Society, 1989.

Hughes, W. A., A. A. Rezazadeh, and C. E. C. Wood. *GaAs Integrated Circuits*. Oxford: BSP Professional Books, 1988.

Jaeger, R. C. *Modular Series on Solid State Devices: Vol. V. Introduction to Microelectronic Fabrication*. Reading, MA: Addison-Wesley, 1988.

Levenson, M. D. "Wavefront Engineering for Photolithography." *Physics Today* 46 (July 1993): 28– 36.

Moreau, W. M. *Semiconductor Lithography: Principles, Practices, and Materials*. New York: Plenum Press, 1988.

Moslehi, M. M., R. A. Chapman, M. Wong, A. Paranjpe, H. N. Najm, J. Kuehne, R. F. Yeakley, and C. J. Davis. "Single-Wafer Integrated Semiconductor Device Processing." *IEEE Transactions on Electron Devices* 39 (January 1992):4– 32.

Oberai, A. S. "Lithography—Challenges of the Future." *Solid State Technology* 30 (September 1987): 123– 8.

Runyan, W. R., and K. E. Bean. *Semiconductor Integrated Circuit Processing Technology.* Reading, MA: Addison-Wesley, 1990.

Ruska, W. S. *Microelectronic Processing: An Introduction to the Manufacture of Integrated Circuits.* New York: McGraw-Hill, 1987.

Seraphim, D. P., R. C. Lasky, and C. Y. Li, eds. *Principles of Electronic Packaging.* New York: McGraw-Hill, 1989.

Tandon, U. S. "An Overview of X-ray Lithography for Use in Semiconductor Device Preparation." *Vacuum* 42 (1991): 1219–28.

Uyemura, J. P. *Fundamentals of MOS Digital Integrated Circuits*. Reading, MA: Addison-Wesley, 1988.

Wolf, S., and R. N. Tauber. *Silicon Processing for the VLSI Era.* Sunset Beach, CA: Lattice Press, 1986.

Zucker, J. E. "Quantum Effects Enhance Integrated Optics Performance." *Laser Focus World* 29 (March 1993): 101–2+.

Chapter 10
Negative Conductance Microwave Devices

We have discussed a number of devices that are useful in microwave circuits, such as the varactor and specially designed high-frequency transistors, which can provide amplification and other functions at microwave frequencies up to 10^{11} Hz. However, transit time and other effects limit the application of transistors beyond the 10^{11} Hz range. Therefore, other devices are required to perform electronic functions such as switching and d-c to microwave power conversion at higher frequencies.

Several important devices for high-frequency applications use the instabilities that occur in semiconductors. An important type of instability involves *negative conductance*. Here we shall concentrate on three of the most commonly used negative conductance devices: *Esaki* or *tunnel* diodes, which depend on quantum-mechanical tunneling; transit time diodes, which depend on a combination of carrier injection and transit time effects; and *Gunn* diodes, which depend on the transfer of electrons from a high-mobility state to a low-mobility state. Each is a two-terminal device that can be operated in a negative conductance mode to provide amplification or oscillation at microwave frequencies in a proper circuit.

10.1 TUNNEL DIODES

The tunnel diode is a p-n junction device that operates in certain regions of its *I–V* characteristic by the quantum mechanical tunneling of electrons through the potential barrier of the junction (see Sections 2.4.4 and 5.4.1). The tunneling process for reverse current is essentially the Zener effect, although negligible reverse bias is needed to initiate the process in tunnel diodes. This device can be used in many applications, including high-speed switching and logic circuits. As we shall see in this section, the tunnel diode (often called the Esaki diode after L. Esaki, who in 1973 received the Nobel prize for his work on the effect) exhibits the important feature of *negative resistance* over a portion of its *I–V* characteristic.

10.1.1 Degenerate Semiconductors

Thus far we have discussed the properties of relatively pure semiconductors; any impurity doping represented a small fraction of the total atomic density of the material. Since the few impurity atoms were so widely spaced throughout the sample, we could be confident that no charge transport could take place within the donor or acceptor levels themselves. What happens, however, if we continue to dope a semiconductor with impurities of either type? As might be expected, a point is reached at which the impurities become so closely packed within the lattice that interactions between them cannot be ignored. For example, donors present in high concentrations (e.g., 10^{20} donors/cm^3) are so close together that we can no longer consider the donor level as being composed of discrete, noninteracting energy states. Instead, the donor states form a band, which may overlap the bottom of the conduction band. If the conduction band electron concentration n exceeds the effective density of states N_c, the Fermi level is no longer within the band gap but lies within the conduction band. When this occurs, the material is called *degenerate* n-type. The analogous case of degenerate p-type material occurs when the acceptor concentration is very high and the Fermi level lies in the valence band. We recall that the energy states below E_F are mostly filled and states above E_F are empty, except for a small distribution dictated by the Fermi statistics. Thus in a degenerate n-type sample the region between E_c and E_F is for the most part filled with electrons, and in degenerate p-type the region between E_v and E_F is almost completely filled with holes.

10.1.2 Tunnel Diode Operation

A p-n junction between two degenerate semiconductors is illustrated in terms of energy bands in Fig. 10–1a. This is the equilibrium condition, for which the Fermi level is constant throughout the junction. We notice that E_{Fp} lies below the valence band edge on the p side and E_{Fn} is above the conduction band edge on the n side. Thus the bands must overlap on the energy scale in order for E_F to be constant. This overlapping of bands is very important; it means that with a small forward or reverse bias, filled states and empty states appear opposite each other, separated by essentially the width of the depletion region. If the metallurgical junction is sharp, as in an alloyed junction, the depletion region will be very narrow for such high doping concentrations, and the electric field at the junction will be quite large. Thus the conditions for electron tunneling are met—filled and empty states separated by a narrow potential barrier of finite height.

As mentioned previously, the filled and empty states are distributed about E_F according to the Fermi distribution function; thus there are some filled states above E_{Fp} and some empty states below E_{Fn}. In Fig. 10–1 the bands are shown filled to the Fermi level for convenience of illustration, with the understanding that a distribution is implied.

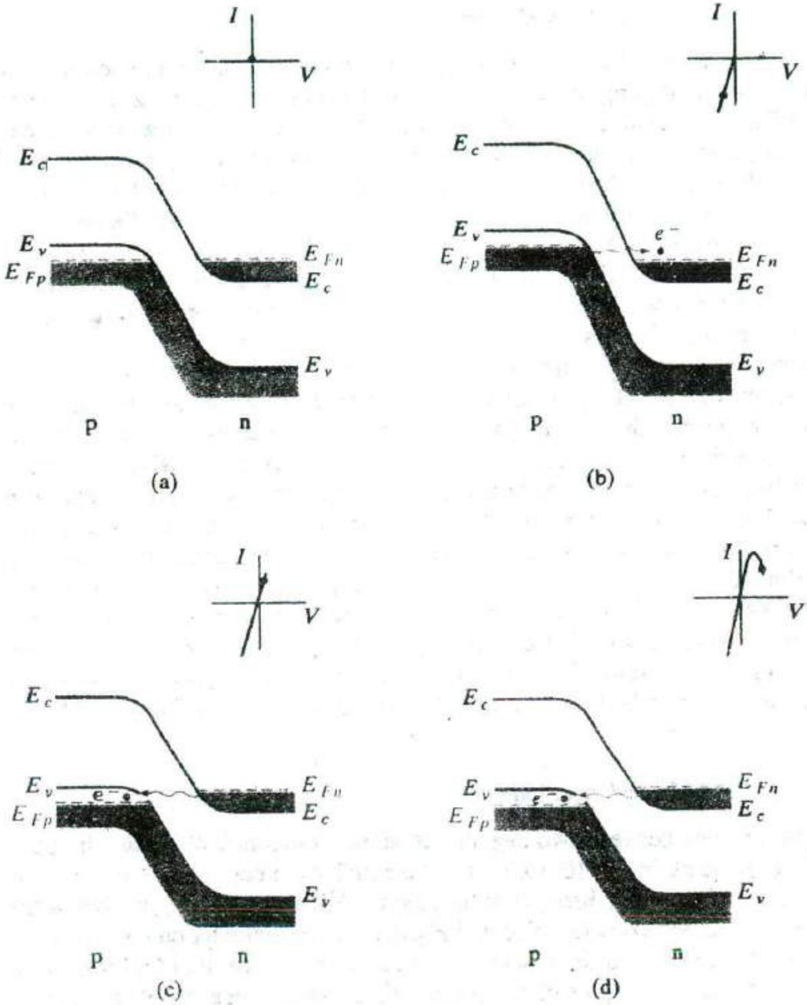


Figure 10–1 Tunnel diode band diagrams and I–V characteristics for various biasing conditions: (a) equilibrium (zero bias) condition, no net tunneling; (b) small reverse bias, electron tunneling from p to n; (c) small forward bias, electron tunneling from n to p; (d) increased forward bias, electron tunneling from n to p decreases as bands pass by each other.

Since the bands overlap under equilibrium conditions, a small reverse bias (Fig. 10–1b) allows electron tunneling from the filled valence band states below E_{Fp} to the empty conduction band states above E_{Fn}. This condition is similar to the Zener effect except that no bias is required to create the condition of overlapping bands. As the reverse bias is increased, E_{Fn} continues to move down the energy scale with respect to E_{Fp}, placing more filled states on the p side opposite empty states on the n side. Thus the tunneling of electrons from p to n increases with increasing reverse bias. The resulting conventional current is opposite to the electron flow, that is, from n to p. At equilibrium (Fig. 10–1a) there is equal tunneling from n to p and from p to n, given a zero net current.

When a small forward bias is applied (Fig. 10–1c), E_{Fn} moves up in energy with respect to E_{Fp} by the amount qV. Thus electrons below E_{Fn} on the n side are placed opposite empty states above E_{Fp} on the p side. Electron tunneling occurs from n to p as shown, with the resulting conventional current from p to n. This forward tunneling current continues to increase with increased bias as more filled states are placed opposite empty states. However, as E_{Fn} continues to move up with respect to E_{Fp}, a point is reached at which the bands begin to pass by each other. When this occurs, the number of filled states opposite empty states decreases. The resulting decrease in tunneling current is illustrated in Fig. 10–1d. This region of the I–V characteristic is important in that the *decrease* of tunneling current with *increased* bias produces a region of negative slope; that is, the *dynamic resistance* dV/dI is negative. This negative resistance region is useful in a number of applications.

If the forward bias is increased beyond the negative resistance region, the current begins to increase again (Fig. 10–2). Once the bands have passed each other, the characteristic resembles that of a conventional diode. The forward current is now dominated by the diffusion current—electrons surmounting the potential barrier from n to p and holes surmounting their potential barrier from p to n. Of course, the diffusion current is present in the forward tunneling region, but it is negligible compared to the tunneling current.

The total tunnel diode characteristic (Fig. 10–3) has the general shape of an N (if a little imagination is applied); therefore, it is common to refer to this characteristic as exhibiting a *type N negative resistance*. It is also called a *voltage-controlled negative resistance*, meaning that the current decreases rapidly at some critical voltage (in this case the *peak voltage* V_p, taken at the point of maximum forward tunneling).

The values of *peak tunneling current* I_p and *valley current* I_v (Fig. 10–3) determine the magnitude of the negative resistance slope for a diode of given material. For this reason, their ratio I_p/I_v is often used as a figure of merit for

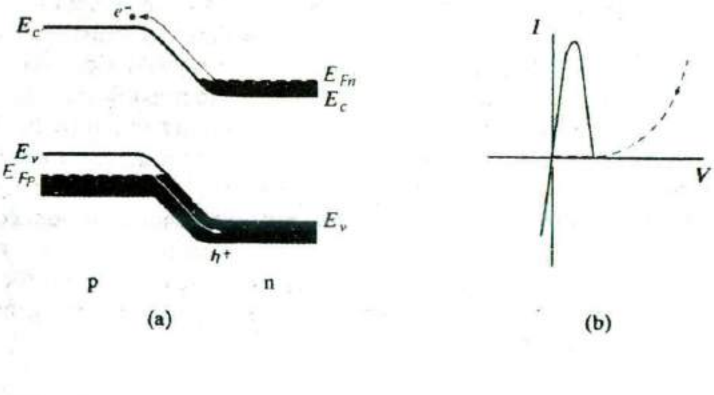


Figure 10–2
Band diagram (a) and I–V characteristic (b) for the tunnel diode beyond the tunnel current region. In (b) the tunneling component of current is shown by the solid curve and the diffusion current component is dashed.

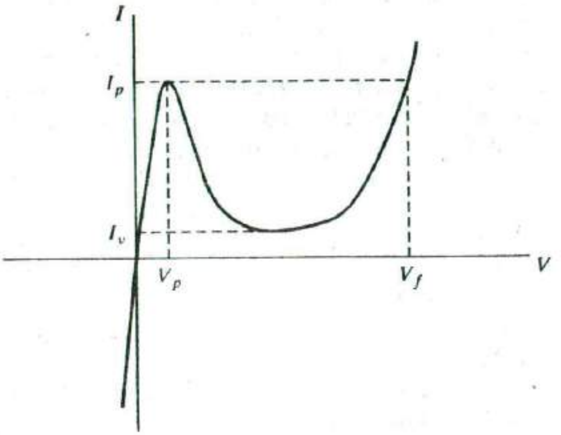


Figure 10–3 Total tunnel diode characteristic.

the tunnel diode. Similarly, the ratio V_p/V_f is a measure of the voltage spread between the two positive resistance regions.

10.1.3 Circuit Applications

The negative resistance of the tunnel diode can be used in a number of ways to achieve switching, oscillation, amplification, and other circuit functions. This wide range of applications, coupled with the fact that the tunneling process does not present the time delays of drift and diffusion, makes the tunnel diode a natural choice for certain high-speed circuits. However, the tunnel diode has not achieved widespread application, because of its relatively low current operation and competition from other devices.

10.2 THE IMPATT DIODE

In this section we describe a type of microwave negative conductance device that operates by a combination of carrier injection and transit time effects. Diodes with simple p-n junction structure, or with variations on that structure, are biased to achieve tunneling or avalanche breakdown, with an a-c voltage superimposed on the d-c bias. The carriers generated by the injection process are swept through a drift region to the terminals of the device. We shall see that the a-c component of the resulting current can be approximately 180° out of phase with the applied voltage under proper conditions of bias and device configuration, giving rise to negative conductance and oscillation in a resonant circuit. Transit time devices can convert d-c to microwave a-c signals with high efficiency and are very useful in the generation of microwave power for many applications.

The original suggestion for a microwave device employing transit time effects was made by W. T. Read and involved an n$^+$-p-i-p$^+$ structure such as that shown in Fig. 10–4. This device operates by injecting carriers into the drift region and is called an *impact avalanche transit time (IMPATT)* diode. Although IMPATT operation can be obtained in simpler structures, the Read diode is best suited for illustration of the basic principles. The device consists essentially of two regions: (1) the n$^+$-p region at which avalanche multiplication occurs and (2) the i (essentially intrinsic) region through which generated holes must drift in moving to the p$^+$ contact. Similar devices can be built in the p$^+$-n-i-n$^+$ configuration, in which electrons resulting from avalanche multiplication drift through the i region, taking advantage of the higher mobility of electrons compared to holes.

Although detailed calculations of IMPATT operation are complicated and generally require computer solutions, the basic physical mechanism is simple. Essentially, the device operates in a negative conductance mode when the a-c component of current is negative over a portion of the cycle during which the a-c voltage is positive, and vice versa. The negative conductance occurs because of two processes, causing the current to lag behind the voltage in time: (1) a delay due to the avalanche process and (2) a further delay due to the transit time of the carriers across the drift region. If the sum of these delay times is approximately one-half cycle of the operating frequency, negative conductance occurs and the device can be used for oscillation and amplification.

From another point of view, the a-c conductance is negative if the a-c component of carrier flow drifts opposite to the influence of the a-c electric field. For example, with a d-c reverse bias on the device of Fig. 10–4, holes drift from left to right (in the direction of the field) as expected. Now, if we superimpose an a-c voltage such that $\mathscr{E}$ decreases during the negative half-cycle, we would normally expect the drift of holes to decrease also. However,

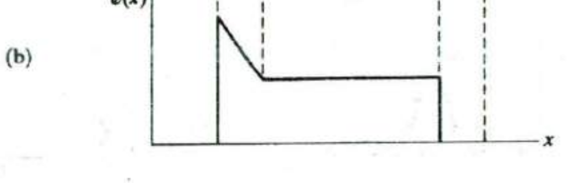


Figure 10–4
The Read diode: (a) basic device configuration; (b) electric field distribution in the device under reverse bias.

in IMPATT operation the drift of holes through the i region actually increases while the a-c field is decreasing. To see how this happens, let us consider the effects of avalanche and drift for various points in the cycle of applied voltage (Fig. 10–5).

To simplify the discussion, we shall assume that the p region is very narrow and that all the avalanche multiplication takes place in a thin region near the n^+-p junction. We shall approximate the field in the narrow p region by a uniform value. If the d-c bias is such that the critical field for avalanche $\mathscr{E}_a$ is just met in the n^+-p space charge region (Fig. 10–5a), avalanche multiplication begins at $t = 0$. Electrons generated in the avalanche move to the n^+ region, and holes enter the i drift region. We assume that device is mounted in a resonant microwave circuit so that an a-c signal can be maintained at a given frequency. As the applied a-c voltage goes positive, more and more holes are generated in the avalanche region. In fact, the pulse

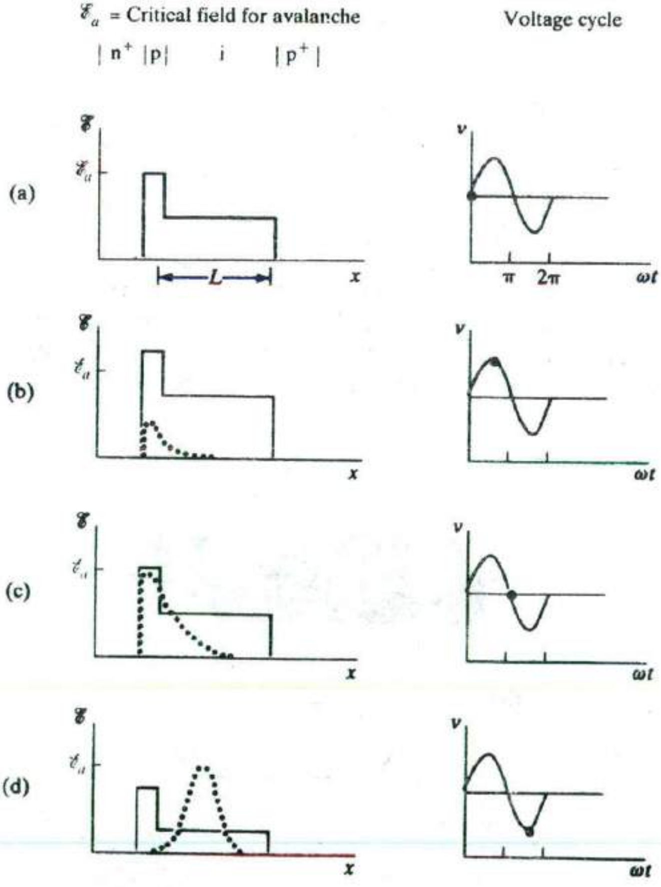


Figure 10–5 Time dependence of the growth and drift of holes during a cycle of applied voltage for the Read diode: (a) $\omega t = 0$; (b) $\omega t = \pi/2$; (c) $\omega t = \pi$; (d) $\omega t = 3\pi/2$. The hole pulse is sketched as a dotted line on the field diagram.

of holes (dotted line) generated by the multiplication process continues to grow as long as the electric field is above $\mathscr{E}_a$ (Fig. 10–5b). It can be shown that the particle current due to avalanche increases exponentially with time while the field is above the critical value. The important result of this growth is that the hole pulse reaches its peak value not at $\pi/2$ when the voltage is maximum, but at π (Fig. 10–5c). Therefore, there is a phase delay of $\pi/2$ inherent in the avalanche process itself. A further delay is provided by the drift region. Once the avalanche multiplication stops ($\omega t > \pi$), the pulse of holes simply drifts toward the p^+ contact (Fig. 10–5d). But during this period the a-c terminal voltage is negative. Therefore, the dynamic conductance is negative, and energy is supplied to the a-c field.

If the length of the drift region is chosen properly, the pulse of holes is collected at the p^+ contact just as the voltage cycle is completed, and the cycle then repeats itself. The pulse will drift through the length L of the i region during the negative half-cycle if we choose the transit time to be one-half the oscillation period

$$\frac{L}{v_d} = \frac{1}{2}\frac{1}{f}, \quad f = \frac{v_d}{2L} \tag{10–1}$$

where f is the operating frequency and v_d is the drift velocity for holes.[1] Therefore, for a Read diode the optimum frequency is one-half the inverse transit time of holes across the drift region v_d/L. In choosing an appropriate resonant circuit for this device, the parameter L is critical. For example, taking $v_d = 10^7$ cm/s for Si, the optimum operating frequency for a device with an i region length of 5 μm is $f = 10^7/2(5 \times 10^{-4}) = 10^{10}$ Hz. Negative resistance is exhibited by an IMPATT diode for frequencies somewhat above and below this optimum frequency for exact 180° phase delay. A careful analysis of the small-signal impedance shows that the minimum frequency for negative conductance varies as the square root of the d-c bias current for frequencies in the neighborhood of that described by Eq. (10–1).

Although the Read diode of Fig. 10–4 displays most directly the operation of IMPATT devices, simpler structures can be used, and in some cases they may be more efficient. Negative conductance can be obtained in simple p-n junctions or in p-i-n devices. In the case of the p-i-n, most of the applied voltage occurs across the i region, which serves as a uniform avalanche region and also as a drift region. Therefore, the two processes of delay due to avalanche and drift, which were separate in the case of the Read diode, are distributed within the i region of the p-i-n. This means that both electrons and holes participate in the avalanche and drift processes.

[1] In general, v_d is a function of the local electric field. However, these devices are normally operated with fields in the i region sufficiently large that holes drift at their scattering limited velocity (Fig. 3–24). For this case the drift velocity does not vary appreciably with the a-c variations in the field.

10.3 THE GUNN DIODE

Microwave devices that operate by the *transferred electron* mechanism are often called *Gunn diodes* after J. B. Gunn, who first demonstrated one of the forms of oscillation. In the transferred electron mechanism, the conduction electrons of some semiconductors are shifted from a state of high mobility to a state of low mobility by the influence of a strong electric field. Negative conductance operation can be achieved in a diode[2] for which this mechanism applies, and the results are varied and useful in microwave circuits.

First, we shall describe the process of electron transfer and the resulting change of mobility. Then we shall consider some of the modes of operation for diodes using this mechanism.

10.3.1 The Transferred Electron Mechanism

In Section 3.4.4 we discussed the nonlinearity of mobility at high electric fields. In most semiconductors the carriers reach a scattering limited velocity, and the velocity vs. field plot saturates at high fields (Fig. 3–24). In some materials, however, the energy of electrons can be raised by an applied field to the point that they transfer from one region of the conduction band to another, higher-energy region. For some band structures, negative conductivity can result from this electron transfer. To visualize this process, let us recall the discussion of energy bands in Section 3.1. The band diagrams we usually draw vs. distance in the sample are good approximations when the conduction electrons exist near the minimum energy of the conduction band. However, in the more complete band diagram, electron energy is plotted vs. the propagation vector **k**, as in Fig. 3–5. It was shown in Example 3-1 that the **k** vector is proportional to the electron momentum in the vector direction; therefore, energy bands such as those in Fig. 3–5 are said to be plotted in *momentum space.*

A simplified band diagram for GaAs is shown in Fig. 10–6 for reference; some of the detail has been omitted in this diagram to isolate the essential features of electron transfer between bands. In n-type GaAs the valence band is filled, and the *central valley* (or *minimum*) of the conduction band at Γ ($\mathbf{k} = 0$) normally contains the conduction electrons. There is a set of *subsidiary minima* at L (sometimes called *satellite valleys*) at higher energy,[3] but these minima are many kT above the central valley and are normally unoccupied. Therefore, the direct band gap at Γ and the energy bands centered at $\mathbf{k} = 0$ are generally used to describe the conduction processes in GaAs. This was true of our discussion of GaAs lasers in Section 8.4, for example. The presence of the satellite valleys at L is crucial to the Gunn effect, however. If the material is subjected to an electric field above some critical value (about 3000 V/cm), the

[2] These devices are called diodes, since they are two-terminal devices. No p-n junction is involved, however. Gunn effect and related devices utilize bulk instabilities, which do not require junctions.

[3] We have shown only one satellite valley for convenience; there are other equivalent valleys for different directions in k-space. The effective mass ratio of 0.55 refers to the combined satellite valleys.

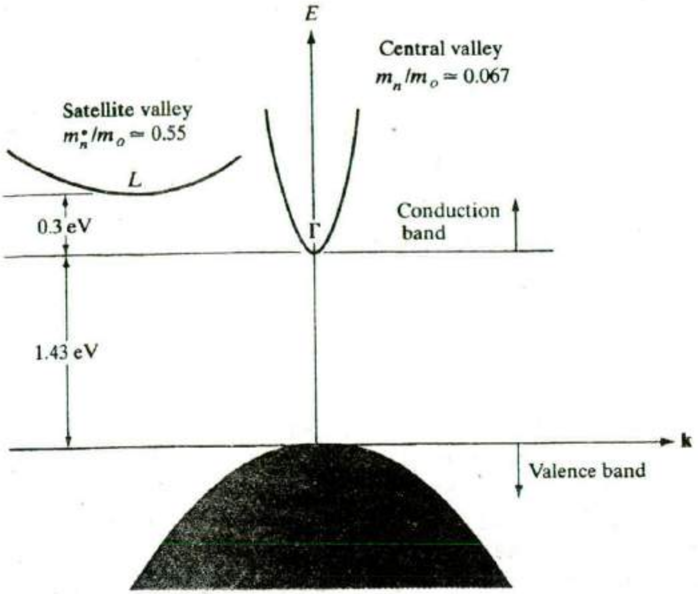


Figure 10–6 Simplified band diagram for GaAs, illustrating the lower (Γ) and upper (L) valleys in the conduction band.

electrons in the central Γ valley of Fig. 10–6 gain more energy than the 0.30 eV separating the valleys; therefore, there is considerable scattering of electrons into the higher-energy satellite valley at L.

Once the electrons have gained enough energy from the field to be transferred into the higher-energy valley, they remain there as long as the field is greater than the critical value. The explanation for this involves the fact that the combined effective density of states for the upper valleys is much greater than for the central valley (by a factor of about 24). Although we shall not prove it here, it seems reasonable that the probability of electron scattering between valleys should depend on the density of states available in each case, and that scattering from a valley with many states into a valley with few states would be unlikely. As a result, once the field increases above the critical value, most conduction electrons in GaAs reside in the satellite valleys and exhibit properties typical of that region of the conduction band. In particular, the effective mass for electrons in the higher L valleys is almost eight times as great as in the central valley, and the electron mobility is much lower. This is an important result for the negative conductivity mechanism: As the electric field is increased, the electron velocity increases until a critical field is reached; then the electrons *slow down* with further increase in field. The electron transfer process allows electrons to gain energy at the expense of velocity over a range of values of the electric field. Taking current density as qv_dn, it is clear that current also drops in this range of increasing field, giving rise to a negative differential conductivity $dJ/d\mathscr{E}$.

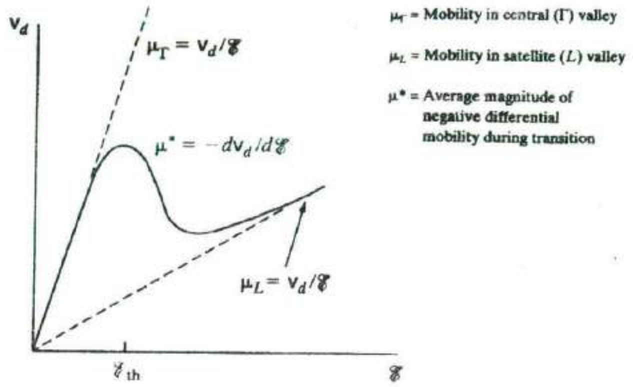


Figure 10–7 A possible characteristic of electron drift velocity vs. field for a semiconductor exhibiting the transferred electron mechanism.

A possible dependence of electron velocity vs. electric field for a material capable of electron transfer is shown in Fig. 10–7. For low values of field, the electrons reside in the lower (Γ) valley of the conduction band, and the mobility ($\mu_\Gamma = v_d/\mathscr{E}$) is high and constant with field. For high values of field, electrons transfer to the satellite valleys, where their velocity is smaller and their mobility lower. Between these two states is a region of negative slope on the v_d vs. $\mathscr{E}$ plot, indicating a negative differential mobility $dv_d/d\mathscr{E} = -\mu^*$.

The actual dependence of electron drift velocity on electric field for GaAs and InP is shown in Fig. 10–8. The negative resistance due to electron transfer occurs at a higher field for InP, and the electrons achieve a higher peak velocity before transfer from Γ to L occurs.

The existence of a drop in mobility with increasing electric field and the resultant possibility of negative conductance were predicted by Ridley and Watkins and by Hilsum several years before Gunn demonstrated the effect in GaAs. The mechanism of electron transfer is therefore often called the Ridley–Watkins–Hilsum mechanism. This negative conductivity effect depends only on the bulk properties of the semiconductor and not on junction or surface effects. It is therefore called a *bulk negative differential conductivity (BNDC)* effect.

10.3.2 Formation and Drift of Space Charge Domains

If a sample of GaAs is biased such that the field falls in the negative conductivity region, space charge instabilities result, and the device cannot be maintained in a d-c stable condition. To understand the formation of these instabilities, let us consider first the dissipation of space charge in the usual semiconductor. It can be shown from treatment of the continuity equation that a localized space charge dies out exponentially with time in a homogeneous sample with positive resistance (Prob. 10.3). If the initial space charge is Q_0, the instantaneous charge is

$$Q(t) = Q_0 e^{-t/\tau_d} \qquad (10\text{–}2)$$

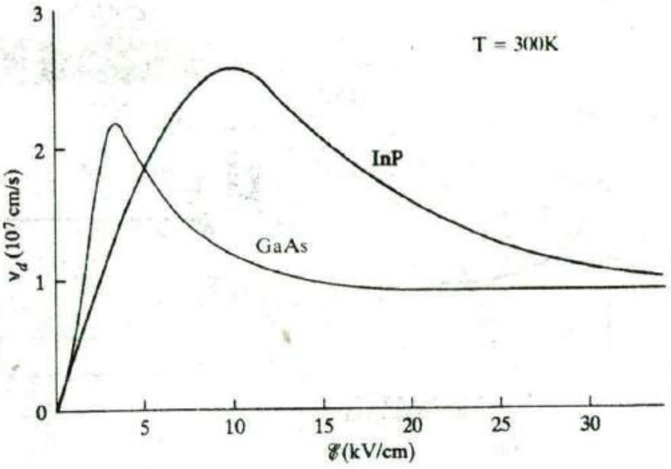


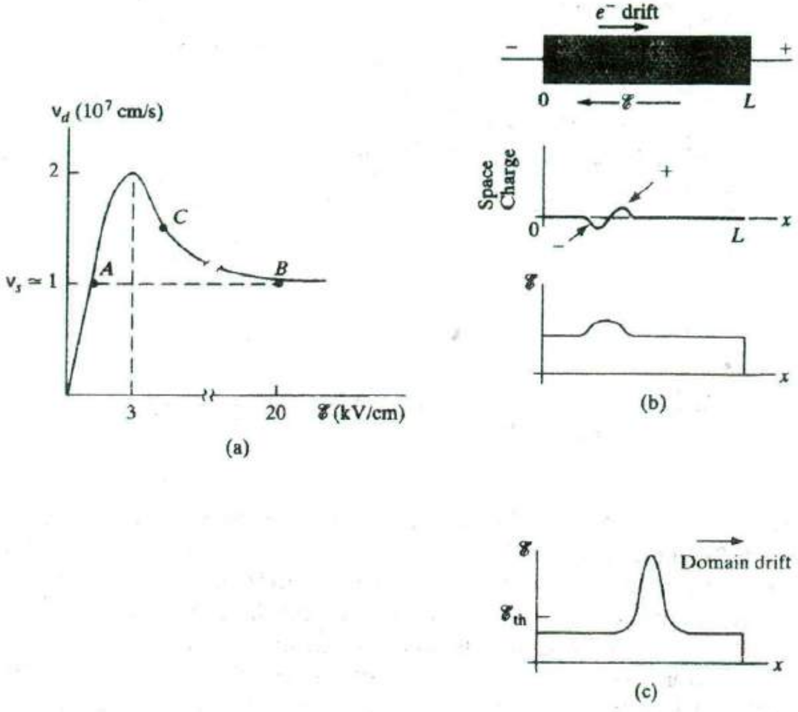
Figure 10–8 Electron drift velocity vs. field for GaAs and InP.

where $\tau_d = \epsilon/\sigma$ is called the *dielectric relaxation time.* Because of this process, random fluctuations in carrier concentration are quickly neutralized, and space charge neutrality is a good approximation for most semiconductors in the usual range of conductivities. For example, the dielectric relaxation time for a 1.0 Ω-cm Si or GaAs sample is approximately 10^{-12} s.

Equation (10–2) gives a rather remarkable result for cases in which the conductivity is negative. For these cases τ_d is negative also, and *space charge fluctuations build up* exponentially in time rather than dying out. This means that normal random fluctuations in the carrier distribution can grow into large space charge regions in the sample. Let us see how this occurs in a GaAs sample biased in the negative conductivity regime. The velocity–field diagram for n-type GaAs is illustrated in Fig. 10–9a. If we assume a small shift of electron concentration in some region of the device, a dipole layer can form as shown in Fig. 10–9b. Under normal conditions this dipole would die out quickly. However, under conditions of negative conductivity the charge within the dipole, and therefore the local electric field, builds up as shown in Fig. 10–9c. Of course, this buildup takes place in a stream of electrons drifting from the cathode to the anode, and the dipole (now called a *domain*) drifts along with the stream as it grows. Eventually the drifting domain will reach the anode, where it gives up its energy as a pulse of current in the external circuit.

During the initial growth of the domain, an increasing fraction of the applied voltage appears across it, at the expense of electric field in the rest of the bar. As a result, it is unlikely that more than one domain will be present in the bar at a time; after the formation of one domain, the electric field in the rest of the bar quickly drops below the threshold value for negative conductivity. If the bias is d-c, the field outside the moving domain will stabilize at a positive conductivity point such as A in Fig. 10–9a, and the field in the domain will stabilize at the high-field value B.

Figure 10–9 Buildup and drift of a space charge domain in GaAs: (a) velocity–field characteristic for n-type GaAs; (b) formation of a dipole; (c) growth and drift of a dipole for conditions of negative conductivity.



Let us follow the motion of a single domain as illustrated by Fig. 10–9. A small dipole forms from a random noise fluctuation (or more likely at a permanent *nucleation site* such as a crystal defect, a doping inhomogeneity, or the cathode itself), and this dipole grows and drifts down the bar as a domain. During the early stages of domain development, we can assume a uniform electric field in the bar, except just at the small dipole layer. If the field is in the negative mobility region, such as point C of Fig. 10–9a, the slightly higher field within the dipole results in a lower value of electron drift velocity inside the dipole than outside. As a result, electrons on the right (downstream) of the domain drift away, while electrons pile up on the left (upstream) side. This causes the accumulation and depletion layers of the dipole to grow, thereby further increasing the electric field in the domain. This is obviously a runaway process, in which the electric field within the domain grows while that outside the domain decreases. A stable condition is realized when the domain field increases to point B in Fig. 10–9a, and the field outside drops to point A. When this condition is met, the electrons drift at a constant velocity v_s everywhere, and the domain moves down the bar without further growth.

In this discussion we have assumed that the domain has time to grow to its stable condition before it drifts out of the bar. This is not always the case; for example, in a short bar with a low concentration of electrons, a dipole can drift the length of the bar before it develops into a domain. We can specify limits on the electron concentration n_0 and sample length L for successful domain formation by requiring the transit time (L/v_s) to be greater than the dielectric relaxation time (absolute value) in the negative mobility region. This requirement gives

$$\frac{L}{\mathrm{v}_s} > \frac{\epsilon}{q\mu^* n_0}$$

$$Ln_0 > \frac{\epsilon \mathrm{v}_s}{q\mu^*} \simeq 10^{12}\mathrm{cm}^{-2} \qquad (10\text{-}3)$$

for n-type GaAs, where the average negative differential mobility[4] is taken to be -100 cm^2/V-s. Therefore, for successful domain formation there is a critical product of electron concentration and sample length.

The type of domain motion we have described here was the first mode of operation observed by Gunn. In the observation of current vs. voltage for a GaAs sample, Gunn found a linear ohmic relation up to a critical bias, beyond which the current came in sharp pulses. The pulses were separated in time by an amount proportional to the sample length. This length dependence was due to the transit time L/v_s required for a domain nucleated at the cathode to drift the length of the bar. Gunn performed an interesting experiment in which he used a tiny capacitive probe to measure the electric field at various positions down the bar. By scanning the field distribution in the bar at various times in the cycle, he was able to plot out the growth and drift of the domains.

The formation of stable domains is not the only mode of operation for transferred electron devices. Nor is it the most desirable mode for most applications, since the resulting short pulses of current are inefficient sources of microwave power.

10.3.3 Fabrication

Devices utilizing the Gunn effect and its variations can be made in a number of materials which have appropriate band structures. Although GaAs and InP are the most common materials, transferred electron effects have been observed in CdTe, ZnSe, GaAsP, and other materials. The band structure of some materials can be altered to exhibit properties appropriate for electron transfer. For example, the energy bands of InAs can be distorted by the application of pressure to the crystal, such that a set of satellite valleys becomes available for electron transfer, although these upper valleys are too far above the lower valley at normal pressures. We have discussed the device behavior in terms of GaAs, since this material can be prepared with good purity and is most widely used in microwave applications.

[4] This is a rather crude approximation, since μ^* is not a constant but varies considerably with field; the negative dielectric relaxation time therefore changes with time as the domain grows.

Gunn diodes and related devices are simple structures in principle, since they are basically homogeneous samples with ohmic contacts on each end. In practice, however, considerable care must be taken in fabricating and mounting workable devices. In addition to the obvious requirements on doping, carrier mobility, and sample length, there are important problems with contacts, heat sinking, and parasitic reactances of the packaged device.

The samples must have high mobility, few lattice defects, and homogeneous doping in the range giving carrier concentrations $n_0 \simeq 10^{13} - 10^{16}\ \text{cm}^{-3}$. Devices can be made from GaAs or InP bulk samples cut from an ingot, but it is more common to use ingot material as a substrate for an epitaxial layer, which serves as the active region of the device. The material properties of epitaxial layers are often superior to bulk samples, and the precise control of layer thickness is helpful in these devices, which require exact sample lengths. In a typical configuration, an n-type epitaxial layer about 10 μm thick is grown on an n^+ substrate wafer, which is perhaps 100 μm thick. The substrate serves as one of the contacts to the active region. A thin n^{++} (very heavily doped) layer is grown on top of the n region, so that an n^+-n-n^{++} sandwiched structure results. External contacts can be made by evaporating a thin layer of Au–Sn or Au–Ge on each surface, followed by a brief alloying step in a hydrogen atmosphere. The wafer is divided into individual devices by cutting or cleaving (giving a cube structure) or by selective etching (giving a mesa structure). Each device is mounted with the n^{++} side down on a copper stud or other heat sink, so that the active region can dissipate heat to the mount in one direction and to the substrate layer in the other direction. Then the substrate side can be contacted by a wire or pressure contact. In other configurations, planar fabrication techniques can be used to produce lateral devices in an n-type epitaxial layer grown on a high-resistivity substrate.

Removal of heat is a very serious problem in these devices. The power dissipation may be $10^7\ \text{W/cm}^3$ or greater (Prob. 10.5), giving rise to considerable heating of the sample. As the temperature increases, the device characteristics vary because of changes in n_0 and mobility. As a result of such heating effects, these devices seldom reach their theoretical maximum efficiency. Pulsed operation allows better control of heat dissipation than does continuous operation, and efficiencies near the theoretical limits can sometimes be achieved in the pulsed mode. If the application does not require continuous operation, peak powers of hundreds of watts can be achieved in pulses of microwave oscillation.

PROBLEMS

10.1 Sketch the band diagram for an abrupt junction in which the doping on the p side is degenerate and the Fermi level on the n side is aligned with the bottom of the conduction band. Draw the forward and reverse bias band diagrams and sketch the I–V characteristic. This diode is often called a *backward diode*. Can you explain why?

10.2 What determines the peak tunneling voltage V_p of a tunnel diode? Explain.

If a large density of trapping centers is present in a tunnel diode (Fig. P10–2), tunneling can occur from the n-side conduction band to the trapping level (A–B). Then the electrons may drop to the valence band on the p side (B–C), thereby completing a two-step process of charge transport across the junction. In fact, if the density of trapping centers is large, it is possible to observe an increase in current as the states below E_{Fn} pass by the trapping level with increased bias. In Fig. P10–2, the trapping level E_t is located 0.3 eV above the valence band. Assume $E_g = 1$ eV, and $E_{Fn} - E_c$ on the n side equals $E_v - E_{Fp}$ on the p side, equals 0.1 eV.

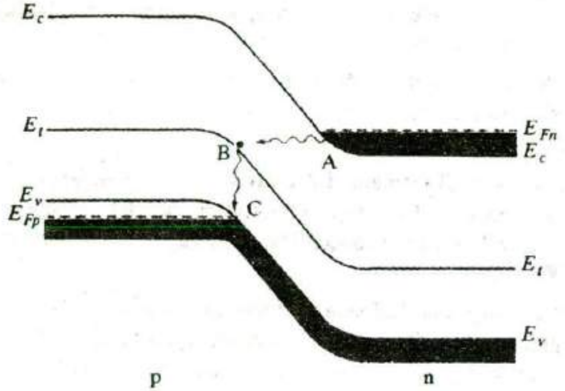


Figure P10–2

(a) Calculate the minimum forward bias at which tunneling through E_t occurs.

(b) Calculate the maximum forward bias for tunneling via E_t.

(c) Sketch the I–V curve for this tunnel diode. Assume the maximum tunneling current via E_t is about one-third of the peak band-to-band tunneling current.

10.3 (a) Use Poisson's equation, the continuity equation, and the definition of current density in terms of the gradient of electrostatic potential to relate the time variation of space charge density ρ to the conductivity σ and the permittivity ϵ of a material, neglecting recombination.

(b) Assuming a space charge density ρ_0 at $t = 0$, show that $\rho(t)$ decays exponentially with a time constant equal to the dielectric relaxation time τ_d.

(c) Given a sample of thickness L and area A, calculate the inherent RC time constant if the conductivity is σ and the permittivity is ϵ.

10.4 Assuming that n_Γ electrons/cm^3 are in the lower (central) valley of the GaAs conduction band at time t and n_L are in the satellite (L) valleys, show that the criterion for negative differential conductivity ($dJ/d\mathscr{E} < 0$) is

$$\frac{\mathscr{E}(\mu_\Gamma - \mu_L)\dfrac{dn_\Gamma}{d\mathscr{E}} + \mathscr{E}\left(n_\Gamma\dfrac{d\mu_\Gamma}{d\mathscr{E}} + n_L\dfrac{d\mu_L}{d\mathscr{E}}\right)}{n_\Gamma\mu_\Gamma + n_L\mu_L} < -1$$

where μ_Γ and μ_L are the electron mobilities in the Γ and L valleys, respectively. *Note:* $n_0 = n_\Gamma + n_L$. Discuss the conditions for negative differential conductivity, assuming the mobilities are approximately proportional to $\mathscr{E}^{-1}$.

10.5 We wish to estimate the d-c power dissipated in a GaAs Gunn diode. Assume the diode is 5 μm long and operates in the stable domain mode.

(a) What is the minimum electron concentration n_0? What is the time between current pulses?

(b) Using data from Fig. 10–9a, calculate the power dissipated in the sample per unit volume when it is biased just below threshold, if n_0 is chosen from the calculation of part (a). In general, does operation at a higher frequency result in greater power dissipation?

10.6 (a) Calculate the ratio N_L/N_Γ of the effective density of states in the upper (L) valleys to the effective density of states in the lower (Γ) valley of the GaAs conduction band (Fig. 10–6).

(b) Assuming a Boltzmann distribution $n_L/n_\Gamma = (N_L/N_\Gamma)\exp(-\Delta E/kT)$, calculate the ratio of the concentration of conduction band electrons in the upper valley to the concentration in the central valley in equilibrium at 300 K.

(c) As a rough calculation, assume an electron at the bottom of the central valley has kinetic energy kT. After it is promoted to the satellite (L) valley, what is its approximate equivalent temperature?

READING LIST

Bailey, M. J. "Heterojunction IMPATT Diodes: Using New Material Technology in a Classic Device." *Microwave Journal* 36 (June 1993): 76+.

Bayraktaroglu, B. "Monolithic IMPATT Technology." *Microwave Journal* 32 (April 1989); 73–4+.

Bose, B. K. "Recent Advances in Power Electronics." *IEEE Transactions on Power Electronics* 7 (January 1992): 2–16.

Esaki, L. "Discovery of the Tunnel Diode." *IEEE Trans. Elec. Dev.*, ED-23 (1976): 644+.

Gunn, J. B. "Microwave Oscillations of Current in III-V Semiconductors." *Solid State Comm.*, 1 (1963): 88+.

Herman, M. A. *Molecular Beam Epitaxy: Fundamentals and Current Status.* Berlin: Springer-Verlag, 1989.

Hughes, W. A., A. A. Rezazadeh, and C. E. C. Wood. *GaAs Integrated Circuits.* Oxford: BSP Professional Books, 1988.

Kearney, M. J., N. R. Couch, and J. Stephens. "Heterojunction Impact Avalanche Transit-time Diodes Grown by Molecular Beam Epitaxy." *Semiconductor Science and Technology* 8 (April 1993) 560–7.

Lesurf, J. "The Rise and Fall of Negative Resistance." *New Scientist* 31 (31 March 1990): 56–60.

Neamen, D. A. *Semiconductor Physics and Devices: Basic Principles.* Homewood, IL: Irwin, 1992.

Read, W. T. "A Proposed High Frequency, Negative Resistance Diode." *Bell Syst. Tech. J.*, 37 (1958): 401+.

Ridley, B. K., and T. B. Watkins. "The Possibility of Negative Resistance Effects in Semiconductors." *Proc. Phys. Soc. Lond.* 78 (161): 293+.

Shockley, W. "Negative Resistance Arising From Transit Time in Semiconductor Diodes." *Bell Syst. Tech. J.*, 33 (1954): 799+.

Shur, M. *GaAs Devices and Circuits.* New York: Plenum Press, 1987.

Singh, J. *Semiconductor Devices.* New York: McGraw-Hill, 1994.

Sze, S. M. *High-Speed Semiconductor Devices.* New York: Wiley, 1990.

Sze, S. M. *Physics of Semiconductor Devices.* New York: Wiley, 1981.

Voelcker, J. "The Gunn Effect." *IEEE Spectrum* 26 (July 1989): 24.

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Wood, J., and D. V. Morgan. "Gallium Arsenide and Related Compounds for Device Applications." *Acta Physica Polonica A* 79 (January 1991): 97–116.

Chapter 11
Power Devices

One of the most common applications of electronic devices is in switching, which requires the device to change from an "off" or *blocking* state to an "on" or *conducting* state. We have discussed the use of transistors in this application, in which base current drives the device from cutoff to saturation. Similarly, diodes and other devices can be used to serve as certain types of switches. There are a number of important switching applications that require a device remain in the blocking state under forward bias until switched to the conducting state by an external signal. Several devices which fulfill this requirement have been developed, and we shall discuss a family of switches in this chapter, the *semiconductor controlled rectifier (SCR)*[1] and related devices. These devices are typified by a high impedance ("off" condition) under forward bias until a switching signal is applied; after switching they exhibit low impedance ("on" condition). The signal required for switching can be varied externally; therefore, these devices can be used to block or pass currents at predetermined levels. In this chapter we shall discuss the physical operation of the SCR and a combination FET and SCR called an *insulated gate bipolar transistor.*

11.1 THE p-n-p-n DIODE

The SCR is a four-layer (p-n-p-n) structure that effectively blocks current through two terminals until it is turned on by a small signal at a third terminal. There are many varieties of the basic p-n-p-n structure, and we shall not attempt to cover all of them; however, we can discuss the basic operation and physical mechanisms involved in these devices. We shall begin by investigating the current flow in a two-terminal p-n-p-n device and then extend the discussion to include triggering by a third terminal. We shall see that the p-n-p-n structure can be considered for many purposes as a combination of p-n-p and n-p-n transistors, and the analysis in Chapter 7 can be used as an aid in understanding its behavior.

Before discussing the control of an SCR using a third terminal, it is important to understand the basic transistor action at work in a p-n-p-n structure. Therefore, in this section we analyze the four-layer structure with only two terminals.

[1] Since Si is the material commonly used for this device, it is often called a *silicon controlled rectifier.*

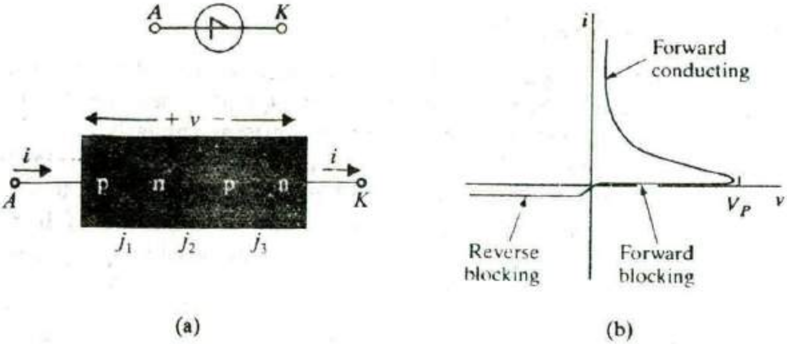


Figure 11–1
A two-terminal p-n-p-n device: (a) basic structure and common circuit symbol; (b) I–V characteristic.

11.1.1 Basic Structure

First we consider a four-layer diode structure with an *anode* terminal A at the outside p region with a *cathode* terminal K at the outside n region (Fig. 11–1a). We shall refer to the junction nearest the anode as j_1, the center junction as j_2, and the junction nearest the cathode as j_3. When the anode is biased positively with respect to the cathode (v positive), the device is forward biased. However, as the I–V characteristic of Fig. 11–b indicates, the forward-biased condition of this diode can be considered in two separate states, the high-impedance or *forward-blocking* state and the low-impedance or *forward-conducting* state. In the device illustrated here the forward I–V characteristic switches from the blocking to the conducting states at a critical peak forward voltage V_p.

We can anticipate the discussion of conduction mechanisms to follow by noting that an initial positive voltage v places j_1 and j_3 under forward bias and the center junction j_2 under reverse bias. As v is increased, most of the forward voltage in the blocking state must appear across the reverse-biased junction j_2. After switching to the conducting state, the voltage from A to K is very small (less than 1 V), and we conclude that in this condition all three junctions must be forward biased. The mechanism by which j_2 switches from reverse bias to forward bias is the subject of much of the discussion to follow.

In the *reverse-blocking* state (v negative), j_1 and j_3 are reverse biased and j_2 is forward biased. Since the supply of electrons and holes to j_2 is restricted by the reverse-biased junctions on either side, the device current is limited to a small saturation current arising from thermal generation of EHPs near j_1 and j_3. The current remains small in the reverse-blocking condition until avalanche breakdown occurs at a large reverse bias. In a properly designed device, with guards against surface breakdown, the reverse breakdown voltage can be several thousand volts.

We shall now consider the mechanism by which this device, often called a *Shockley diode*, switches from the forward-blocking state to the forward-conducting state.

11.1.2 The Two-Transistor Analogy

The four-layer configuration of Fig. 11–1a suggests that the p-n-p-n diode can be considered as two coupled transistors: j_1 and j_2 form the emitter and collector junctions, respectively, of a p-n-p transistor; similarly, j_2 and j_3 form the collector and emitter junctions of an n-p-n (note the emitter of the n-p-n is on the right, which is the reverse of what we usually draw). In this analogy, the collector region of the n-p-n is in common with the base of the p-n-p, and the base of the n-p-n serves as the collector region of the p-n-p. The center junction j_2 serves as the collector junction for both transistors.

This two-transistor analogy is illustrated in Fig. 11–2. The collector current i_{C1} of the p-n-p transistor drives the base of the n-p-n, and the base current i_{B1} of the p-n-p is dictated by the collector current i_{C2} of the n-p-n. If we associate an emitter-to-collector current transfer ratio α with each transistor, we can use the analysis in Chapter 7 to solve for the current i. Using Eq. (7–37b) with $\alpha_1 = \alpha_N$ for the p-n-p, $\alpha_2 = \alpha_N$ for the n-p-n, and with I_{CO1} and I_{CO2} for the respective collector saturation currents, we have

$$i_{C1} = \alpha_1 i + I_{C01} = i_{B2} \tag{11–1a}$$

$$i_{C2} = \alpha_2 i + I_{C02} = i_{B1} \tag{11–1b}$$

But the sum of i_{C1} and i_{C2} is the total current through the device:

$$i_{C1} + i_{C2} = i \tag{11–2}$$

Taking this sum in Eq. (11–1) we have

$$i(\alpha_1 + \alpha_2) + I_{C01} + I_{C02} = i$$

$$i = \frac{I_{C01} + I_{C02}}{1 - (\alpha_1 + \alpha_2)} \tag{11–3}$$

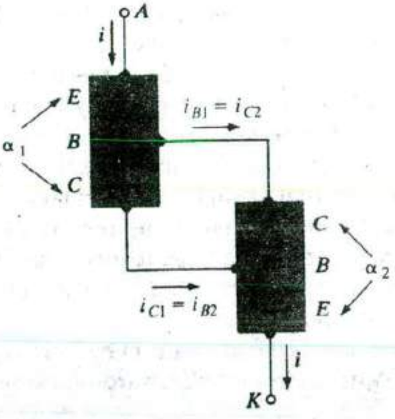


Figure 11–2 Two-transistor analogy of the p-n-p-n diode.

As Eq. (11–3) indicates, the current i through the devices is small (approximately the combined collector saturation currents of the two equivalent transistors) as long as the sum $\alpha_1 + \alpha_2$ is small compared with unity. As the sum of the alphas approaches unity, the current i increases rapidly. The current does not increase without limit as Eq. (11–3) implies, however, because the derivation is no longer valid as $\alpha_1 + \alpha_2$ approaches unity. Since j_2 becomes forward biased in the forward-conducting state, both transistors become saturated after switching. The two transistors remain in saturation while the device is in the forward-conducting state, being held in saturation by the device current.

11.1.3 Variation of α with Injection

Since the two-transistor analogy implies that switching involves an increase in the alphas to the point that $\alpha_1 + \alpha_2$ approaches unity, it may be helpful to review how alpha varies with injection for a transistor. The emitter-to-collector current transfer ratio α is given in Section 7.2 as the product of the emitter injection efficiency γ and the base transport factor B. An increase in α with injection can be caused by increases in either of these factors, or both. At very low currents (such as in the forward-blocking state of the p-n-p-n diodes), γ is usually dominated by recombination in the transition region of the emitter junction (Section 7.7.4). As the current is increased, injection across the junction begins to dominate over recombination within the transition region (Section 5.6.2) and γ increases. There are several mechanisms by which the base transport factor B increases with injection, including the saturation of recombination centers as the excess carrier concentration becomes large. Whichever mechanism dominates, the increase in $\alpha_1 + \alpha_2$ required for switching of the p-n-p-n diode is automatically accomplished. In general, no special design is required to maintain $\alpha_1 + \alpha_2$ smaller than unity during the forward-blocking state; this requirement is usually met at low currents by the dominance of recombination within the transition regions of j_1 and j_3.

11.1.4 Forward-Blocking State

When the device is biased in the forward-blocking state (Fig. 11–3a), the applied voltage v appears primarily across the reverse-biased junction j_2. Although j_1 and j_3 are forward biased, the current is small. The reason for this becomes clear if we consider the supply of electrons available to n_1 and holes to p_2. Focusing attention first upon j_1, let us assume a hole is injected from p_1 into n_1. If the hole recombines with an electron in n_1 (or in the j_1 transition region), that electron must be resupplied to the n_1 region to maintain space charge neutrality. The supply of electrons in this case is severely restricted, however, by the fact that n_1 is terminated in j_2, a reverse-biased junction. In a normal p-n diode the n region is terminated in an ohmic contact, so that the supply of electrons required to match recombination (and injection into p)

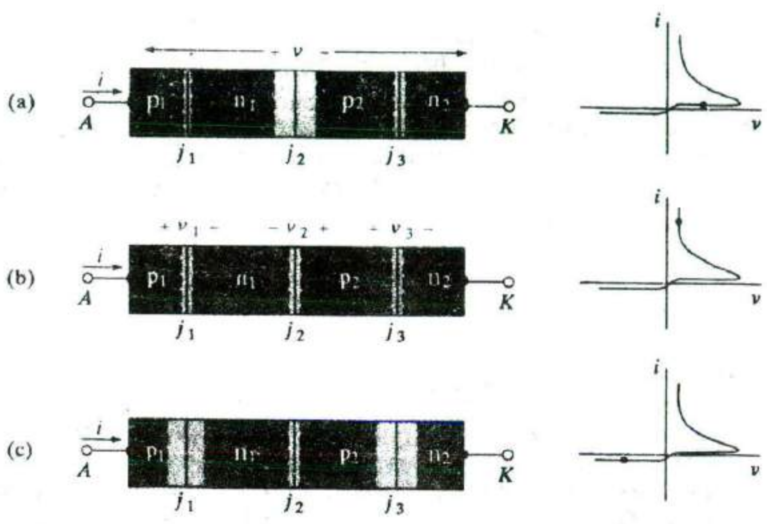


Figure 11–3 Three bias states of the p-n-p-n diode: (a) the forward-blocking state; (b) the forward-conducting state; (c) the reverse-blocking state.

is unlimited. In this case, however, the electron supply is restricted essentially to those electrons generated thermally within a diffusion length of j_2. As a result, the current passing through the j_1 junction is approximately the same as the reverse saturation current of j_2. A similar argument holds for the current through j_3; holes required for injection into n_2 and to feed recombination in p_2 must originate in the saturation current of the center junction j_2. The applied voltage v divides appropriately among the three junctions to accommodate this small current throughout the device.

In this discussion we have tacitly assumed that the current crossing j_2 is strictly the thermally generated saturation current. This implies that electrons injected by the forward-biased junction j_3 do not diffuse across p_2 in any substantial numbers, to be swept across the reverse-biased junction into n_1 by transistor action. This is another way of saying that α_2 (for the "n-p-n transistor") is small. Similarly, the supply of holes to p_2 is primarily thermally generated, since few holes injected at j_1 reach j_2 without recombination (i.e., α_1 is small for the "p-n-p"). Now we can see physically why Eq. (11–3) implies a small current while $\alpha_1 + \alpha_2$ is small: Without the transport of charge provided by transistor action, the thermal generation of carriers is the only significant source of electrons to n_1 and holes to p_2.

11.1.5 Conducting State

The charge transport mechanism changes dramatically when transistor action begins. As $\alpha_1 + \alpha_2$ approaches unity by one of the mechanisms described above, many holes injected at j_1 survive to be swept across j_2 into p_2. This helps to feed the recombination in p_2 and to support the injection of holes into

n_2. Similarly, the transistor action of electrons injected at j_3 and collected at j_2 supplies electrons for n_1. Obviously, the current through the device can be much larger once this mechanism begins. The transfer of injected carriers across j_2 is regenerative, in that a greater supply of electrons to n_1 allows greater injection of holes at j_1 while maintaining space charge neutrality; this greater injection of holes further feeds p_2 by transistor action, and the process continues to repeat itself.

If $\alpha_1 + \alpha_2$ is large enough, so that many electrons are collected in n_1 and many holes are collected in p_2, the depletion region at j_2 begins to decrease. Finally the reverse bias disappears across j_2 and is replaced by a forward bias, in analogy with a transistor biased deep in saturation. When this occurs, the three small forward-bias voltages appear as shown in Fig. 11–3b. Two of these voltages essentially cancel in the overall v, so that the forward voltage drop of the device from anode to cathode in the conducting state is not much greater than that of a single p-n junction. For Si this forward drop is less than 1 V, until ohmic losses become important at high current levels.

We have discussed the current transport mechanisms in the forward-blocking and forward-conducting states, but we have not indicated how switching is initiated from one state to the other. Basically, the requirement is that the carrier injection at j_1 and j_2 must somehow be increased so that significant transport of injected carriers across j_2 occurs. Once this transport begins, the regenerative nature of the process takes over and switching is completed.

11.1.6 Triggering Mechanisms

There are several methods by which a p-n-p-n diode can be switched (or *triggered*) from the forward-blocking state to the forward-conducting state. For example, an increase in the device temperature can cause triggering, by sufficiently increasing the carrier generation rate and the carrier lifetimes. These effects cause a corresponding increase in device current and in the alphas discussed above. Similarly, optical excitation can be used to trigger a device by increasing the current through EHP generation.[2] The most common method of triggering a two-terminal p-n-p-n, however, is simply to raise the bias voltage to the peak value V_p. This type of *voltage triggering* results in a breakdown (or significant leakage) of the reverse-biased junction j_2; the accompanying increase in current provides the injection at j_1 and j_3 and transport required for switching to the conducting state. The breakdown mechanism commonly occurs by combination of *base-width narrowing* and *avalanche multiplication*.

When carrier multiplication occurs in j_2, many electrons are swept into n_1 and holes into p_2. This process provides the majority carriers to these regions needed for increased injection by the emitter junctions. Because of transistor action, the full breakdown voltage of j_2 need not be reached. As we

[2] Four-layer devices that can be triggered by a pulse of light are useful in many optoelectronic systems. This type of device is often called a *light-activated SCR*, or *LASCR*.

showed in Eq. (7-52), breakdown occurs in the collector junction of a transistor with $i_B = 0$ when $M\alpha = 1$. In the coupled transistor case of the p-n-p-n diode, breakdown occurs at j_2 when

$$M_p\alpha_1 + M_n\alpha_2 = 1 \tag{11-4}$$

where M_p is the hole multiplication factor and M_n is the multiplication factor for electrons.

As the bias v increases in the forward-blocking state, the depletion region about j_2 spreads to accommodate the increased reverse bias on the center junction. This spreading means that the neutral base regions on either side (n_1 and p_2) become thinner. Since α_1 and α_2 increase as these base widths decrease, triggering can occur by the effect of base-width narrowing. A true punch-through of the base regions is seldom required, since moderate narrowing of these regions can increase the alphas enough to cause switching. Furthermore, switching may be the result of a combination of avalanche multiplication and base-width narrowing, along with possible leakage current through j_2 at high voltage. From Eq. (11–4) it is clear that with avalanche multiplication present, the sum $\alpha_1 + \alpha_2$ need not approach unity to initiate breakdown of j_2. Once breakdown begins, the increase of carriers in n_1 and p_2 drives the device to the forward-conducting state by the regenerative process of coupled transistor action. As switching proceeds, the reverse bias is lost across j_2 and the junction breakdown mechanisms are no longer active. Therefore, base narrowing and avalanche multiplication serve only to start the switching process.

If a forward-bias voltage is applied rapidly to the device, switching can occur by a mechanism commonly called *dv/dt triggering*. Basically, this type of triggering occurs as the depletion region of j_2 adjusts to accommodate the increasing voltage. As the depletion width of j_2 increases, electrons are removed from the n_1 side and holes are removed from the p_2 side of the junction. For a slow increase in voltage, the resulting flow of electrons toward j_1 and holes toward j_3 does not constitute a significant current. If *dv/dt* is large, however, the rate of charge removal from each side of j_2 can cause the current to increase significantly. In terms of the junction capacitance (C_{j2}) of the reverse-biased junction, the transient current is given by

$$i(t) = \frac{dC_{j2}v_{j2}}{dt} = C_{j2}\frac{dv_{j2}}{dt} + v_{j2}\frac{dC_{j2}}{dt} \tag{11-5}$$

where v_{j2} is the instantaneous voltage across j_2. This type of current flow is often called *displacement current*. The rate of change of C_{j2} must be included in calculating current, since the capacitance varies with time as the depletion width changes.

The increase in current due to a rapid rise in voltage can cause switching well below the steady state triggering voltage V_P. Therefore, a *dv/dt* rating is usually specified along with V_P for p-n-p-n diodes. Obviously, *dv/dt* triggering can be a disadvantage in circuits subjected to unpredictable voltage transients.

The various triggering mechanisms discussed in this section apply to the two-terminal p-n-p-n diode. As we shall see in the following section, the semiconductor controlled rectifier is triggered by an external signal applied to a third terminal.

11.2 THE SEMICONDUCTOR CONTROLLED RECTIFIER

The semiconductor controlled rectifier (SCR) is useful in many applications, such as in power switching and in various control circuits. This device can handle currents from a few milliamperes to hundreds of amperes. Since it can be turned on externally, the SCR can be used to regulate the amount of power delivered to a load simply by passing current only during selected portions of the line cycle. A common example of this application is the light-dimmer switch used in many homes. At a given setting of this switch, an SCR is turned on and off repetitively, such that all or only part of each power cycle is delivered to the lights. As a result, the light intensity can be varied continuously from full intensity to dark. The same control principle can be applied to motors, heaters, and many other systems. We shall discuss this type of application in this section, after first establishing the fundamentals of device operation.

11.2.1 Gate Control

The most important four-layer device in power circuit applications is the three-terminal SCR[3] (Fig. 11–4). This device is similar to the p-n-p-n diode,

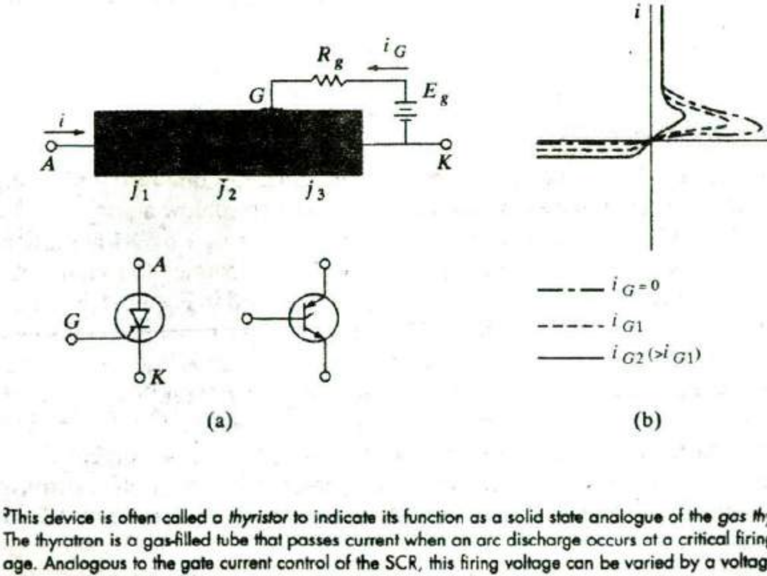


Figure 11–4 A semiconductor controlled rectifier: (a) four-layer geometry and common circuit symbols; (b) *I–V* characteristics.

[3] This device is often called a *thyristor* to indicate its function as a solid state analogue of the gas *thyratron*. The thyratron is a gas-filled tube that passes current when an arc discharge occurs at a critical firing voltage. Analogous to the gate current control of the SCR, this firing voltage can be varied by a voltage applied to a third electrode.

except that a third lead (*gate*) is attached to one of the base regions. When the SCR is biased in the forward-blocking state, a small current supplied to the gate can initiate switching to the conducting state. As a result, the anode switching voltage V_P decreases as the current i_G applied to the gate is increased (Fig. 11–4b). This type of turn-on control makes the SCR a useful and versatile device in switching and control circuits.

To visualize the *gate triggering* mechanism, let us assume the device is in the forward-blocking state, with a small saturation current flowing from anode to cathode. A positive gate current causes holes to flow from the gate into p_2, the base of the n-p-n transistor. This added supply of holes and the accompanying injection of electrons from n_2 into p_2 initiates transistor action in the n-p-n. After a transit time τ_{t2}, the electrons injected by j_3 arrive at the center junction and are swept into n_1, the base of the p-n-p. This causes an increase of hole injection of j_1, and these holes diffuse across the base n_1 in a transit time τ_{t1}. Thus, after a delay time of approximately $\tau_{t1} + \tau_{t2}$, transistor action is established across the entire p-n-p-n and the device is driven into the forward-conducting state. In most SCRs the delay time is less than a few microseconds, and the required gate current for turn-on is only a few milliamperes. Therefore, the SCR can be turned on by a very small amount of power in the gate circuit. On the other hand, the device current i can be many amperes, and the power controlled by the device may be very large.

It is not necessary to maintain the gate current once the SCR switches to the conducting state; in fact, the gate essentially loses control of the device after regenerative transistor action is initiated. For most devices a gate current pulse lasting a few microseconds is sufficient to ensure switching. Ratings of minimum gate pulse height and duration are generally provided for particular SCR devices.

11.2.2 Turning off the SCR

Turning off the SCR, changing it from the conducting state to the blocking state, can be accomplished by reducing the current i below a critical value (called the *holding current*) required to maintain the $\alpha_1 + \alpha_2 = 1$ condition. In some SCR devices, gate turn-off can be used to reduce the alpha sum below unity. For example, if the gate voltage is reversed in Fig. 11–4, holes are extracted from the p_2 base region. If the rate of hole extraction by the gate is sufficient to remove the n-p-n transistor from saturation, the device turns off. However, there are often problems involving the lateral flow of current in p_2 to the gate; nonuniform biasing of j_3 can result from the fact that the bias on this emitter junction varies with position when a lateral current flows. Therefore, SCR devices must be specifically designed for turn-off control; at best, this turn-off capability can be utilized only over a limited range for a given device.

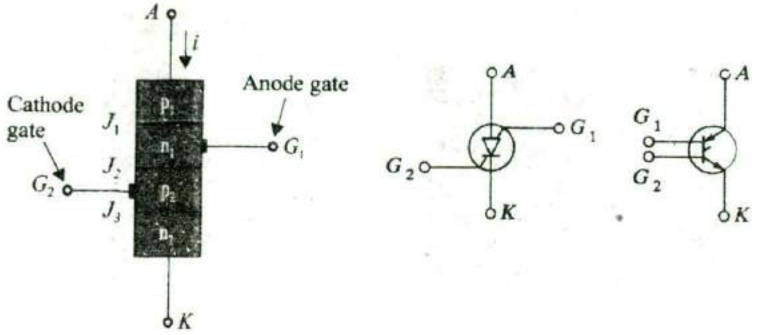


Figure 11–5 Semiconductor controlled switch: schematic configuration and common circuit symbols.

Some four-layer devices have two gate leads, one attached to n_1 and the other to p_2 (Fig. 11–5). This type of device is often called a *semiconductor controlled switch (SCS)*. The availability of the second gate electrode provides additional flexibility in circuit design. The SCS biased in the forward-blocking state can be switched to the conducting state by a positive current pulse applied to the *cathode gate* (at p_2) or by a negative pulse at the *anode gate* (n_1). If the device is designed for turn-off capability, separate circuits can be employed for turn-on at one gate and turn-off at the other. Other advantages of the SCS configuration include the possibility of minimizing unwanted *dv/dt* switching; for example, the gate not used for triggering can be capacitively coupled to the nearest current terminal (G_1 to A or G_2 to K) to allow a charging path for j_2 during a voltage transient without causing inadvertent switching.

11.2.3 Bilateral Devices

In many applications it is useful to employ devices which switch symmetrically with forward and reverse bias. This type of *bilateral* device is particularly useful in a-c circuits in which sinusoidal signals are switched on and off during positive and negative portions of the cycle. A typical bilateral p-n-p-n diode configuration is shown in Fig. 11–6a. This device differs from the p-n-p-n diode of Fig. 11–1 in that the n_2 region extends over only half the width of the cathode, and a new region n_3 is diffused into half of the anode region. In effect, this *bilateral diode switch* consists of two separate p-n-p-n diodes: the p_1-n_1-p_2-n_2 section and the p_2-n_1-p_1-n_3 section. We notice that the device shown in Fig. 11–6a is symmetrical. With the anode A biased positively with respect to the cathode K, junction j_1 is forward biased, while j_2 is reverse biased. Junction j_3 is shorted at one end (as is j_4) by the metal contact. When j_2 is biased to breakdown, however, a lateral current flows in p_2 biasing the left edge of j_3 into injection, and the device switches. This *shorted-emitter* design is commonly used in SCRs to enhance triggering control. During this operation, junction j_4 remains dormant. Because the device is symmetrical, j_4 serves

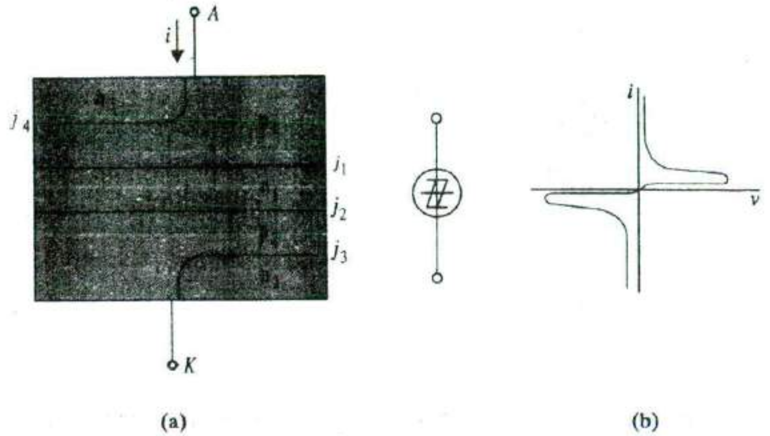


Figure 11–6 A bilateral diode switch: (a) schematic of the device configuration and a common circuit symbol; (b) typical I–V characteristic.

the shorted emitter function when the polarity is reversed (K positive with respect to A), and j_1 is the junction which is biased to break down in initiating the switching operation. If the bilateral diode is constructed properly, the forward and reverse characteristics are symmetrical as shown in Fig. 11–6b.

Bilateral triode switches (sometimes called *triacs*) can be constructed with SCR characteristics that can be triggered in either the forward- or reverse-bias mode. A good discussion of these devices is presented in the book by Gentry et al. in the reading list for this chapter.

11.2.4 Fabrication and Applications

Many variations of diffusion, implantation, and epitaxial growth are used in the fabrication of p-n-p-n devices. The type of fabrication process depends largely on the power rating and intended use of the device. For high-current devices the anode is attached to a heavy copper stud, and the cathode is contacted by a large cable. In high-current operation, heat is carried away from the junction by the massive metal substrate. The entire device is hermetically sealed in a housing, which provides protection from the atmosphere and from thermal and mechanical shock. Devices with this type of mounting can be rated at several hundred amperes in the conducting state. Of course, SCR devices intended for small-signal applications can be made in simpler and smaller packages.

Applications of SCRs and other four-layer devices are quite varied and extend into many fields of electronics, switching, and control. As a simple example, let us consider the problem of delivering variable power to a load from a constant line source (Fig. 11–7). The load may be the heater windings of a furnace, a light bulb, or another circuit. The amount of power delivered to the load during each half-cycle depends on the switching of the SCR. If pulses are delivered to the gate near the beginning of each half-cycle, essen-

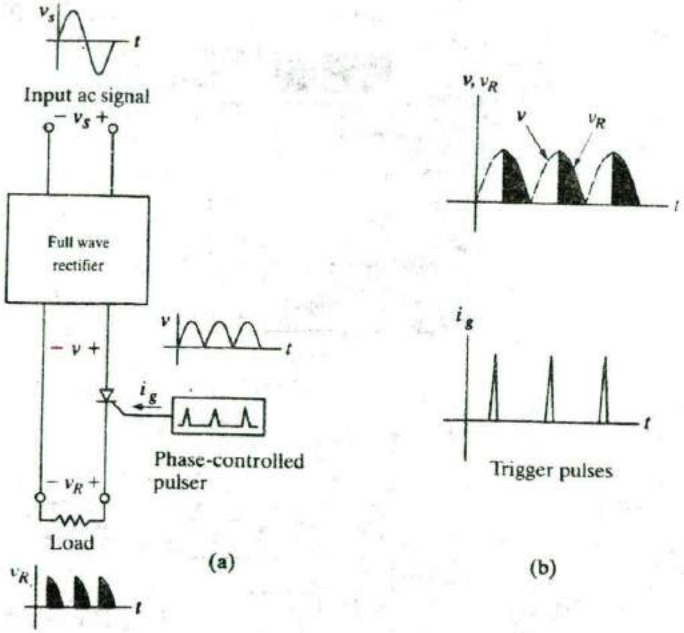


Figure 11–7 Example of the use of an SCR to control the power delivered to a load: (a) schematic diagram of the circuit; (b) waveforms of the delivered signal and the phase-variable trigger pulse.

tially the full power of the input is delivered to the load. On the other hand, if the trigger pulses are delayed, the SCR does not turn on until later in the half cycle. As a result, the amount of power delivered to the load can be varied from almost full power to no power.

Many examples of SCR and other four-layer device applications can be found in the reading list of this chapter and in the current literature.

11.3 INSULATED GATE BIPOLAR TRANSISTOR

We saw in Section 11.2 that the SCR has difficulty in efficiently turning off the device using the gate. We need to use additional circuitry to reduce the anode-to-cathode current below the holding current to change the SCR from the conducting state to the blocking state. This is, of course, clumsy and expensive.

Hence, the *insulated-gate bipolar transistor* (*IGBT*) was invented by Baliga in 1979 to address this issue. This variation on the SCR can easily be turned off from the conducting to the blocking state by the action of the gate. This device is also known by several other names such as conductivity-modulated FET (COMFET), insulated gate transistor (IGT), insulated gate rectifier (IGR), gain-enhanced MOSFET (GEMFET) and bipolar FET (BiFET).

The basic structure is shown for an n-channel device in Fig. 11–8. It basically combines an SCR with a MOSFET able to connect or disconnect the

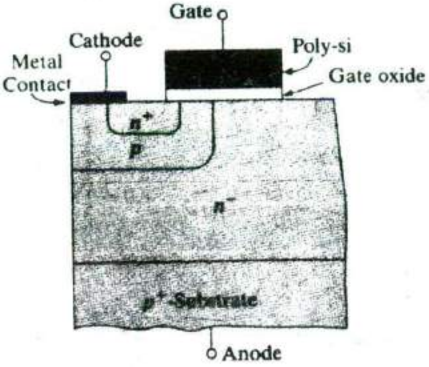


Figure 11–8 Structure of an insulated gate bipolar transistor.

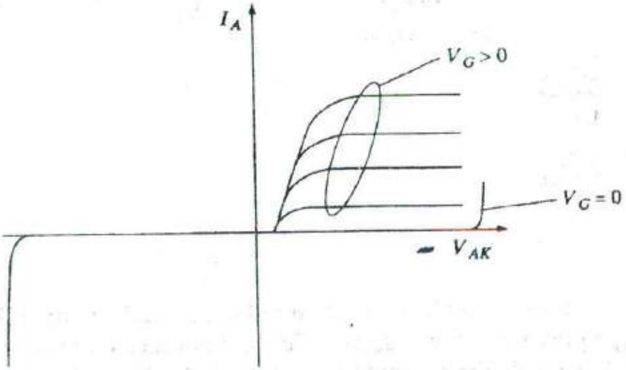


Figure 11–9 Output current–voltage characteristics of an insulated gate bipolar transistor (n-channel).

n^+ cathode to the n^- base region, depending on the gate bias of the MOSFET. The MOSFET channel length is determined by the p-region, which is formed by diffusion of the acceptors implanted in the same region as the n^+ cathode. In other words, the channel length is not determined by lithography of the gate, as in a conventional MOSFET, but rather by the diffusion of the acceptors. Such a MOSFET structure is known as a double-diffused MOSFET (DMOS). The DMOS device is essentially an NMOSFET.

The main part of the IGBT is the n^- region, which acts as the drain of the DMOS device. This is generally a thick (~50 μm) epitaxial region with a low doping (~10^{14} cm^{-3}) grown on a heavily p^+-doped substrate which forms the anode. This n^- region can therefore support a large blocking voltage in the OFF state. In the ON state, the conductivity of this lightly doped region is modulated (increased) by the electrons injected from the n^+ cathode and the holes injected from the p^+ anode; hence, the alternative name, conductivity modulated FET (COMFET). This increased conductivity allows the voltage drop across the device to be minimal in the ON state.

The current–voltage characteristics are shown in Fig. 11–9. If the DMOS gate voltage is zero (or below the threshold voltage), an n-type inversion region is not formed in the p-type channel region and the n^+ cathode is not shorted to the n^- base. The structure then looks exactly like a conventional SCR which allows minimum current flow in either polarity until breakdown is reached. For positive anode-to-cathode bias V_{AK}, avalanche breakdown occurs at the n^--p junction, while for negative V_{AK} avalanche occurs at the n^--p^+ junction.

When a gate bias is applied to the DMOS gate, for positive V_{AK}, we see that there is significant current flow (Fig. 11–9). The characteristics look like that of a MOSFET, with one difference. Instead of the current starting to increase from the origin, there is an offset or cut-in voltage of ~0.7V, just like for a diode. The reason for this can be understood by looking at the equivalent circuit in Fig. 11–10a. For small V_{AK} up to the offset voltage, the structure looks like a DMOS in series with a p-i-n diode made up of the p^+ substrate (anode), the n^- blocking region (base) which is essentially like an intrinsic region, and the n^+ cathode. In this regime, there is negligible voltage drop across the DMOS device, and the p-i-n device is in forward bias. The injected carriers from the anode and the cathode recombine in the n^- region. As we saw in Chapter 5, for a diode dominated by recombination in the depletion region, the current–voltage characteristics show an exponential behavior, with a diode ideality factor of $\mathbf{n} = 2$. Therefore, we get in this region

$$I_A \propto \exp(qV_{AK}/2\,kT) \qquad (11\text{–}6)$$

On the other hand when V_{AK} is larger than the offset voltage (~0.7 V), the characteristics look like that of a MOSFET, multiplied by a p-n-p bipolar junction transistor gain term. The equivalent circuit in this region is shown in Fig. 11–10b. In this regime, not all the injected carriers recombine in the near-intrinsic n^- region. This current, which is essentially the DMOSFET current, I_{MOS}, acts as the base current of the vertical p-n-p BJT formed between the p^+ substrate (anode), the n^- base and the p^- channel of the DMOS device. Hence, the current now is given by

$$I_A = (1 + \beta_{\text{pnp}})I_{MOS} \qquad (11\text{–}7)$$

The shape of the characteristics looks like that of the DMOS device. This is the preferred mode of operation of the IGBT.

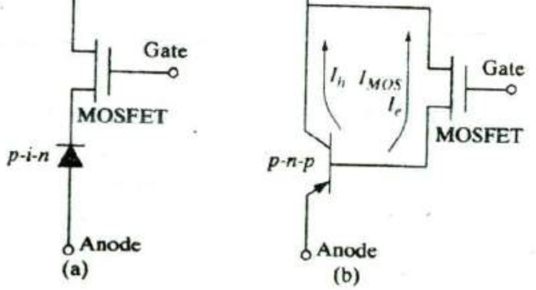


Figure 11–10 IGBT equivalent circuit: (a) below the offset voltage, for low V_{AK}; (b) above the offset voltage, for high V_{AK}.

Finally, if the current levels are too high, the IGBT latches into a low impedance state like that of a conventional SCR in the ON state. This is undesirable because it means that the gate of the DMOS device has now lost control.

The IGBT clearly incorporates some of the best features of MOSFETs and BJTs. Like a MOSFET, it has high input impedance and low input capacitance. On the other hand, in the ON state, it has low resistance and high current handling capability, like a BJT or SCR. Because of these factors, and because it can turn off more easily than a SCR, the IGBT is gradually becoming the power device of choice, in place of the more traditional SCR.

PROBLEMS

11.1 Explain why two separate transistors cannot be connected as in Fig. 11–2 to achieve the p-n-p-n switching action of Fig. 11–1.

11.2 In the p-n-p-n diode (Fig. 11–3a), the junction j_3 is forward biased during the forward-blocking state. Why, then, does the forward bias provided by the gate-to-cathode voltage in Fig. 11–4 cause switching?

11.3 (a) Sketch the energy band diagrams for the p-n-p-n diode in equilibrium; in the forward-blocking state; and in the forward-conducting state.

(b) Sketch the excess minority carrier distributions in regions n_1 and p_2 when the p-n-p-n diode is in the forward-conducting state.

11.4 Use schematic techniques such as those illustrated in Fig. 7–3 to describe the hole flow and electron flow in a p-n-p-n diode for the forward-blocking state and for the forward-conducting state. Explain the diagrams and be careful to define any new symbols (e.g., those representing EHP generation and recombination).

11.5 Using the coupled transistor model, rewrite Eqs. (11–1) to include avalanche multiplication in j_2, and show that Eq. (11–4) is valid for the p-n-p-n diode.

READING LIST

Baliga, B. J. *Power Semiconductor Devices.* Boston, MA: PWS, 1996.

Gentry, F. E., F. W. Gutzwiller, N., Holonyak, Jr., and E. E. Von Zastrow. *Semiconductor Controlled Rectifiers: Principles and Application of p-n-p-n Devices.* Englewood Cliffs, NJ: Prentice Hall, 1964.

Jaeklin, A. A., ed. *Power Semiconductor Devices and Circuits.* New York: Plenum Press, 1992.

Moll, J. L., M. Tanenbaum, J. M. Goldey and N. Holonyak. "P-N-P-N Transistor Switches." *Proc. IRE*, 44 (1956), 1174+.

Taylor, P. D. *Thyristor Design and Realization.* Chichester: Wiley, 1992.

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.