

Bipolar Transistors

Michael Shur
Rensselaer Polytechnic Institute

I. Principle of Operation
II. BJT Circuit Configurations and Current Gains
III. BJT Modes of Operation and Current–Voltage Characteristics
IV. BJTs, FETs, and BiCMOS
V. Small-Signal Parameters
VI. Heterojunction Bipolar Transistors

GLOSSARY

Band diagram Plot of potential energies of electrons and holes versus distance.
Bipolar junction transistor Semiconductor device with three contacts to regions of different conductivity types: emitter, base, and collector.
Carrier injection Carrier transport from one region of semiconductor to another causing an excess concentration in that region.
Current gain Ratio of output current over input current.
Heterojunction bipolar transistor Bipolar transistor with wide band-gap emitter.
***p–n* junction** Contact between *n*-type semiconductor and *p*-type semiconductor, usually (but not always) of same chemical composition.
Transistor Three-terminal semiconductor device where the input signal applied to the two of terminals controls the output signal

A BIPOLAR junction transistor is a semiconductor device with three contacts to the three regions of the device of different conductivity types: emitter, base, and collector. The emitter and collector have one conductivity type (these regions are usually doped by donors; i.e., they are *n*-type), and the base has the opposite conductivity type (the base is usually doped by acceptors; i.e., the base is *p*-type). The charge carriers are injected from the emitter into the base. They diffuse across the base and reach the collector region. This injection, and hence the collector current, are strong functions of the base current, so that a small change in the base current causes a much larger change in the collector current. Therefore, the bipolar junction transistor has a large current gain, which is defined as the ratio of the collector current over the base current.

I. PRINCIPLE OF OPERATION

A bipolar junction transistor (BJT) is a three-terminal semiconductor device. An input signal is fed to the input terminals, and a load is connected to output terminals, as shown in Fig. 1. Transistors are designed in such a way that a small change in the input transistor current or voltage can result in a large change in the output current or

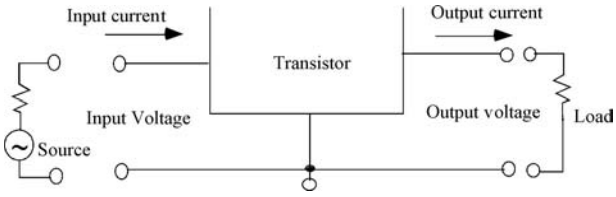


FIGURE 1 Transistor input and output.

voltage. Thus, a transistor may provide a voltage and/or current *gain* and a power gain. A change in the input signal can also switch the output current between a high value and a low value. This property is used in all digital electronic circuits for switching the output current on and off by a small input signal.

Brattain, Bardeen, and Shockley of Bell Laboratories invented the BJT in 1947. The word "transistor" was coined by combining the two words "transfer" and "resistor". Figure 2 shows the famous photograph of the first transistor, which started the modern age, the age of microelectronics.

For this invention, Brattain, Bardeen, and Shockley received a Nobel Prize in 1956.

A BJT is composed of two back-to-back p–n junctions. These devices can be made in two polarities, as shown in Fig. 3, and are called n–p–n and p–n–p transistors. Figure 4 shows the circuit symbols for n–p–n and p–n–p BJTs.

In order to understand how a BJT works, let us first review the principles of operation of a p–n junction. A p–n junction is a contact between an n-type semiconductor and a p-type semiconductor usually (but not always) of the same chemical composition (a vast majority of semiconductor devices are made of silicon). In n-type and p-type semiconductors, mobile charge carriers are electrons and holes, respectively. If p-type and n-type semi-

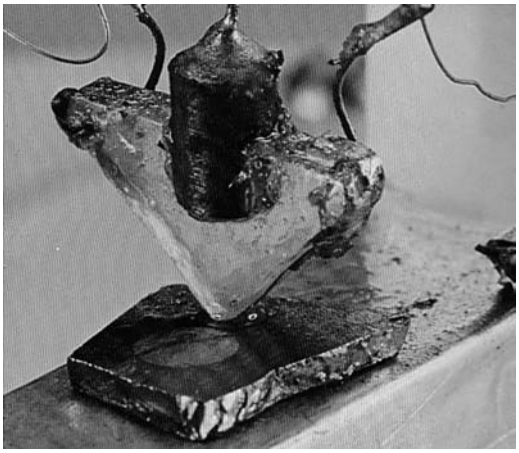


FIGURE 2 The first bipolar transistor, invented by Brattain, Bardeen, and Shockley.

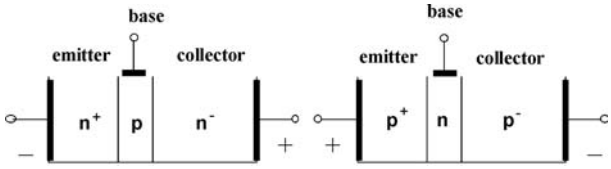


FIGURE 3 Schematic diagrams of n–p–n and p–n–p BJTs. Signs indicate voltage polarities with respect to the base.

conductors are brought together and form an ideal p–n junction, the holes would diffuse from the p-type region, where their concentration is high, into the n-type region, where their concentration is low. The deficit of positively charged holes creates a layer of negatively charged *acceptors* in the p-region close to the junction. In a similar way, electrons diffuse from the n-type region, where their concentration is high, into the p-type region, where their concentration is low. The deficit of negatively charged electrons creates a layer of positively charged *donors* in the n-type region near the junction. Figure 5 shows the distribution of charges in a p–n junction and a p–n junction band diagram, which shows potential energies of electrons (E_c) and holes (E_v) versus distance. The positive and negative charge densities in Fig. 6 correspond to ionized donors and ionized acceptors, respectively.

The charged region, nearly devoid of holes in the p-region and nearly devoid of electrons in the n-region, is called a *space charge region* or *depletion region.* The charges in the depletion region create a potential barrier that prevents more electrons from coming into the p-region and more holes from coming into the n-region. As a result, the net flux of electrons and holes is zero under thermal equilibrium conditions. The *built-in voltage*, V_{bi}, is typically of the order of E_g/q, where E_g is the *energy gap* and q is the electronic charge.

The potential barrier at the p–n junction can be increased or decreased by applying an external voltage to the junction. A positive potential, V (with respect to the n-region), applied to the p-region (forward bias), attracts negatively charged electrons from the n-region and pushes positively charged holes into the n-region, reducing the potential barrier to $V_{bi} - V$. A negative potential applied to the p-region (reverse bias) pushes electrons away from the p-region and pushes holes away from the n-region. Such a potential increases the potential barrier to $V_{bi} + |V|$. As a result, the diode resistance under the forward and reverse bias will be dramatically different, and the diode

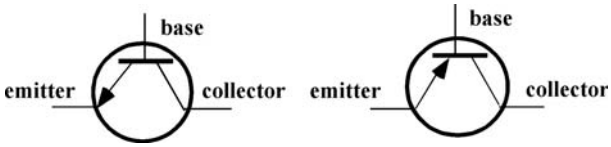


FIGURE 4 Circuit symbols for n–p–n and p–n–p BJTs.

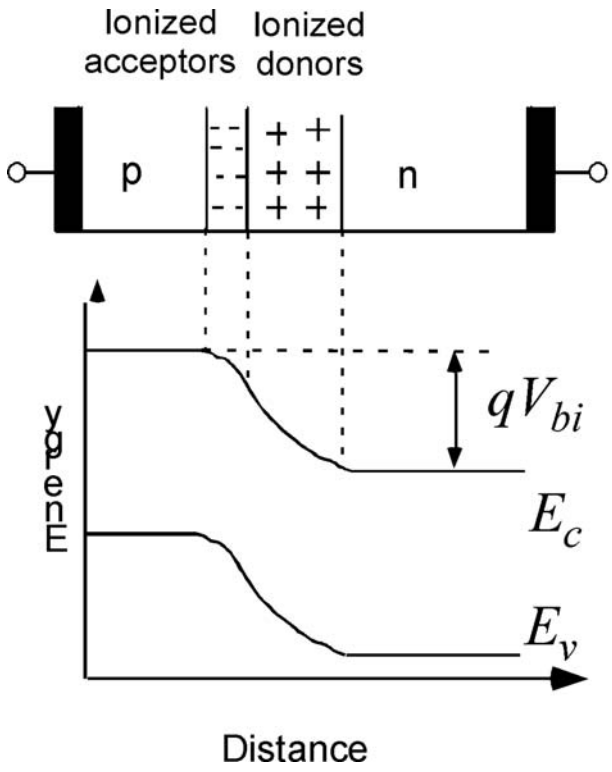


FIGURE 5 Schematic structure and band diagram of *p–n* junction.

current–voltage characteristic will look like that shown in Fig. 6.

Three important regions of operation are shown in Fig. 6. At reverse bias, the diode resistance is very large. At the forward bias below or close to the turn-on voltage,

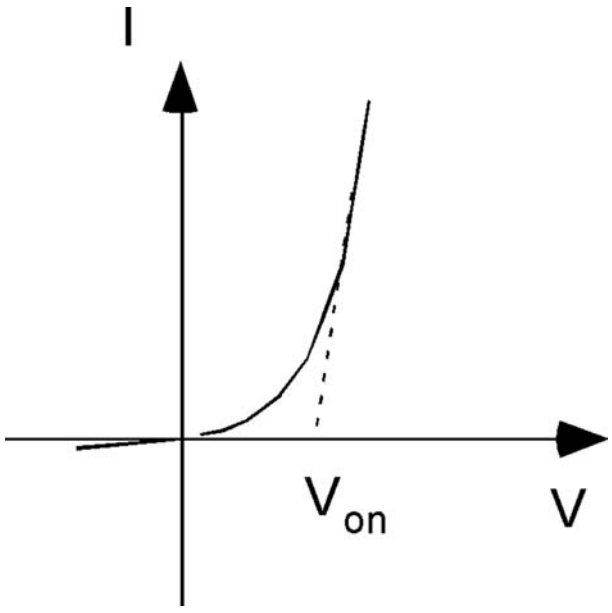


FIGURE 6 Qualitative current–voltage characteristic of a *p–n* junction diode.

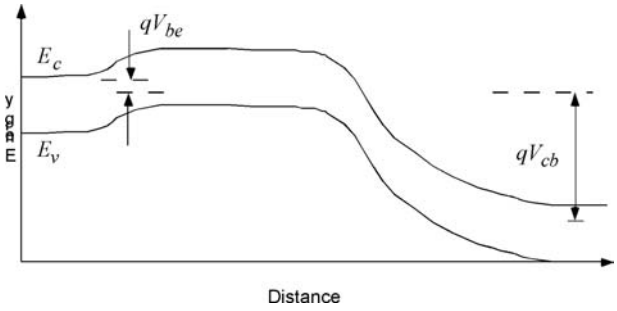


FIGURE 7 Schematic energy band diagram for *n–p–n* BJT.

V_{on}, the current increases exponentially with an increase in the forward bias. At forward bias substantially higher than V_{on}, the diode series resistance limits the current.

In a typical regime of BJT operation, the emitter-base *p–n* junction (see Fig. 3) is forward biased and the collector-base *p–n* junction is reverse biased. (This mode of operation is called the *active forward mode.*) Figure 7 shows a qualitative energy band diagram for an *n–p–n* transistor for this case.

The forward emitter–base voltage reduces the barrier for the electrons from the emitter region to cross into the base and the barrier for the holes to cross into the emitter region. The holes injected from the base into the emitter region flow in the emitter–base circuit loop, entering the base from the base ohmic contacts. These holes constitute the base current. If the base region were very long, the emitter–base and collector–base junctions would be independent. In this case, the electrons injected from the emitter into the base would have also contributed to the base current. In fact, the base region is made to be much shorter than the *diffusion length*, L_n, of the minority carriers in the base (i.e., electrons in an *n–p–n* BJT base region). The diffusion length is the characteristic length that determines the extent of diffusion of the minority carriers from the forward-biased junction into the semiconductor. As the minority carriers diffuse, they recombine. As a consequence, the diffusion length, L_n, depends on the diffusion coefficient, D_n, and on the lifetime, τ_n (i.e., on the characteristic recombination time of the minority carriers):

$$L_n = (D_n \tau_n)^{1/2} \tag{1}$$

If the base region were much longer than the electron diffusion length, these electrons would all recombine in the base. Since the base width is much shorter than L_n, only a tiny fraction of the electrons recombine, and the base recombination current is small. Most of the electrons injected from the emitter into the base reach and enter the collector region. The holes coming from the base into the emitter region constitute only a small fraction of the emitter current, since the base doping is very small compared to the emitter doping. Also, the hole mobility and diffusion

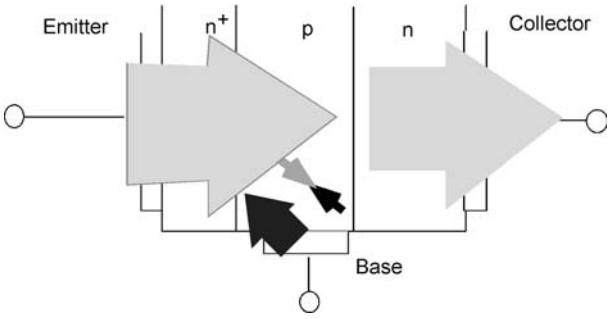


FIGURE 8 Principle of BJT operation.

coefficients (in silicon and in most other important semiconductor materials) are much smaller than the electron mobility and diffusion coefficients.

Figure 8 shows electron and hole fluxes in an n–p–n BJT in the forward active mode. The light-shaded arrows in Fig. 8 represent electron fluxes. The dark-shaded arrows represent hole fluxes. The larger dark arrow corresponds to the holes injected into the emitter region. The two small arrows in the base region indicate recombination of electrons and holes in the base.

Figure 9 shows a more realistic BJT structure and the streamlines of the base current. The lateral current flow in the base leads to a voltage drop along the base related to the resistance, called the *base spreading resistance*. It is very important to minimize the base spreading resistance, which affects the maximum frequency of the transistor operation and might lead to a nonuniform distribution of the emitter current (called *emitter current crowding*). As described below, this resistance can be drastically reduced in heterostructure bipolar transistors.

II. BJT CIRCUIT CONFIGURATIONS AND CURRENT GAINS

A bipolar junction transistor can operate in three possible circuit configurations: the common base, common emitter, and common collector (see Fig. 10). A small variation of the emitter–base voltage, v_e, caused by the signal, leads to a nearly equal variation of the emitter and the collector currents ($\delta I_c \sim \delta I_e$). Let us first consider the common-base configuration. If the load resistance is much larger than the source resistance ($R_L \gg R_S$), the voltage variation $\delta V_L = \delta I_c R_L$ across R_L can be much greater than v_e. (With an appropriate value of R_L, the maximum value of δV_L is limited by the collector voltage supply, V_{cc}.) Hence, we can have a small-signal voltage gain of the order of R_L/R_s, and a small-signal, common-base current gain,

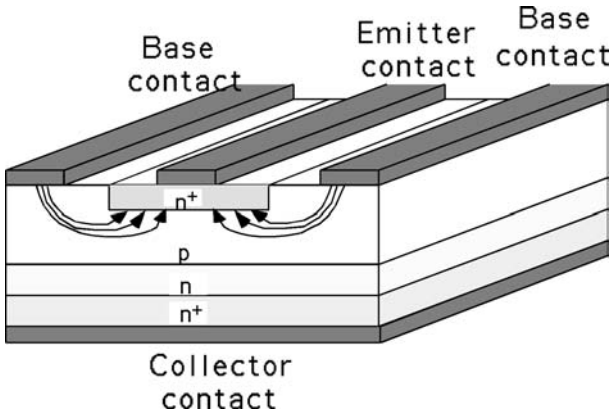


FIGURE 9 BJT structure and streamlines of base current.

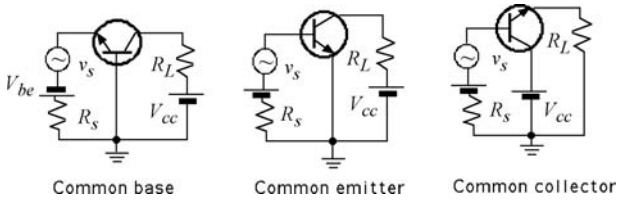


FIGURE 10 Circuit configurations for a bipolar junction transistor (BJT).

$$\alpha = \delta I_c / \delta I_e \tag{2}$$

close to 1. A small-signal power gain is nearly the same as the small-signal voltage gain. The voltage and power gains occur because electrons are injected from the emitter region through the base into the collector region, where the voltage drop is large compared to the small forward emitter–base voltage.

The common emitter configuration provides not only a high voltage gain but also a high current gain,

$$\beta = \delta I_c / \delta I_b \tag{3}$$

as well, since the input current is the base current and the output current is the collector current. This is the most popular BJT circuit configuration. Since $\delta I_b = \delta I_e - \delta I_c$, from Eqs. (2) and (3), we obtain:

$$1/\beta = (\delta I_e - \delta I_c)/\delta I_c = 1/\alpha - 1$$

Hence,

$$\beta = \frac{\alpha}{1-\alpha}, \qquad \alpha = \frac{\beta}{1+\beta} \tag{4}$$

In the common collector configuration, the current gain is equal to $\beta + 1$, but the voltage gain is close to unity.

Modern BJTs have a nonuniform doping in the base, with the acceptor density decreasing towards the collector–base junction (such transistors are called *graded base* transistors.) The nonuniform doping leads to a nonuniform hole concentration along the base. Hence, the holes diffuse from the emitter side of the base toward the collector side of the base. This creates an excess positive charge at the collector side of the base, an excess

negative charge at the emitter side of the base, and, therefore, a built-in electric field. This electric field pushes the minority carriers (electrons) injected into the base toward the collector. As a consequence, the minority carriers take less time to traverse the base, and their recombination is less effective than for a uniformly doped base. Hence, graded-base transistors have a higher gain.

III. BJT MODES OF OPERATION AND CURRENT–VOLTAGE CHARACTERISTICS

So far, we considered BJT operation when the emitter–base junction is forward biased and the collector–base junction is reverse biased. This mode of operation is called the *forward active mode.* In the *reverse active mode*, the emitter–base junction is reverse biased, and the collector–base junction is forward biased. In the *saturation mode*, both junctions are forward biased. Finally, in the *cutoff mode*, both junctions are reverse biased. These four modes cover all possible regimes of a BJT operation.

In the saturation mode, the injection of electrons from the collector region is in the opposite direction compared to the injection of the electrons from the emitter region. Therefore, the collector current in this regime is smaller than in the forward active mode. In the cutoff regime, the transistor currents are very small. In the reverse active mode, more holes are injected into the collector region than electrons into the base because of the higher doping in the base. This leads to a very low transistor gain in this regime, and this regime is of little practical importance. This mode of operation may be used for the extraction of transistor parameters (transistor characterization).

Figure 11 shows current–voltage characteristics of a BJT simulated using AIM-Spice™—a software program that allows simulation of electronic devices and circuits. (It runs on personal computers under Windows and can be downloaded for free from http://www.aimspice.com.) The AIM-Spice™ default parameters were used in this simulation, except for the parameter called *early voltage*, VAF, which determines the slope of the current–voltage characteristics at large values of the collector–emitter voltage, V_{ce}. (The value of $VAF = 15$ V was used.)

At large collector–emitter voltages, V_{ce}, the collector–base junction is reverse biased and the emitter–base junction is forward biased. Hence, the transistor operates in the active forward mode. As can be seen from Fig. 11, in this regime the collector current is approximately proportional to the base current ($I_c \sim \beta I_b$). When V_{ce} is decreased to values considerably smaller than the built-in voltage (approximately 0.6 V for a silicon BJT), the emitter–collector bias is not large enough to forward bias the emitter–base junction and reverse bias the collector–base junction. Both junctions become forward biased, and the BJT operates in the saturation mode.

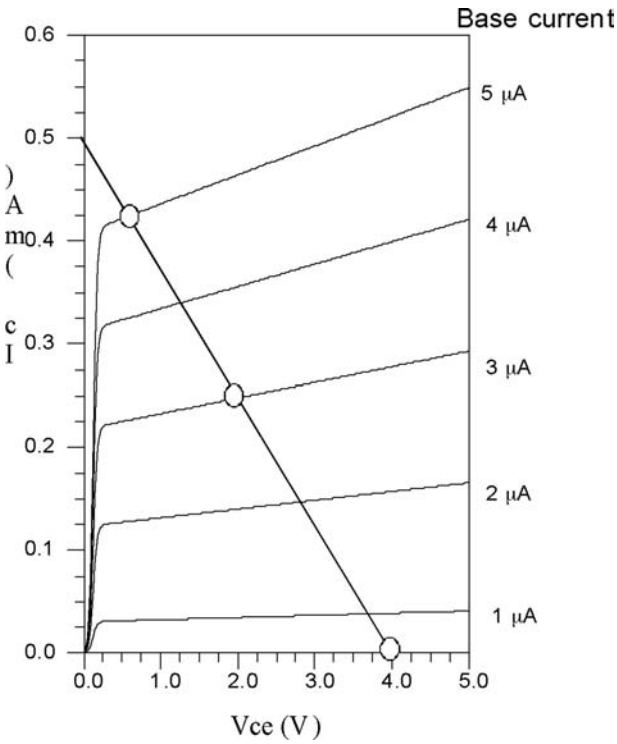


FIGURE 11 BJT current–voltage output characteristics for common emitter configuration. Also shown is a load line for the load resistance $R_L = 8$ kΩ and two operating points corresponding to the base currents of 1 μA and 5 μA.

As can be seen from Fig. 11, in this mode, the collector current decreases with the decrease in the collector–emitter bias. At zero base current, the BJT is in the cutoff mode, and the collector current is extremely low. By changing the base current, a BJT can be switched from the cutoff mode in the saturation mode or vice versa. If the BJT emitter is grounded and the collector is attached via a load resistance, R_L, to a power supply, V_{cc}, the collector–emitter voltage is given by:

$$V_{ce} = V_{cc} - I_c R_L \tag{5}$$

The straight line in Fig. 11 (called the *load line*) represents this dependence. The intersection of the load line with the transistor current–voltage characteristics yields operating points. As can be seen from the figure, for zero base current, the operating point corresponds to a high collector–emitter voltage and practically zero collector current state. For a larger base current (5 μA), the operating point corresponds to a low collector–emitter voltage and high collector current state. Switching between these two states can be achieved by a small change in the base

current. For applications in digital circuits utilizing BJTs as switches, these devices may be fabricated with multiple collector or emitter contacts, allowing for many outputs. Even though, in most digital circuits, silicon field effect transistors have replaced the BJTs, recently developed silicon–germanium heterostructure bipolar junction transistors might find many applications in ultra-high-speed circuits.

IV. BJTS, FETS, AND BICMOS

Since BJT currents are exponential functions of the emitter–base voltage, a BJT transconductance (which is the derivative of the collector current with respect to the emitter–base voltage) is quite high compared to that in a *field effect transistor* (FET). For a 1-μA collector current at room temperature, a typical BJT transconductance, g_m, is close to 40 msec. For an equal area device, a typical FET transconductance is at least several times smaller. High transconductance and current swing make the bipolar junction transistor a device of choice for many high-speed and high-power applications in both discrete and integrated circuits.

Bipolar junction transistors, however, circuits usually require more power than similar circuits implemented using FET technology. This limits the maximum number of transistors one can put on a semiconductor chip (i.e., the power requirements limit the integration scale). The very strong exponential dependence of the BJT current on temperature may also be a negative factor. If a certain transistor in an integrated circuit would have a higher temperature because of an unavoidable parameter variation or because of poor cooling conditions in this particular spot, the current through this particular transistor will be considerably higher, leading to a larger power dissipation in this particular spot and to a further temperature increase.

Field effector transistor technologies are more suitable for very large scale integrated (VLSI) circuits. Recently, a combined BJT–FET technology became a possible choice for fast integrated circuits. In this technology, called BiCMOS (bipolar complementary metal oxide semiconductor) field effect transistors. BJTs and FETs are fabricated next to each other in the same integrated circuit; however, even this technology faces stiff competition from more conventional field effect technology utilizing very small device sizes.

V. SMALL-SIGNAL PARAMETERS

Figure 12 shows the common-emitter output characteristics of the same transistor with superimposed small-signal variations of the base and collector currents. The operating

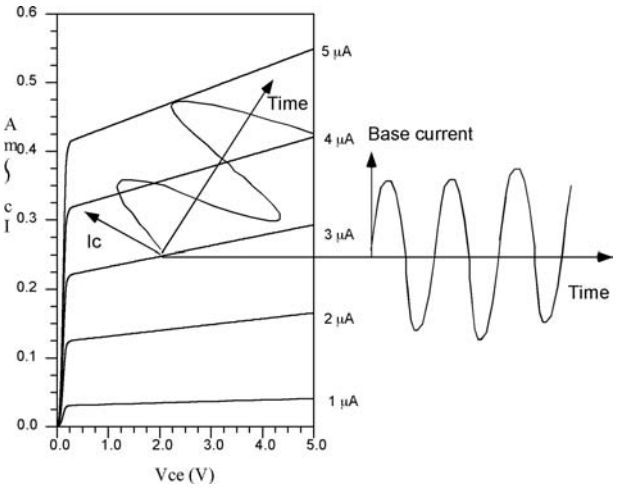


FIGURE 12 Small-signal (linear) regime of transistor operation.

point corresponds to a base current of 3 μA. The amplitude of the base current variation is 1 μA so that the base current changes with time in a sinusoidal fashion from 2 to 4 μA. A small variation of the base current leads to a relatively large variation of the collector current (from 120 to 310 μA, with the operating point at 200 μA) corresponding to a current gain of approximately 90.

For small base-current variations, the transistor gain and other parameters remain nearly the same. Under such conditions, the transistor response to a small-signal (i.e., an ac response) may be described in terms of a linear two-port network. Of course, this is no longer true for larger variations of the base current, since the current gain depends on the collector current. Let us denote the input small-signal variations of the current and voltage as i_1 and v_1, and the output small-signal variations of the current and voltage as i_2 and v_2. Each port of a linear two-port network may be represented by a series combination of an impedance and a voltage source or a parallel combination of an admittance and a current source. An *h-parameter* equivalent circuit shown in Fig. 13 describes a small-signal transistor response as follows:

$$v_1 = h_{11} i_1 + h_{12} v_2 \tag{6}$$

$$i_2 = h_{21} i_1 + h_{22} v_2 \tag{7}$$

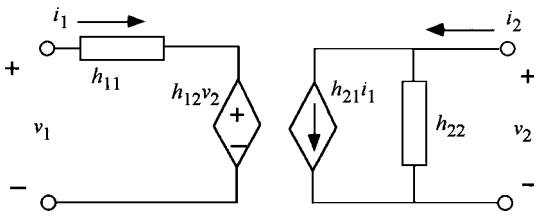


FIGURE 13 *h*-parameter representation of BJT small-signal equivalent circuit.

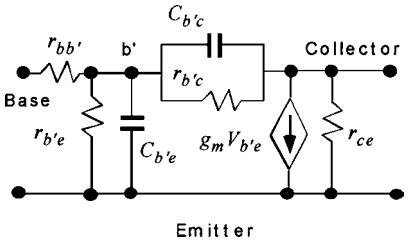


FIGURE 14 Hybrid-π equivalent circuit.

The h-parameters are most frequently used for transistor characterization at relatively low frequencies (below 100 MHz or so) and are specified by transistor manufacturers in their data sheets.

The parameter h_{11} (h_i) is called the *short-circuit input impedance*, h_{12} is called the *open-circuit reverse voltage ratio* (h_r), h_{21} is called the *short-circuit forward current ratio* (h_f), and h_{22} is called the *open-circuit output admittance* (h_o). In the alternate notation (h_i, h_r, h_f, and h_o,), a second subscript is often used to denote the transistor configuration. For example, the h-parameters for the common-emitter transistor circuit configuration are denoted as h_{ie}, h_{re}, h_{fe}, and h_{oe}. Manufacturers' data sheets usually provide only common-emitter h-parameters; however, h-parameters for other BJT circuit configurations can be expressed in terms of common-emitter h-parameters.

At low frequencies, h_{fe} is equal to β and h_{fb} is equal to $-\alpha$. For other h-parameters, such relationships are more complicated. Therefore, the circuit analysis of BJTs often relies on other equivalent circuits whose parameters are more directly related to the physical device parameters. The equivalent circuit called the *hybrid-π equivalent circuit* is often used for analysis of the small-signal response of the common-emitter configuration (see Fig. 14).

The definitions of the small-signal parameters show that these parameters can be introduced using open-circuit and short-circuit conditions at the input and output. However, at high frequencies (such as frequencies corresponding to the microwave frequency range starting from 1 GHz and above), it may be very difficult to realize such conditions. At such frequencies, interconnects between devices can no longer be considered as lumped resistances since their dimensions become comparable to the wavelength of the electromagnetic field. These interconnects behave as transmission lines. In this case, small-signal parameters called *s-parameters* (or scattering parameters) are used to describe the transistor response; s-parameters are very useful for describing BJT behavior at high frequencies.

The transistor gain depends on the frequency of operation. At high frequencies, the gain decreases. The frequency at which the common emitter current gain drops to unity (for zero load resistance) is called the *cutoff frequency*, f_T. This frequency is related to the effective delay time, τ_{eff}, by:

$$f_T = 1/(2\pi\tau_{\text{eff}})$$

where τ_{eff} is determined by the transit time of minority carriers across the base, the transit time of minority carriers across the collector-depletion region, and by the collector capacitance charging time. Another important frequency is the maximum frequency of oscillations, f_{max}, which is the frequency at which the power gain under optimum matching conditions for the input and output signals is equal to unity.

VI. HETEROJUNCTION BIPOLAR TRANSISTORS

William Shockley proposed the idea of the heterojunction bipolar transistor (HBT) in 1951. He suggested that an emitter region with a wider energy band gap than the base can present an additional barrier for the minority carrier injection from the base region (see Fig. 15). This should result in an increased current gain. What is even more important, the presence of this barrier allows us to dope the base region higher without ruining the transistor gain. A higher base doping results in a much smaller base-spreading resistance and greatly improves the transistor frequency response.

A related device is a tunneling emitter bipolar transistor (TEBT) (see Fig. 16). In this device, a barrier between the emitter and the base region is relatively transparent for electrons that have a small effective mass but blocks hole injection into the emitter region, since the holes are much heavier than electrons.

Most HBTs use AlGaAs/GaAs heterostructures since an Al atom has almost the same size as a Ga atom. Therefore, the lattice constants of AlGaAs and GaAs are very close, and AlGaAs/GaAs heterostructures have very high-quality interfaces. However, as mentioned above, Si–Ge HBT technology has recently emerged to challenge the

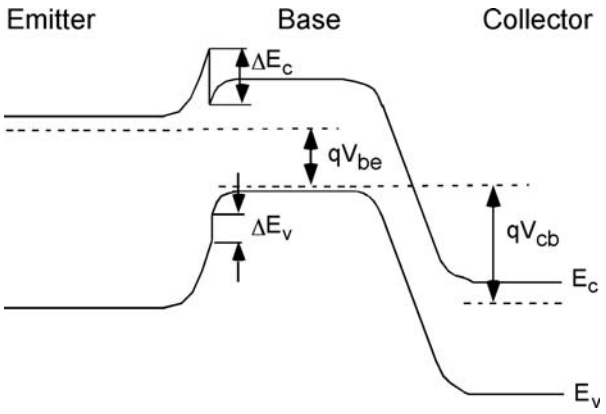


FIGURE 15 Schematic band structure of *n–p–n* heterojunction bipolar transistor (HBT).

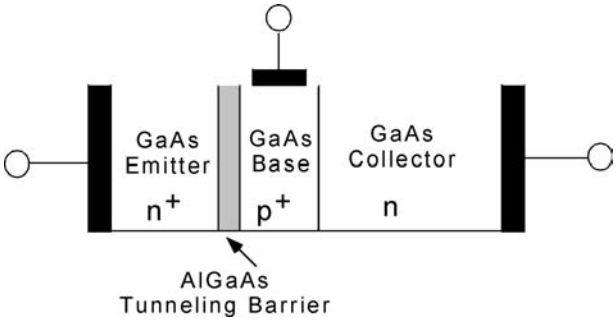


FIGURE 16 Schematic diagram of TEBT.

AlGaAs/GaAs and related technologies. A great advantage of Si–Ge HBT technology is its compatibility with conventional silicon processing, which is the mainstream of modern microelectronics technology.

SEE ALSO THE FOLLOWING ARTICLES

FIELD-EFFECT TRANSISTORS • INTEGRATED CIRCUIT PACKAGING • SEMICONDUCTOR ALLOYS • THIN FILM TRANSISTORS

BIBLIOGRAPHY

Asbeck, P. M., Chang, M. G., Wang, K. C., and Miller, D. L. (1990). "Heterojuction bipolar transistor technology," *In* "Introduction to Semiconductor Technology: GaAs and Related Compounds" (C. T. Wang, ed.), John wiley & Sons, New York.

Getreu, I. (1970). "Modeling the Bipolar Transistor," Part #062-2841-00, Tektronix, Inc.

Lee, K., Shur, M. S., Fieldly, T. A., and Ytterdal, T. (1993). "Semiconductor Device Modeling for VLSI," Prentice-Hall, Englewood Cliffs, NJ.

Neudeck, G. W. (1983). "The Bipolar Junction Transistors," Vol. III, Modular Series on Solid State Devices, Addison–Wesley, Reading, MA.

Shur, M. S. (1990). "Physics of semiconductor Devices," Solid State Physical Electronics Series, Prentice-Hall, Englewood Cliffs, NJ.

Shur, M. S. (1996). "Introduction to Electronic Devices," John Wiley & Sons, New York.

Sze, S. M. (1981). "Physics of semiconductor Devices," 2nd ed., Wiley, New York.

Sze, S. M. (1985), "Semiconductor Devices: Physics and Technology," Wiley, New York.

Sze, S. M., ed. (1998). "Modern Semiconductor Device Physics," John Wiley & Sons, New York.

Warner, R. M., and Grung, B. L. (1983), "Transistor Fundamentals for the Indegrated-Circuit Engineer," Wiley, New York.

Xu, J., and Shur, M. S. (1986). "Tunneling bipolar transistors," *IEEE Electron. Dev. Lett.* **7,** 416-418.



Acoustic Wave Devices

Erhard K. Sittig
Bergische Universität Wuppertal

I. Types of Acoustic Wave Devices and Their Evolution
II. Dielectric and Elastic Materials
III. Piezoelectric and Elasto-Optic Materials
IV. Sound and Light Waves in Solids
V. Diffraction, Directivity, and Interference
VI. Piezoelectric Transducers
VII. Piezoelectric Resonators
VIII. Bulk Wave Delay Lines
IX. Acousto-Optic Light Deflectors and Modulators
X. Surface Acoustic Wave Devices
XI. Technology for Bulk and Surface Acoustic Wave Devices

GLOSSARY

Anisotropic Having properties that vary with direction.
Elastic Responding reversibly with forces to deformations.
Elasto-optic Responding with changes in light propagation data to deformations.
Inhomogeneous Having properties that vary with location.
Piezoelectric Responding with deformations to applied electric fields.
Single crystal Material built up from replicated cells of ordered atoms.
Solid material Material resisting shear deformations with elastic forces.
Strain Array of deformation components.
Stress Array of elastic forces.
Tensor Array of constants or variables with certain mathematical properties.
Transducer Device for transforming electrical to sound signals and vice versa.

IN PRESENT-DAY electronics, acoustic wave devices (AWDs) are electronic components of a system that use sound waves as part of a signal transmission path. They do not include such things as sound generators from sirens to musical instruments, electroacoustic devices such as microphones and loudspeakers, and ultrasonic drills. Here, *sound* means any type of elastic wave at frequencies of human hearing, that is, 10,000–15,000 and above (also ultrasound); light means electromagnetic waves in the

wavelength range below ~100 μm. AWD components exploit the fact that sound travels at velocities v_s of 1–10 km/sec in most solid materials, but electric signals travel at velocities v_1 of 100,000–300,000 km/sec. A sonic transmission path for a given delay T is thus scaled down by a factor v_s/v_1 relative to the equivalent length of a coaxial cable. In signal processing, signal delays of many microseconds may be needed, and AWDs offer cost and size advantages. AWDs offer stable and compact frequency and time control. Sound waves in solids can also be used to modulate and scan laser light beams.

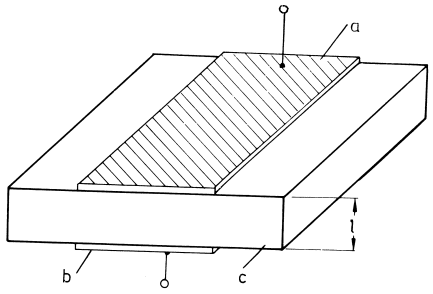


FIGURE 2 Crystal filter. (a, b) Metal thin film electrodes: (c) piezoelectric crystal plate.

I. TYPES OF ACOUSTIC WAVE DEVICES AND THEIR EVOLUTION

The classical AWD components are the ultrasonic delay line and the crystal filter. An ultrasonic delay line consists, in principle, of a sound transmission medium of propagation path length L, to the ends of which two "transducers" are attached (Fig. 1). The one at the transmitting end converts an electrical signal input to a sound beam, which travels along the transmission medium to the transducer at the receiving end, which reconverts the sound signal to an electrical one. In a typical set up the delay path may consist of a glass and the transducers of plates of a piezoelectric material affixed to it. A piezoelectric material responds to an applied electric field with a mechanical deformation or, when deformed, produces a voltage between electrodes attached to it.

The crystal filter belongs to the category of resonators (Fig. 2). If a signal is applied to the electrodes of the piezoelectric transmission medium, it travels along the path between the electrodes, is reflected at either end, and returns. These multiple reflections traveling back and forth may add in phase and thereby produce an enhanced output. This occurs for any frequency component of the signal for which the transmission path length L equals an integer multiple of a half-wavelength. Wavelength λ, frequency f, and velocity v are related by the familiar relation $v = f\lambda$, and thus the resonator will enhance a fundamental resonance frequency $f_0 = v/2L$ and its *overtones*. At this point the principle applies as well to organ pipes as to the resonators for light waves such as the cavity of a laser. We note that for sound and light at the same wavelength their frequencies f_s and f_l and for equal delay time T the necessary delay path lengths L_s and L_1 are in the ratio v_s/v_1 according to the relations

$$\begin{aligned} f_s\lambda_s &= v_s; \qquad f_1\lambda_1 = v_1 \\ L_s &= v_sT; \qquad L_1 = v_1T. \end{aligned} \tag{1}$$

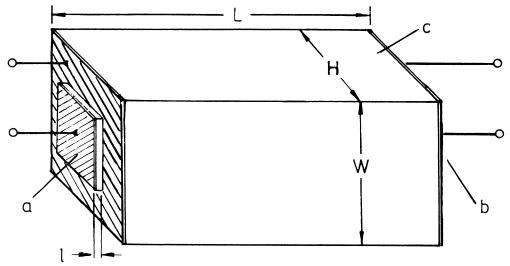


FIGURE 1 Ultrasonic delay line. (a, b) Piezoelectric transducers; (c) delay medium.

Sound wave resonators, unlike electromagnetic wave filters, make use of the scaling ratio v_s/v_1 to obtain device dimensions in the centimeter range at frequencies above ~1 MHz. However, in that frequency range electrical filters can also be had from inductors and capacitors. It turns out that a single LC circuit (i.e., a resonance circuit consisting of a single inductor and capacitor) typically has an inverse relative bandwidth Q of, at most, a few hundred at room temperature. A reasonably well-designed AWD such as a quartz filter offers, however, a Q of the order of 1 million and a much higher stability. One has to admit, however, that wide-band tuning is not, at present, feasible with an AWD.

Historically, AWDs were developed before World War II. In the late 1930s, precision frequency and time generators using quartz crystals were well established, as were adjustable delay lines with liquid transmission media. During World War II, radar systems evolved. The need for target simulation, discrimination of moving targets from stationary ones, and suppression of jamming by the enemy became pressing and stimulated further development greatly on either side. After the war, computer technology provided further impetus as ultrasonic delay lines proved capable of providing fast digital storage. Magnetic storage devices and integrated circuits capable of random access emerged later, and the use of delay line memories subsided, because they were capable only

of volatile—if fast—serial storage. Material development culminated, during the postwar years, in the production of various synthetic crystals and ceramics that were better optimized for specific applications than traditional ones, such as quartz and fused quartz glass. The major work in this area was done by 1970, but the development of computational and technological resources has led to deeper understanding and new approaches.

Today, delay lines and filters have developed into quite sophisticated devices. A delay line storing one television line for phase comparison with the next one is part of every television set operating by the PAL color TV standard common in Europe, making the color control unnecessary. Ceramic intermediate-frequency (IF) filters are part of almost every radio receiver, dispensing with relatively bulky tunable coils. The digital wristwatch uses piezoelectric devices as time base and "beeper." The simple resonator has developed into the complicated monolithic filter containing many resonators integrated on a common substrate. Modern radar systems have, since the days of simple moving-target indicators of World War II vintage, relied on AWDs as signal processors to increase range and decrease sensitivity to jamming. The simple delay line has developed into the transversal filter, a device with very generalized signal-processing capabilities. In the 1960s understanding of acoustic surface waves propagating in a narrow zone near a surface of a suitable material and the concepts of surface acoustic wave (SAW) devices evolved. Since these are more demanding of technological skill but share with silicon integrated circuits the economic advantages of batch fabrication technology, their development has surpassed the development of the latter. Thus, SAW delay lines and filters are destined to supplement the spectrum of components available to designers of electronic hardware.

A third group of AWDs is based on the capacity of a sound wave in elasto-optic materials to modulate the propagation characteristics for light waves. An acousto-optic deflector can scan a laser beam over a range of angles at speeds unattainable with a mechanically moving mirror, or it can modulate the beam at speeds unreachable with mechanical shutters. The fact that a sound beam can affect light propagation was discovered in the 19th century, but useful devices emerged only with the rise of electronics before World War II. In the mid-1930s there existed a television projector using this principle. Since 1970 acousto-optic modulators and deflectors for laser beams have found application in laser printers, pattern generators in microelectronics, and electrically tunable spectral filters in optical spectroscopy.

Integrated devices combining acoustic surface waves, surface-guided light waves, and manipulation of electric charges in surface layers of semiconductor components are beginning to emerge. Here, some aspects of high-energy concentration at lower power levels in surface layers, such as nonlinear effects, modulation, frequency translation, and others, may offer device concepts useful to the evolving field of modern wide-band communication via glass fibers and the like.

II. DIELECTRIC AND ELASTIC MATERIALS

Understanding and fabrication of AWDs require cooperation among materials science, device physics, and electronics. Materials science must provide materials and their formal description for a specific application.

The field of elastodynamics, needed to treat mathematically the propagation of sound waves and their dependence on material data, has a reputation for complexity. However, the basic principles are quite simple. In these days of the computer, we cast the basic principles in a "computer-friendly" from (i.e., a program) and leave the details, special cases, and numerical evaluation to the machine. A suitable notation facilitates grasping the essentials unencumbered by excessive and complicated formalism. We present these essentials to illustrate the problems facing material science in selecting and characterizing AWD materials, since many of them are anisotropic, that is, their properties depend on direction, and their optimal orientation also has to be found. Moreover, variables may be inhomogeneous, that is, depend on location.

A system of, say, three linear equations between variables x_1, x_2, x_3 and x_1', x_2', x_3' can be expressed in the form

$$\begin{aligned} x_1' &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ x_2' &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ x_3' &= a_{31}x_1 + a_{32}x + a_{33}x_3 \end{aligned} \tag{2}$$

We shall do this henceforth in tensor notation with implied summation, namely,

$$x_i' = a_{ij}x_j; \qquad i, j = 1, 2, 3, \tag{3}$$

implying that the subscript i should assume its three values and the equation thus generated should be stated. The subscript j is repeated in both factors of the product on the right-hand side. Wherever that occurs the subscript is to assume all its values, and the resulting terms are to be summed. It is apparent that if Eq. (3) is processed thus, Eq. (2) is reconstructed. Likewise, the statement

$$T_{ij} = c_{ijkl}S_{kl}; \qquad i, j, k, l = 1, 2, 3 \tag{4}$$

will expand into a system of nine equations for the variables T_{ij} and S_{kl} with 81 constants c_{ijkl}. Equations (3) and (4) provide compactness, and programming the algorithm for this "expansion" is straightforward. Writing the symbol, say, c_{ijkl} by itself, we mean an individual component of the array; writing (c_{ijkl}), we refer to the array as a whole.

In the following, let us indicate an array of variables by a letter, say c_{ijkl}, with each subscript assuming, unless stated otherwise, the values 1, 2, 3. Variables with no subscript we call scalar, with one subscript vector, with two subscripts matrix, as is customary. If the arrays follow certain transformation rules to be stated, we shall call them nth-rank tensors, with the value n simply being the number of subscripts appended. Thus, (c_{ijkl}) is a fourth-rank tensor. For brevity we also designate vectors by boldface type and second-rank tensors by boldface sans serif type. A tensor function, say $\mathbf{T}(\mathbf{x})$, is thus simply a short symbol for the array with the components $T_{ij}(x_1, x_2, x_3)$, meaning that each of them may depend on the variables x_1, x_2, x_3.

We now define a Cartesian coordinate system by three unit vectors $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, $\mathbf{e}_3 = (0, 0, 1)$ perpendicular to each other. Any point in space is thus given by its coordinate vector $\mathbf{x} = (x_1, x_2, x_3)$. The expression $\mathbf{T}(\mathbf{x})$ now states that all its components may vary with location, that is, may be inhomogeneous.

A particular point in space may be referred to one coordinate system (x_1, x_2, x_3) or another (x'_1, x'_2, x'_3). If the relation in Eq. (3) holds between the two, with (a_{ij}) being called the transformation matrix, then tensors in general are defined as such by obeying the following transformation rules:

$$y'_{ij} = a_{ip}a_{jq}y_{pq} \quad \text{for a second-rank tensor} \tag{5}$$

$$y'_{ijk} = a_{ip}a_{jq}a_{kr}y_{pqr} \quad \text{for a third-rank tensor} \tag{6}$$

$$y'_{ijkl} = a_{ip}a_{jq}a_{kr}a_{ls}y_{pqrs} \quad \text{for a fourth-rank tensor} \tag{7}$$

In this way tensors can be referred to any other coordinate system.

In a piece of elastic material (for visualization, rubber) single out a small cube-shaped volume element V in Fig. 3. Imagine that forces act on its faces that displace and deform V inside the material. Let the forces per unit area be labeled T_{ij}, with the first subscript indicating the coordinate x_i perpendicular to the face on which it acts, and the second j indicating the coordinate x_j parallel to which it acts. The set of these T_{ij} form the stress tensor $\mathbf{T}$. Let the location of V be given by its coordinate vector $\mathbf{x} = (x_1, x_2, x_3)$ and the displacement of one of its points after applying the stress by $\mathbf{u} = (u_1, u_2, u_3)$. $\mathbf{T}$ and $\mathbf{u}$ may be inhomogeneous. Since only the deformation of the volume element makes it contribute elastic forces of its own, not its rigid translation or rotation, subtracting those from $\mathbf{u}$ yields the components S_{kl} of the strain-tensor $\mathbf{S}$ describing this deformation by

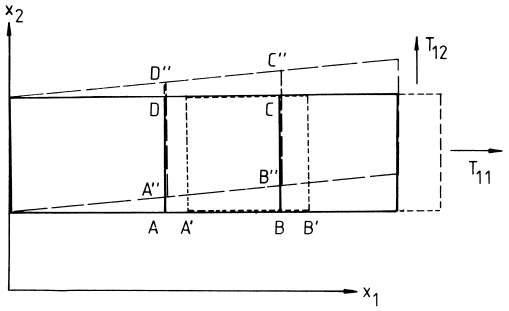


FIGURE 3 Deformation of a solid material. The contour *A, B, C, D* moves to *A′, B′, C′, D′* or *A″, B″, C″, D″* if an extensional or shear stress is applied.

$$S_{kl} = (\partial u_k/\partial x_l + \partial u_l/\partial x_k)/2 \tag{8}$$

A solid material resists, in general, an imposed strain tensor $\mathbf{S}$ with elastic forces, that is, a stress tensor $\mathbf{T}$. The stress–strain relation $\mathbf{S}(\mathbf{T})$ may not be simple. However, if there is no plastic flow, it is at least reversible and, for sufficiently small excursions from zero strain, can be approximated by a system of linear equations known as Hooke's law:

$$T_{ij} = C_{ijkl}S_{kl}; \qquad i, j, k, l = 1, \ldots, 3 \tag{9}$$

Thus, a material appears to need 81 elastic constants c_{ijkl} to describe its elastic properties. Various arguments, however, establish the relations

$$c_{ijkl} = c_{ijkl} = c_{jikl} = c_{jilk} = c_{klij} = \cdots \tag{10}$$

so that, at most, 18 of these constants can be independent of each other and have to be determined by measurement. To avoid these redundancies in the tensor notation, a more compact matrix notation with only two subscripts m, n ranging from 1 to 6 is customary in listings of materials data, in which

$$C_{mn} = C_{nm}; \qquad m, n = 1, \ldots, 6 \tag{11}$$

holds. The prescription for conversion is

$$\begin{aligned} m &= 1, 2, 3 \quad \text{for} \quad i = j = 1, 2, 3 \\ n &= 1, 2, 3 \quad \text{for} \quad k = l = 1, 2, 3 \end{aligned} \tag{12}$$

else

$$m = 9 - i - j; \qquad n = 9 - k - l; \qquad S_m = 2S_{ij}$$

These 18 elastic constants suffice, however, to describe single-crystal materials that are generally anisotropic.

Single crystals are materials with an ordered arrangement of their atoms over distances of centimeters and larger. They consist of unit cells, elementary geometric bodies like cubes, hexagonal prisms, and so on replicated like bricks in a brick wall to form a crystal lattice. The unit cells themselves contain individual atoms or molecules in an order of their own. Among all possible geometric bodies, only a few are admissible for geometric reasons. After all, one cannot lay a tile floor with tiles shaped like an equilateral pentagon. Considering admissible shapes of such unit cells, we find that single crystals can be grouped into seven crystal systems called triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic, ordered by increasing degree of symmetry. Considering also the possible symmetry properties of the atoms or molecules making up the unit cells, we find that all crystals can be assigned to one of 32 crystal classes of different degrees of symmetry.

Symmetry, in this context, means that the crystal lattice is indistinguishable from its former position after certain rotations or mirror imaging. The 32 classes are customarily labeled by symbols such as 6*mm* that denote the symmetry properties of a class and the system the class belongs to.

There are materials that can crystallize in different classes with profound differences in properties. Carbon, for instance, may appear as graphite, a hexagonal crystal, or as diamond, a cubic crystal. In addition, a material may consist of randomly oriented "crystallites" like ceramics. Such a material is globally isotropic, that is, has properties independent of direction on a large scale. Materials like glass are amorphous; that is, their building blocks possess no perceptible order at all. They also are isotropic, even on a small scale, say, below 1 μm.

Crystallographers conventionally use a Cartesian coordinate system with reference to directions of high symmetry of the unit cells and thus to the crystal as a whole. Tensor components are then expressed in these coordinates x_i but can be expressed in other coordinate systems via Eqs. (4)–(7).

The structural symmetry properties of the various crystal classes establish further relations between the elastic constants, thus reducing the number of independent ones further, depending on the individual class. Thus quartz, a crystal in the trigonal system (class 32), has the following set of elastic constants, referred to crystallographic coordinates:

$$\begin{aligned} c_{11} &= c_{22}; c_{33}; \quad c_{44} = c_{55}; \\ c_{66} &= (c_{11} - c_{12})/2 \\ c_{12}; c_{13} &= c_{23}; \quad c_{14} = -c_{24} = c_{56}/2; \\ \text{else} \quad c_{ij} &= 0. \end{aligned} \tag{13}$$

From Eq. (11) it follows that six independent elastic constants are needed to characterize quartz.

For isotropic materials one finds

$$c_{11} = c_{22} = c_{33} = c_{12} + 2c_{44}; \quad c_{44} = c_{55} = c_{66} \tag{14}$$

$$c_{12} = c_{13} = c_{23}; \quad \text{else} \quad c_{ij} = 0. \tag{15}$$

Thus, there are only two independent constants.

It may be instructive to compare this with the second-rank tensor in the relation between the electric flux density **D** and electric field **E** in an electrically insulating material:

$$D_i = \varepsilon_{ij} E_j \tag{16}$$

The crystal class of quartz has for the permittivity tensor ε:

$$\varepsilon_{11} = \varepsilon_{22}; \varepsilon_{33}; \quad \text{else} \quad \varepsilon_{ij} = 0. \tag{17}$$

Thus, there are two independent permittivities. Isotropic materials and those of the cubic system have

$$\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33}; \quad \text{else} \quad \varepsilon_{ij} = 0. \tag{18}$$

Thus, there is only one independent permittivity constant.

We note that in crystallographic coordinates the permittivity tensor has only three nonvanishing components ε_{ii} in all systems except the monoclinic and triclinic ones. Lists of material tensors in matrix notation are usually given in the literature referred to these coordinates. The matrices for higher rank tensors then contain nonvanishing off-diagonal terms. If recomputed for representation in other coordinate systems, even more components may appear but not more independent ones. If crystal plates are cut obliquely to the crystallographic axes, the symmetry properties for new axes along the edges of them can be reduced correspondingly. However, optimized data can be found in such directions, and obliquely cut crystal plates are not unusual in practical applications.

III. PIEZOELECTRIC AND ELASTO-OPTIC MATERIALS

Materials with built-in remanent dipole moments exhibit a piezoelectric effect and can be used as transducers between electric signals and sound waves. A dipole is defined as a distribution of equal positive and negative electrical charges $+q$ and $-q$, the centers of gravity of which do not coincide but are spaced a distance d apart. Let **d** be the vector pointing from $-q$ to $+q$, and

$$\mathbf{p} = q\mathbf{d} \tag{19}$$

the dipole moment. In a homogeneous electric field **E**, forces are exerted on the charges, tending to align **p**

parallel to **E**. If the charges are affixed to atoms of the crystal lattice in a solid material, the alignment tends to deform the crystal lattice and **S** depends now not only on **T** but also on **E**. Conversely, deformation alters the internal charge distribution in the material, and thus its electric flux density **D** now depends not only on **E** but also on **S**.

A functional dependence of the form

$$\mathbf{D} = \mathbf{D}(\mathbf{E}, \mathbf{S}); \qquad \mathbf{T} = \mathbf{T}(\mathbf{E}, \mathbf{S}) \tag{20}$$

can thus be postulated, the choice of **E** and **S** as independent variables being arbitrary. Setting both dependent variables equal to zero if **E** and **S** have all zero components, assuming reversibility and expanding **D** and **T** into a Taylor series, we obtain

$$D_i = \varepsilon_{ij}^S E_j + e_{ijk} S_{jk} + \zeta_{ijk} E_j E_k + f_{ijkl} E_j S_{kl} + \cdots \tag{21}$$

$$T_{ij} = -e_{kij} E_k + c_{ijkl}^E S_{kl} + q_{ijkl} E_k E_l + \cdots \tag{22}$$

where the new material constants are abbreviation defined by

$$\varepsilon_{ij}^S = (\partial D_i / \partial E_j)^S; \quad e_{ijk} = (\partial D_i / S_{jk})^E; \quad \text{etc.} \tag{23}$$

and the superscript indicates the variable held constant when carrying out the partial derivatives.

In this tensor notation there is a formidable array of material constants, since every subscript assumes values from 1 to 3 in the implicit summation. As an approximation for small excursions of the dependent variables, the material tensors up to the fourth rank only will be retained.

The individual tensors bear the following names: (ε_{ij}), dielectric tensor; (e_{ijk}), piezoelectric tensor; (ζ_{ijk}), electro-optic tensor; (f_{ijkl}), elasto-optic tensor; (c_{ijkl}), elastic tensor; (q_{ijkl}), electrostrictive tensor. A material possesses a piezoelectric effect if at least one e_{ijk} differs from zero; similarly a material has an electro-optic, elasto-optic, and electrostrictive effect if at least one of the corresponding tensor components differs from zero.

The same arguments used in simplifying the **c** tensor can be invoked to reduce the new tensors to a matrix notation with subscripts $m, n = 1, \ldots, 6$. Material tensors for other choices of independent variables, such as $\mathbf{E} = \mathbf{E}(\mathbf{D}, \mathbf{T})$ are, of course, equally valid and are used in the literature. Thus, again indicating the second variable to be held constant by a superscript, the piezoelectric effect can now be described by

$$D_i = e_{im} S_m + \varepsilon_{ij}^S E_j; \qquad T_m = c_{mn}^E S_n - e_{im} E_i \tag{24}$$

or other equivalent forms for other choices of variables, for example,

$$E_i = -h_{im} S_m + \beta_{ij}^S D_j; \quad T_m = c_{mn}^D S_n - h_{im} D_i$$

$$D_i = d_{im} T_m + \varepsilon_{ij}^T E_j; \quad S_m = s_{mn}^E T_n + d_{im} E_i \tag{25}$$

$$E_i = -g_{im} T_m + \beta_{ij}^T D_j; \quad S_m = s_{mn}^D T_n + g_{im} D_i$$

The constants in these can, of course, be calculated from those in Eq. (24), and all exhibit the same rules of symmetry as does the first set.

An argument from thermodynamics states that

$$c_{mn} = c_{nm}; \qquad \varepsilon_{ij} = \varepsilon_{ji} \tag{26}$$

and thus a triclinic crystal requires for the characterization of its materials properties 6 dielectric, 21 elastic, and 18 piezoelectric constants. The symmetry properties of the 32 crystal classes reduce this number. In particular, the crystal classes with centrosymmetry are found to have zero tensors of odd rank and thus neither a piezoelectric nor electrooptic effect. Centrosymmetry means indistinguishability of crystal data in a direction and the opposite direction. Among the 32 crystal classes, 10 are centrosymmetric and so are isotropic materials. These, therefore, possess neither a piezoelectric nor electro-optic effect.

They can, however, exhibit an electrostrictive or an elasto-optic effect. According to Eqs. (21) and (22), through these an "indirect" piezoelectric effect can be induced if the material is subjected to a homogeneous electric field. In certain ferroelectric ceramics, a field can be built into the material by a remanent dielectric polarization in analogy to remanent magnetization of a ferromagnetic material. Thereupon they behave like piezoelectric crystals in class 6*mm*, the symbol designating one of the classes in the hexagonal system. Since such ceramics can be produced in almost arbitrary shapes and are less expensive than crystals, this fact is of considerable importance. The remanent polarization is induced in a ferroelectric ceramic by heating it above a critical temperature, the Curie temperature, applying an electric field, and maintaining it while the material cools. However, since such materials depolarize at elevated temperatures and under high stress, their temperature and aging stability tend to be inferior to that of crystals. In piezoelectric crystals as well, the piezoelectric constants are subject to small variations in a stress or electric field, due to these other effects. The frequency stability of a crystal-controlled oscillator can thereby be impaired.

Elasto-optic tensors occur in all crystal classes as well as in isotropic materials. The form of the elasto-optic tensor, however, is not in all classes the same as the form of the elastic tensor. On conversion to matrix notation for the latter, $c_{\mathrm{mn}} = c_{\mathrm{nm}}$ holds, but $f_{\mathrm{mn}} \neq f_{\mathrm{nm}}$.

According to Eq. (21) the electric flux density vector **D** can be written in the form

$$D_i = \left(\varepsilon_{ij}^S + \zeta_{ijk}^S E_k + f_{ijkl} S_{kl}\right) E_k = \varepsilon'_{ij} E_j \qquad (27)$$

saying that $\mathbf{D}(\mathbf{E})$ can also be varied parametrically by E_k and S_{kl}, that is, via the electro-optic and elasto-optic effects. This means that light wave propagation can be modulated by electric or elastic fields in a material having such effects.

Opticists prefer to express the effects on light beam propagation in nonmagnetic materials not by (ε_{ij}), but by its inverse $(\varepsilon_{ij})^{-1}$, namely,

$$B_{ij} = \varepsilon_0 (\varepsilon_{ij})^{-1} \qquad (28)$$

with the absolute permittivity

$$\varepsilon_0 = 8.86 \times 10^{-12} \text{ As V}^{-1} \text{ m}^{-1} \qquad (29)$$

and write the electro-optic and elasto-optic effects, $B(0, 0)$ being the values for all $E = 0$ and $\mathbf{S} = 0$ in tensor notation,

$$\Delta B_{ij} = B_{ij}(\mathbf{E}, \mathbf{S}) - B_{ij}(0, 0) = r_{ijk}^S E_k + p_{ijkl}^E S_{kl} \qquad (30)$$

or an alternative formulation for different choices of independent variables. The symmetry rules prevailing permit the use of matrix notation in the form r_{mi} and p_{mn} $(m, n = 1, \ldots, 6)$. Devices based on the electro-optic effect, not being relevant to the present topic, will not be discussed further.

At this point it is worth mentioning that postulating a dependence of the magnetic flux density $\mathbf{B}$ on magnetic field $\mathbf{H}$ and, say, strain $\mathbf{S}$ yields the formulation of a similar set of effects where $\mathbf{E}$ is replaced by $\mathbf{H}$. Experimental data show, however, that only the magnetostrictive effect has a technically relevant magnitude. Polarized magnetostrictive materials are indeed used as transducers at lower frequencies. Piezoelectric transducers have, however, superior properties, especially at higher frequencies, and they are easier to couple to the peripheral electronic circuitry. Therefore, only these will be discussed further.

Once a complete set of material constants has been determined by measurment and expressed in Cartesian crystal coordinates, it can be recomputed with respect to another, rotated coordinate system by means of the tensor transformations mentioned. In this way we can find rotated crystal cuts, that is, plate orientations with faces cut obliquely to the original coordinate axes with data optimized for certain desiderata such as zero temperature coefficient. The data for such optimized cuts can then be used as inputs by the actual device designer. For illustration, Table I shows the matrix forms for some piezoelectric materials and Table II shows the numerical data describing them.

TABLE I Material Data Matrix Forms

Crystal class $3m$

Elastic constants C_{mn}:

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 0 & 0 \\ c_{12} & c_{11} & c_{13} & -c_{14} & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ c_{14} & -c_{14} & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & c_{14} \\ 0 & 0 & 0 & 0 & c_{14} & (c_{11} - c_{12})/2 \end{bmatrix}$$

Piezoelectric constants e_{im}:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & e_{15} & -e_{22} \\ -e_{22} & e_{22} & 0 & e_{15} & 0 & 0 \\ e_{31} & e_{31} & e_{33} & 0 & 0 & 0 \end{bmatrix}$$

Permittivity constants ε_{ij}:

$$\begin{bmatrix} \varepsilon_{11} & 0 & 0 \\ 0 & \varepsilon_{11} & 0 \\ 0 & 0 & \varepsilon_{33} \end{bmatrix}$$

Crystal class 32

Elastic constants (same as class $3m$)

Piezoelectric constants e_{im}:

$$\begin{bmatrix} e_{11} & -e_{11} & 0 & e_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -e_{14} & -e_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Permittivity constants ε_{ij} (sane as class $3m$)

Crystal class $6mm$

Elastic constants c_{mn} (same as class $3m$ but $c_{14} = 0$)

Piezoelectric constants e_{im} (same as class $3m$ but $e_{22} = 0$)

Permittivity constants ε_{ij} (same as class $3m$)

IV. SOUND AND LIGHT WAVES IN SOLIDS

A. Bulk Sound Waves in Solids

Each volume element in a material moves in dynamic equilibrium between the elastic forces acting on it and its inertial forces, with the latter depending on its mass density, the mass per unit volume. Equating them yields a system of partial differential equations, the wave equations

$$\partial^2 u_i / \partial t^2 = (c_{ijkl}/\rho) \partial^2 u_k / \partial x_j \partial x_l \qquad (31)$$

for the components u_i of the elastic displacement $\mathbf{u}$, with t designating time. These can be solved by inserting trial solutions $\mathbf{u}(\mathbf{x}, t)$ for which one can carry out the derivatives, thus converting Eq. (31) into a system of algebraic equations. One such trial solution has components

$$\begin{aligned} u_i(x, t) &= U \alpha_i \exp[jk(a_i x_i - vt)] \\ k &= 2\pi/\lambda \end{aligned} \qquad (32)$$

The complex exponential function describes a homogeneous plane wave with planes of constant phase

TABLE II Data for Three Piezoelectrical Materials at 20°C

	Lithium niobate $LiNbO_3$	Quartz SiO_2	Ceramic PZT5H[a]
Crystal class	$3m$	32	$6mm$
$\rho(10^3\times \mathrm{kg}^{-3})$	4.36	2.65	7.5
c_{11}^E $(10^9\times \mathrm{N\,m}^{-2})$	203	86.8	126
c_{12}^E	53	7.04	79.5
c_{13}^E	75	11.91	84.1
c_{14}^E	9	−18.04	0
c_{33}^E	245	105.75	117
c_{44}^E	60	58.20	23.0
e_{11} $(\mathrm{As\,m}^{-2})$	0	0.171	0
e_{14}	0	−0.041	0
e_{15}	3.7	0	17
e_{22}	2.5	0	0
e_{31}	0.2	0	−6.5
e_{33}	1.3	0	23.0
$\varepsilon_{11}^S/\varepsilon_0$	44	4.428	1700
$\varepsilon_{33}^S/\varepsilon_0$	29	4.634	1470

[a] Listed by Mattiat [(1971). "Ultrasonic Transducer Materials," Plenum, New York] as trademark of Vernitron Piezoelectric Division, Bedford, OH. A ceramic, it consists of ∼55% lead zirconate and ∼45% lead titanate with some of the lead substituted by lanthanum.

perpendicular to a wave vector **k** of magnitude k and direction cosines a_i against the coordinate axes x_i. An amplitude vector **U** of magnitude U and direction cosines α_i indicates the direction of the displacement. The planes have the spacing λ move with velocity v in the direction of **k** and pass through a fixed point (x_1, x_2, x_3) with frequency f. Thus,

$$v = f\lambda. \tag{33}$$

If we insert Eq. (32) into Eq. (31) and carry out the derivatives, the result assumes, with the auxiliary definitions,

$$\begin{aligned} &\delta_{ii} = 1; \quad \text{else} \quad \delta_{ij} = 0; \\ &\bar{c}_{ik} = a_j a_l c_{ijkl} = \bar{c}_{ki} \end{aligned} \tag{34}$$

the compact form

$$\left(\bar{c}_{ik} - \delta_{ik}\rho v^2\right)\alpha_i = 0 \tag{35}$$

and can be solved for the v^2 if the a_i are given. Equation (35) has nonvanishing solutions if its secular determinant vanishes, that is,

$$\det\left(\bar{c}_{ik} - \delta_{ik}\rho v^2\right) = 0 \tag{36}$$

If solved for v^2 this is a third-order equation and thus has three solutions v_1, v_2, v_3. If any of these are inserted into Eq. (35), the α_i can be found and thus also the direction of **U**, hence **u**.

In an anisotropic material, far from any boundary, three types of sound waves can thus propagate in any direction of **k** chosen, with, in general, three different velocities. One finds their amplitude vectors **U** to be mutually perpendicular to each other, but neither necessarily parallel nor perpendicular to **k**. They are called bulk waves because they travel freely through the bulk of the material. One of them has **k** and **u** nearly parallel and thus is called quasi-longitudinal: the other two have **k** and **u** nearly perpendicular and are thus quasi-transversal.

Boundary conditions (e.g., specified stresses or displacements on plane surfaces of the material) can be satisfied with linear combinations (i.e., sums of such solutions), each multiplied by an unknown factor. These sums one inserts into the boundary conditions. The result is another system of algebraic equations, from which the unknown factors can be determined.

In isotropic materials one finds that bulk sound waves can propagate as longitudinal waves with **u** parallel to **k** or as transversal waves with **u** perpendicular to **k**. Their phase velocities differ, however, being given, respectively, by

$$v = v_{\mathrm{L}} = (c_{11}/\rho)^{1/2}$$

and (37)

$$v = v_{\mathrm{T}} = (c_{44}/\rho)^{1/2}$$

and v_{T} is independent of the orientation of $\mathbf{u} \perp \mathbf{k}$. Since in most materials $v_{\mathrm{T}}/v_{\mathrm{L}}$ is in the range 0.6–0.4, the use of transversal waves results in smaller dimensions of the transmission path for a specified delay time and is thus preferred.

Even smaller delay lines result from folding up the transmission path by multiple reflection from the side walls (Fig. 4). These can be thought of as mirrors, as in optics, but the reflection properties of sound waves tend to be more complicated than those of light waves. In Fig. 5, let a plane sound wave with wave vector **k** impinge at the incidence angle θ on a plane interface between two solid isotropic materials with mass densities ρ_1, ρ_2, longitudinal velocities $v_{\mathrm{L}1}$, $v_{\mathrm{L}2}$, and transversal velocities $v_{\mathrm{T}1}$, $v_{\mathrm{T}2}$. Label the longitudinal wave with $\mathbf{k} \parallel \mathbf{u}$ the L wave, the transversal one with **u** normal to the plane of incidence

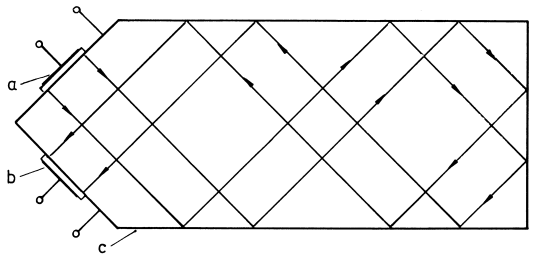


FIGURE 4 Folded ultrasonic delay line. The sound beam emitted by transducer a is reflected several times from side walls and end face of delay medium c before arriving at transducer b.

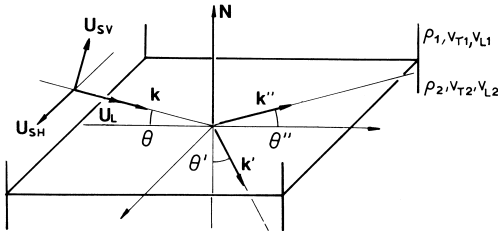


FIGURE 5 Reflection and refraction of a sound beam incident at angle θ and displacement orientation L, SV, or SH on an interface between materials with ρ_1, v_{L1}, v_{T1} and ρ_2, v_{L2}, v_{T2}, respectively.

defined by **k** and the interface normal **n** the SH wave, and the one with **u** in the plane of incidence the SV wave. The SH wave turns out to split into one reflected part with $\mathbf{k}''$ and one refracted part with $\mathbf{k}'$, which is transmitted into the second medium. The angles $\theta, \theta', \theta''$ follow the law of refractions as in optics, namely,

$$\theta'' = \theta; \qquad \sin\theta' = (v_{T2}/v_{T1})\sin\theta \tag{38}$$

and the reflection and transmission factors r and t, the ratios of reflected and transmitted amplitude to the incident one

$$r = (1 - z')(1 + z'); \qquad t = 2/(1 + z') \tag{39}$$

where

$$z' = (\rho_2 v_T \cos\theta')/(\rho_1 v_{T1} \cos\theta) \tag{40}$$

With the L wave or SV wave incident, more complicated relations result. In general, and dependent on incidence angle and materials data, an L wave and an SV wave may be reflected simultaneously and both wave types transmitted. The angles of reflection and transmission depend on the velocity ratios, and the amplitude reflection and transmission factors also depend on the mass density ratios of both materials. This phenomenon, called mode conversion, is troublesome in delay lines with folded transmission paths in that it may cause off-angle sound beams, giving rise to spurious, unwanted signals with differing transmission times, and, at best, cause increased transmission losses. The use of SH wave propagation is thus preferred if no other considerations intervene.

In anisotropic materials additional complications are found. They can be summarized as follows:

1. The two quasi-transversal waves may propagate with different velocities, which depend on orientation relative to crystallographic axes, as already stated.
2. The power propagation direction, defined by a vector **P** with components (**A** being the beam cross-sectional area),

$$P_i = AT_{ij}\partial u_j/\partial t \tag{41}$$

may deviate from **k**. In isotropic materials **k** and **P** are parallel.
3. Any of these waves, when impinging on an interface between two anisotropic materials, may launch up to three transmitted and reflected waves.

In spite of all the complexities attending the use of anisotropic materials, they are worth using for the following reasons:

1. The loss in single crystals due to absorption in a given frequency band tends to be lower than in isotropic materials. This is important at very high frequencies, because the absorption universally increases with frequency.
2. Single-crystal materials can, as a rule, be produced with higher purity and homogeneity than isotropic materials. Thus, better constancy and stability, particularly against aging, may result from the use of crystals.
3. In integrated designs, transducers and transmission medium are made from the same piezoelectric material. The piezoelectric effect does not exist in isotropic materials.
4. Desiderata, such as independence from temperature or optimized magnitudes of piezoelectric effects, may exist only in certain directions of certain crystal materials.

B. Surface Sound Waves in Solids

Another class of solutions can be obtained by inserting a chosen value for v into the determinant in Eq. (36) and solving for the a contained in the c_{ik}. For these, the determinant is a sixth-order equation. Since $\bar{c}_{ik} = \bar{c}_{ki}$, it then has three pairs of conjugate complex solutions. For those, the imaginary parts produce real exponents in Eq. (32), which can be factored out and are then seen to describe waves the amplitudes of which decrease or increase exponentially with distance. Only the former, called evanescent waves, are of interest here since they do not produce arbitrarily large energy deep in the material.

An important example of evanescent waves are Rayleigh or surface waves. In Fig. 6 the material has a plane surface normal to the x_3 axis, say, to air or vacuum, and is thus force-free there. The boundary conditions prevailing at $x_3 = 0$ are thus

$$T_{3j} = c_{3jkl}\partial u_k/\partial x_l = 0 \tag{42}$$

and can be satisfied with a linear combination of solutions with negative imaginary part for a_3 but real a_1 and a_2, thus causing the **k** vector to lie in the surface. This case describes a wave that travels along a direction on the

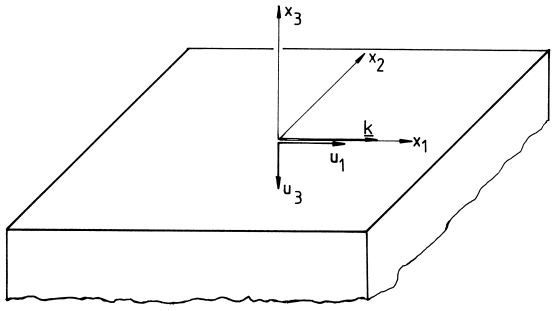


FIGURE 6 Orientation of surface wave displacements u_1 and u_3 for wave crest propagation direction **k** parallel to x_1 on the surface of an isotropic solid.

surface with "straight-crested" wave fronts but an amplitude vector the components of which decrease with depth. The wave thus appears to be "guided" by the surface and travels with v_R, the surface wave velocity.

For isotropic materials, v_R/v_T can be approximated by

$$\begin{aligned} v_R/v_T &= (0.87 + 1.12\sigma)/(1+\sigma) \\ \sigma &= (v_L/v_T - 2)/2(v_L/v_T - 1) \end{aligned} \tag{43}$$

Poisson's ratio σ is for all isotropic materials in the range 0–0.5, and $v_R < v_T$. For $\sigma = 0.25$, Fig. 7 shows the components u_1 and u_3. The component u_2 vanishes. What Fig. 7 does not display is the fact that u_1 and u_3 are out of phase by 90°. The amplitude vector thus performs an elliptical motion not unlike that of a small piece of wood floating on a water surface as a wavelet passes through it.

In anisotropic materials the situation is more complicated. The value of v may vary markedly with the direction of **k**, which, of course, it does not do in isotropic materials. The amplitude need not stay closely confined to the surface. Over a certain range of directions the wave motion may assume a very large penetration depth and turn into a surface-skimming bulk transversal wave. The amplitude may also not decrease monotonically with depth but may undergo a number of oscillations before it does. These surface acoustic waves are finding increasing application in AWDs for technical reasons mentioned later.

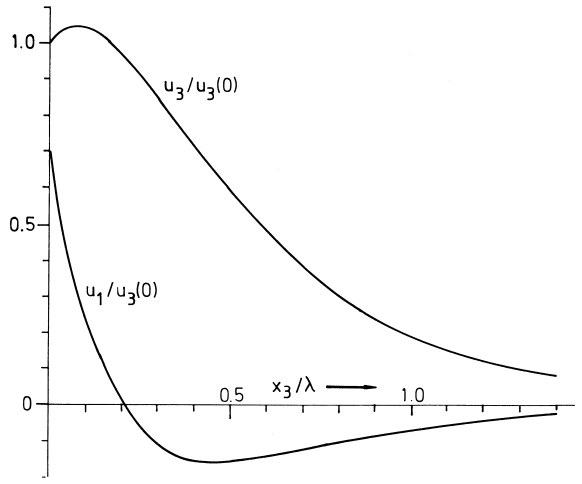


FIGURE 7 Surface wave displacement amplitudes u_1 and u_3 relative to u_3 at $x_3 = 0$ for a material with Poisson's ratio $\sigma = 0.25$.

Finally, sound waves can be "guided" in a layer of material sandwiched between two materials or deposited on top of another. in the layer, propagation modes exist with complicated distributions of the u_i. These may have velocities that not only depend on the geometric dimensions of the layer, but also depend on wavelength and are then called dispersive. The phenomenology somewhat resembles that of light propagation in optical fibers.

C. Waves in Piezoelectric Solids

With the constitutive relations for a piezoelectric solid [Eq. (24)], the wave equations become

$$\begin{aligned} \rho\partial^2 u_i/\partial t^2 - c^E_{ijkl}\partial^2 u_k/\partial x_j\partial x_l - e_{kil}\partial^2\phi/\partial x_k\partial x_l = 0 \\ e_{ikl}\partial^2 u_k/\partial x_i\partial x_l - \varepsilon^S_{ik}\partial^2\phi/\partial x_i\partial x_k = 0 \end{aligned} \tag{44}$$

with the electrostatic potential ϕ. These can be solved substantially in the same way as for nonpiezoelectric materials, but with a few additions. These can be summarized as follows.

Use the same trial solutions for the u_i as before. Define ϕ by

$$\phi = \alpha_4 \exp[jk(a_i x_i - vt)] \tag{45}$$

Insert

$$\begin{aligned} \bar{c}_{i4} &= e_{ijk}a_j a_k; \qquad \bar{c}_{44} = -\varepsilon^S_{ik}a_i a_k \\ \delta_{44} &= 0 \end{aligned} \tag{46}$$

into the four equations for the α_r to obtain

$$\left(\bar{c}_{rs} - \delta_{rs}\rho v^2\right)\alpha_r = 0; \quad r,s = 1,\ldots,4 \tag{47}$$

and solve its secular determinant either as a third-order equation for v if the a_i are specified, or solve it as an eighth-order equation for a_3 if v is specified. Insert the trial solution with the v's found into Eq. (46). Solve for the α_r. Proceed to solve stated boundary conditions with a linear combination of the wave types now determined, but note that additional boundary conditions must be stated for ϕ.

The result shows that the same types of waves exist as in nonpiezoelectric materials, but their velocities depend on the electrical boundary conditions and the permittivities on the mechanical ones. These waves are relevant for an understanding of piezoelectric crystal filters and SAW devices.

D. Light Waves in Solids

It is instructive to compare the relations describing sound wave propagation with those for light wave popagation. In transparent, nonmagnetic materials the light wave equations for the electric field vector can be derived from Maxwell's equations and written with Eq. (30) in the following form, with v_{10} being the light velocity *in vacuo*:

$$\partial^2 E_i/\partial t^2 = v_{10}^2 B_{ij}\delta_{kl}\partial^2 E_j \big/ \partial x_k \partial x_l \qquad v_{10} = 2.998 \times 10^8 \text{ msec}^{-1} \tag{48}$$

Proceeding in a similar way as for sound waves one finds the salient features of light wave propagation, as follows:

1. In isotropic materials there exists only one type of light wave with **E** perpendicular to **k**, that is, a transversal wave. In materials of the cubic system, where all ε_{ii} are also equal, the same is true. The power propagation direction is parallel to **k**.
2. At interfaces between isotropic materials one must distinguish between TE or *s* waves with **E** perpendicular to the plane of incidence and TM or *p* waves with **E** in the plane of incidence. For both there exists only one reflected and one transmitted wave, and the law of refraction for the angles $\theta, \theta', \theta''$ of the incident, refracted, and reflected beams is that of Eq. (38) with the light velocities instead of the sound velocities inserted, that is,

$$\theta'' = \theta; \qquad \sin\theta' = (v_{12}/v_{11})\sin\theta \tag{49}$$

However, the two types have different reflection and transmission factors.

3. In crystals of hexagonal or lesser symmetry two types of waves propagate in any direction of **k** with different velocities, say v_{1a} and v_{1b}. This phenomenon is called birefringence. There are certain optical axes, directions in which $v_{1a} = v_{1b}$ independent of the orientation of **E** $\perp$ **k**. Mode conversion exists between the two types, and the power propagation direction may deviate from the direction of **k**.
4. There are no surface waves, but guided waves exist in layers of lower velocity than that of the adjacent ones. Thus, in acousto-optic SAW devices such a layer must be provided on the surface to keep the light in the interaction region with the sound wave.

Evidently, the design of acousto-optic devices, in which light wave propagation is controlled by interaction with sound waves, can be of formidable complexity. The task of finding optimized materials, light and sound propagation directions, and the orientations of **E** and **u** is left to the materials scientist, who presents the results in a form useful to the device designer.

V. DIFFRACTION, DIRECTIVITY, AND INTERFERENCE

The wave equations for sound and light alike prescribe certain conditions of continuity on surfaces where the material data have discontinuities. These give rise to boundary waves, of which the reflections at interfaces were an example. But "stops" limiting the diameter of a light or sound beam do likewise. The resulting waves originating at lateral boundaries illuminate the geometric shadow zone and cause interference effects in the light zone due to phase-dependent addition and subtraction effects.

For the following assume that the entire surface of a planar emitter radiates in phase. In optics this can be realized by apertures illuminated with "coherent" light, in sonics by piezoelectric transducers with electrodes driven appropriately.

Let P_0 be the power emitted by a planar emitter normal to its rectangular radiating area with sides a and b. An approximate but elementary theory gives the angular range over which a wave of wavelength λ can be emitted. At an angle θ against the normal the emitter radiates the power P given by its directivity

$$P/P_0 = (\sin\alpha/\alpha)^2(\sin\beta/\beta)^2 \tag{50}$$

with (Fig. 8)

$$\alpha = (ka/2)\sin\theta_a; \qquad \beta = (kb/2)\sin\theta_b \tag{51}$$

A transducer with finite cross-dimensions transmits thus over an angular main lobe and sidelobes according to Eq. (50). In delay lines this causes beam spread loss since part of the power emitted by the transmitter does not reach the receiver. In addition, sidelobes being reflected off the side walls of the transmission medium can lead to unwanted spurious signals from the receiver. A receiver, a coherent detector in optics, a piezoelectric transducer in sonics, has the same directivity for waves impinging on it.

A grating consisting of p line emitters of width a, length $b \gg a$, and equidistant spacing Λ (Fig. 9) has the directivity

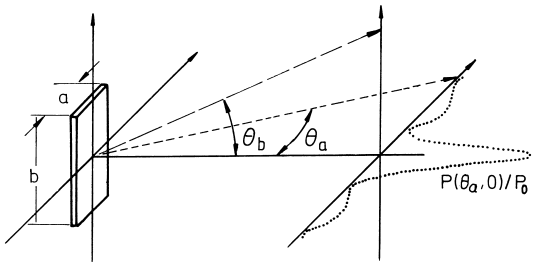


FIGURE 8 Directivity of a radiating plane surface $P(\theta_a, 0)/P_0$.

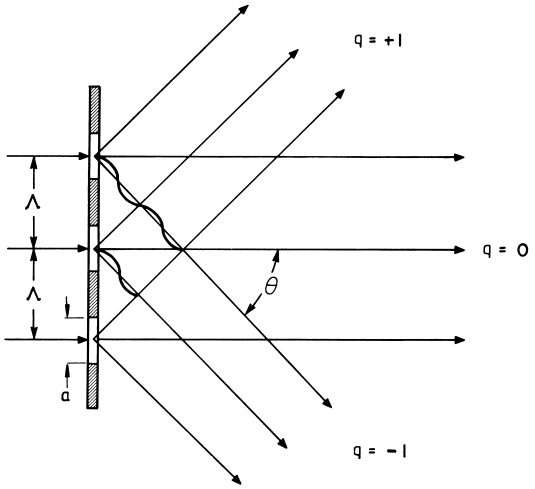


FIGURE 9 Diffraction orders q of a diffraction grating consisting of radiating lines of width a and spacing Λ, separated by gaps that do not radiate (amplitude grating) or radiate with different phase (phase grating).

$$P/P_0 = \eta(\sin\alpha/\alpha)^2(\sin p\gamma/p\sin\gamma)^2$$
$$\gamma = (k\Lambda/2)\sin\theta \tag{52}$$

with the diffraction efficiency $0 < \eta < 1$. Because the second factor equals 1 for

$$\sin\theta = q\lambda/\Lambda$$
$$q = \ldots, -2, -1, 0, +1, +2, \ldots \tag{53}$$

this diffraction grating emits diffraction orders indexed by q, that is, beams, the angles θ of which vary with frequency, except the zeroth-order beam $q = 0$, the direction of which is constant. The available power is, of course, split among several orders if such exist, so that for any order the diffraction efficiency η_q is less than unity. If only one diffracted beam is desired, the grating should not "amplitude-modulate" the input wave. Gratings that modulate the phase of the incident wave front are more efficient.

The angles θ, at which a particular diffraction order is emitted for a planar diffraction grating, do not depend on the in-plane position of the grating. Therefore, to suppress the zeroth-order beam, two gratings emitting at opposite phase can be interdigitated. If the line spacings are then chosen to be equal to $\lambda/2$, the first-order beams are emitted in the plane of the grating ($\theta = 90°$). This can be accomplished with metal electrodes on one surface of a piezoelectric medium and is the basic idea of the interdigitated transducer for surface waves.

The diffraction grating affords the possibility of scanning a sound beam over a range of angles as its frequency is varied. This has found application in dispersive delay lines, the delay time of which varies with signal frequency.

A sound wave traveling through a transparent medium can modulate the refractivity $n = v_{10}/v_1$ of the latter via the elasto-optic effect. It thus acts as a moving phase grating, the line spacing of which, the sound wavelength Λ, can be varied by the frequency of the sound beam. If a light beam of wavelength λ impinges on this grating, it can be switched and scanned by an appropriate amplitude and frequency modulation of the sound beam. This is the principle of acousto-optic light modulation and deflection.

VI. PIEZOELECTRIC TRANSDUCERS

Piezoelectric transducers can assume many shapes, including cylinders and spheres (e.g., for hydrophones or other electroacoustic devices), in which piezoelectric ceramics can be used. In contrast, the transducers in AWDs tend to be bars or plates with mostly plane faces, since these can be readily fabricated in crystals by standard sawing, grinding, and polishing methods. A designer of an AWD presented with specifications of the frequency band of operation would have to select among configurations such as those shown in Fig. 10, select material and crystal cut, and determine dimensions and then the resulting

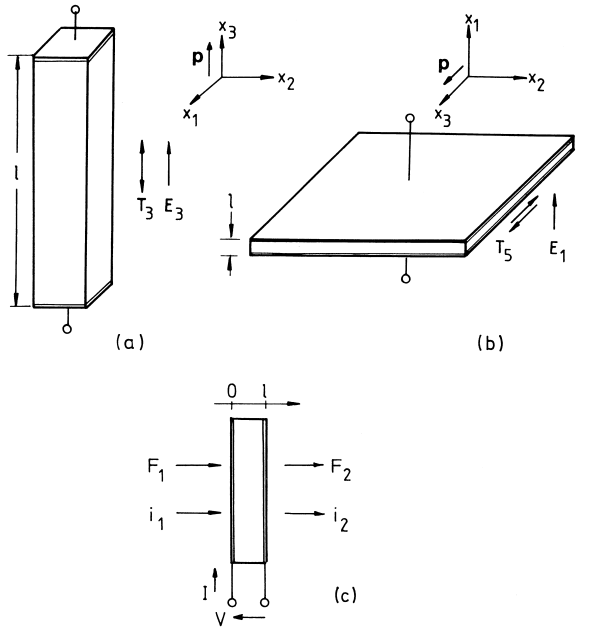


FIGURE 10 Basic piezoelectric transducers. (a) In-line field, length extension. (b) In-line field, thickness extension or in-line field, thickness shear configurations. Subscripts are appropriate for materials of crystal class 6*mm*. (c) Terminal variables for a selected configuration.

electrical data, such as input and output impedance and transmission loss. These, the designer would find, depend also on electrode layers and the method of bonding the transducer to the transmission medium and, of course, on the data of the latter.

In view of the exceedingly large number of material constants entering the problem, starting from first principles is difficult. Better insight results from evolving a procedure to describe the transducer as a "black box," the terminals of which are then connected to further "black boxes" representing the outside world to which the transducer is connected. This description yields in appropriately simplified cases an equivalent circuit, which can be inserted into the circuit simulation programs and aids electronic designers are familiar with.

To this end the terminal parameters of a transducer have to be determined, that is, the data that can be measured at its various interfaces with the outside world. These can be obtained from solutions of Eq. (44) with appropriate boundary conditions for the case at hand. For a piezoelectric plate of thickness l and cross-sectional area A, relations result between voltage V and current I at the electrodes, and the forces F_1 and F_2 and displacement velocities i_1 and i_2 at the faces (Fig. 10c), the latter being defined for the relevant subscripts m and i by

$$\begin{aligned} F_1 &= AT_m(0); \qquad F_2 = AT_m(l); \\ i_1 &= \partial u_i(0)/\partial t; \qquad i_2 = \partial u_i(l)/\partial t \end{aligned} \tag{54}$$

Relations of this form can be developed if essentially only one type of sound wave interacts with the transducer.

For frequencies much above 1 MHz, only the plate configurations in Fig. 10b with the electric field **E** between the major surfaces are practical. To produce longitudinal waves choose the inline field, thickness extension configuration; for transversal waves choose the in-line field, thickness shear configuration. In crystals of hexagonal class 6*mm* and piezoelectric ceramics poled in the direction of the crystallographic x_3 axis, the driving field component E_3 produces longitudinal wages with **k** also parallel to the x_3 axis. With major faces normal to and **E** parallel to the x_1 axis, the plate produces transversal waves with **u** parallel to the x_3 axis and **k** parallel to the x_1 axis. Thus, we arrive at the orientations in Fig. 10b.

With the following abbreviating definitions, for longitudinal waves,

$$\begin{aligned} &\varepsilon = \varepsilon_{33}^S; \quad c'_{33} = c_{33}^E + e_{33}^2/\varepsilon; \\ &K = e_{33}(c'_{33}\varepsilon)^{-1/2}; \quad l = l_3; \quad v = (c'_{33}/\rho)^{1/2} \end{aligned} \tag{55}$$

and for transversal waves,

$$\begin{aligned} &\varepsilon = \varepsilon_{11}^S; \quad c'_{55} = c_{55}^E + e_{15}^2/\varepsilon; \\ &K = e_{15}(c'_{55}\varepsilon)^{-1/2}; \quad l = l_1; \quad v = (c'_{55}/\rho)^{1/2} \end{aligned} \tag{56}$$

dielectric and elastic losses can be included in this model by entering the appropriate complex numerical values for ε and v. Once the terminal equations are developed, an equivalent circuit having the same equations can be found.

Mason's equivalent circuit (MEC) shown in Fig. 11a is widely used with component values for the plate given by

$$\omega_0 = \pi v/l; \qquad \gamma = \pi\omega/\omega_0 = \pi f/f_0; \qquad C_0 = A\varepsilon/l \tag{57}$$

$$Z_0 = A\rho v; \qquad N = K(\omega_0 C_0 Z_0/\pi) \tag{58}$$

$$Z_A = jZ_0 \tan(\gamma/2); \qquad Z_C = jZ_0/\sin\gamma \tag{59}$$

The electromechanical coupling factor K is an important figure of merit for a piezoelectric material.

The equivalent circuit for nonpiezoelectric layers can now be obtained by omitting the transformer $1:N$ and connecting the bottom terminal of Z_C directly to node 0, resulting in Fig. 11b. Layers of "infinite" lengths, that is, those that do not return reflections from the other end, are represented by a simple resistance $Z = A\rho v$, as in Fig. 11c. In this way a stack of piezoelectric and inactive layers, such as a transducer with electrodes, bonding layers, and possibly a load on its back face (Fig. 12) can now be modeled as an electrical circuit, as can more complicated systems of multiple piezoelectric layers and surface wave transducers.

Now standard techniques for circuit analysis can be used to determine the frequency characteristics of input admittance Y_i, insertion loss TL, and soon for transducer configurations driven by a source with impedance R_s,

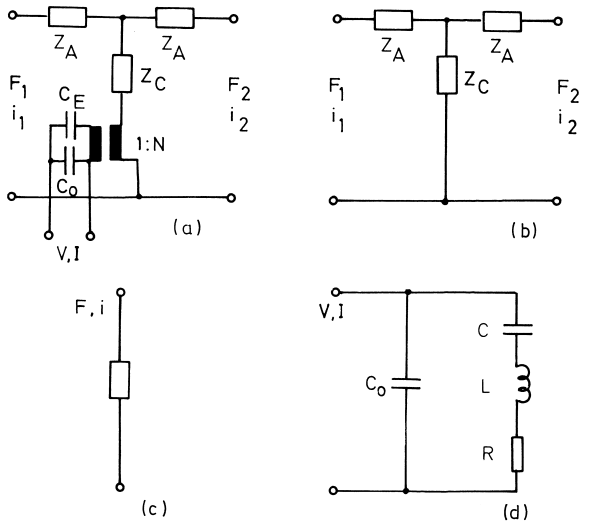


FIGURE 11 Mason's equivalent circuit (a) for in-line field, thickness extension or shear configurations of a piezoelectric plate; (b) for a nonpiezoelectric layer; (c) for a nonreflecting medium. (d) Approximation of (a) for an unloaded transducer (Cady's equivalent circuit).

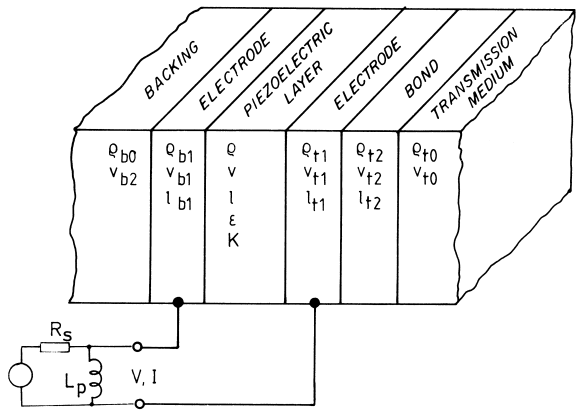


FIGURE 12 Stack of inactive layers around a piezoelectric one representing a transducer of a delay line. The first subscript indicates the backing or transmission side, the second the layer number counting outward from the piezoelectric one; the outermost ones have "infinite" length. [Adapted from Sittig, E. K. (1972). In "Physical Acoustics" (W. P. Mason and R. N. Thurston, eds.), Vol. 9, pp. 221–275, Academic Press, New York.]

possibly with an electrical matching network, in the simplest case a parallel tuning inductance L_p. To remove unnecessary parameters it is convenient to normalize the input and output variables to some of the piezoelectric layer. It is convenient to chose as input variables

$$\begin{aligned} z_{ab} &= A_{ab}\rho_{ab}v_{ab}/A\rho v; & t_{ab} &= vl_{ab}/v_{ab}l; \\ r_s &= R_s\omega_0 C_0; & x_p &= \omega_0^2 C_0 L_p \end{aligned} \tag{60}$$

the subscripts ab meaning the ones in Fig. 12.

The example of Fig. 13 is computed with source resistance $r_s = 1$, no L_p, and an infinitely long transmission medium of impedance ratio $z_{to} = 1$. The backing impedance has been chosen to be $z_{b0} = 0$. One layer with $z_{t1} = 0.1$ and t_{t1} is interposed between transducer and transmission medium. As t increases from zero, the frequency response is distorted markedly. Since the impedance ratio z_{t1} approximately represents a bonding material like epoxy resin, it becomes apparent that at high frequencies such bonds have to be made extremely thin or a material with $z_{t1} = z_{t0}$ has to be used to suppress the influence of a bonding layer where necessary in the fabrication of AWDs.

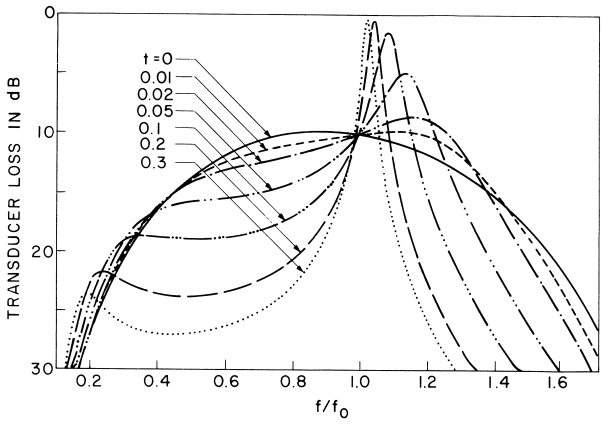


FIGURE 13 Computed transducer loss of an unbacked transducer layer with $K = 0.2$, $z_{t0} = 1$, $r_s = 1$ and a bonding layer interposed between transducer and transmission medium having $z_{t1} = 0.1$ and t_{t1} as parameter. [Adapted from Sittig, E. K. (1972). In "Physical Acoustics" (W. P. Mason and R. N. Thurston, eds.), Vol. 9, pp. 221–275, Academic Press, New York.]

The following general features emerge from such computations:

1. The transducer frequency response is that of a bandpass with maximum response near $f = f_0$. There is zero response at even multiples of f_0. There are higher passbands near odd multiples of f_0. The higher passbands are useful if the transducer is used as a resonator, that is, with $z_{b0} = z_{t0} = 0$, but for wide-band application the fundamental band gives lower loss and higher relative bandwidth.
2. The bandwidth increases with z_{b0} and z_{t0}, but the frequency response becomes double-humped if these assume values near 2 and larger.
3. The transducer loss decreases with increasing K. Because K in a piezoelectric material usually is larger for transversal waves than for longitudinal waves, the use of the former is to be preferred where a choice exists.
4. Matching to source and load impedances R_s, R_1 of external circuitry can be done at high K without tuning networks and $r = 1$. At low K, loss can be traded in against bandwidth if tuning networks are used, for instance, parallel inductances for which $x_p = 1$ and increasing r to match the real part of the input impedance of the transducer.
5. Electrode and other intermediate layers have marked influence on the frequency response unless they have very small t or their impedances are matched to that of either transmission medium or transducer.

Table III shows data that can be inserted into MEC computations for some transducer materials in current use. The symbols in the column labeled "Cut" give the direction of **E** in terms of the crystal coordinate axes, here X, Y, Z and the angles at which the plate is cut.

For application at frequencies above 100 MHz a piezoelectric material should thus have high K, high v, low ε, and technological properties allowing the fabrication of plates or layers with a thickness of a few micrometers. Plate cuts for single-mode operation should be available. The materials listed in Table III satisfy these conditions adequately, although most ferroelectric ceramics have high permittivities and thus require excessively low interconnection impedances, at high frequencies.

TABLE III Mason's Equivalent Circuit Data of Some Transducer Materials

Transducer	Cut	Mode	$\varepsilon/\varepsilon_0$	K	v (mm μ sec^{-1})	ρ g cm^{-3}
Quartz	X	L	4.58	0.098	5.70	15.1
	Y	T	—	0.137	3.80	10.1
Cadmium sulfide	0	L	9.5	0.15	4.50	21.1
	40	T	—	0.24	2.11	10.1
Zinc oxide	0	L	8.8	0.28	6.40	36.4
	40	T	—	0.38	3.32	18.3
Lithium niobate	Y/35	L	39	0.49	7.35	34.8
	Y/−20	T	40	0.62	4.52	21.12
Sodium potassium niobate	Ceramic	L	300	0.45	7.10	31.4
		T	—	0.65	3.68	16.6

VII. PIEZOELECTRIC RESONATORS

A resonator, in this cotext, is any device in which a wave can travel back and forth between reflectors. At frequencies above ~1 MHz a convenient implementation (Fig. 2) is a plate of piezoelectric material, crystal or ceramic, as the transmission medium for sound waves, with force-free end faces at which the sound waves, with force-free end faces at which the sound wave is reflected. At lower frequencies, nonpiezoelectric bar or rod-shaped media are used, the amplitude of which is coupled to separate transducers. If successive reflections of a single-frequency wave arrive at the end faces in phase, they sum to a large amplitude. This occurs at frequencies f_0 for which the length l of the travel path obeys the relation

$$l = q\lambda/2 = qv/2f_0; \qquad q = 1, 2, \ldots \tag{61}$$

and can be used for frequency selection (filtering). Thus, a wave or distributed resonator also has overtone responses, in contrast to a lumped resonator, in which energy exchanges between two stores such as a mass-spring pendulum or the like. In a three-dimensional body such as a rectangular thick plate, each pair of sides can act as the reflector pair, and many guided modes can propagate within the body. The resonator then has a complicated spectrum of resonances.

A freely vibrating piezoelectric plate can be modeled by Mason's equivalent circuit, if its vibration modes are not coupled with each other. In this case one enters $z_{t0} = z_{b0} = 0$. It is, of course, impossible to define a transducer loss, but the behavior of its input impedance $Z_i = V/I$ can be observed. One then finds

$$Z_i = (1/j\omega C_0)[1 - K^2 \tan(\gamma/2)/(\gamma/2)] \tag{62}$$

an expression assuming a very large value—limited by loss—at the frequency $f = f_0$ and a very small value at $f = f_1$, for which

$$K^2 \tan(\gamma/2)/(\gamma/2) = 1 \tag{63}$$

Thus, $f_1 < f_0$, with the difference depending on K. The losses mentioned arise from those intrinsic in the material, but also from those introduced by having to mount the crystal in a support, which may damp its vibration. To obtain a resonator with high quality factor Q, the mounting has to minimize this damping.

The electrodes at which Z appears can be modeled in Fig. 12 as additional layers with z_{b1}, t_{b1} and z_{t1}, t_{t1} and $z_{b0} = z_{t0} = 0$. They thus present imaginary impedances to the faces of the piezoelectric plate. Computation shows that thin layers f_0 and f_1 appear to be lowered a small amount by the mass loading of these layers. A resonator can thus be tuned precisely to a desired frequency by making its f_0 slightly too high (i.e., l too small) and then adjusting the electrode layer thickness by plating or etching until the desired frequency is reached.

In the vicinity of these frequencies the impedance of a narrow-band crystal resonator can be represented by Cady's equivalent circuit (Fig. 11d). Its component values are chosen to fit MEC at f_0 and f_1, and R models the losses. While simpler to discuss, its validity is restricted to narrow-frequency bands. Nevertheless, it is useful to circuit designers who want to obtain prescribed frequency response curves for filters in which crystal resonators are embedded as components of electronic circuits.

A practical resonator must fulfill the following:

1. The resonance frequency f_0 should be independent of temperature and other environmental variables and should be stable with passing time.
2. Side-resonance suppression should be high; that is, there should be no additional resonances near the desired one with amplitude sufficient to cause its oscillator circuit to "jump modes" (to change its frequency spontaneously to the adjacent one).

3. The bandwidth Δf over which the resonance amplitude stays within at most 3 dB below the maximum should be as narrow as possible, or the quality factor $Q = f_0/\Delta f$ should be as high as possible. This enhances stability against changing driving conditions of the accessory electronic circuitry.

As to the first requirement, temperature independence of f_0 means that thermal expansion of the resonator dimensions and temperature dependence of sound velocity cancel each other. In crystals of the trigonal system, such is the case in certain directions oblique to the crystal coordinate axes. Plates cut with faces perpendicular to those directions are known for quartz as AT, GT, and SC cuts. Other environmental effects, such as corrosion and surface deposits, can be suppressed by encapsulation. Aging due to material migration, diffusion, recrystallization of metal layers, or stresses in mounting supports can be suppressed by proper technological procedures in fabrication.

The second and third requirements are interrelated. An undamped plate, like any three-dimensional body, can resonate in any of its three dimensions in various modes (i.e., transversal, longitudinal, flexure), each with a not necessarily harmonic spectrum of overtones. Each resonance frequency has associated with it a standing wave with an amplitude distribution possessing local maxima and minima. In the latter, termed "nodes," the mounting supports have to be attached to the crystal since there they impart minimal damping. Losses of vibrational energy into the supports as well as radiation of airborne sound tend to exceed the intrinsic losses of the resonator material by far and thus limit the attainable Q. Intrinsic losses in, say, quartz would permit a $Q > 10^7$ at frequencies $f <$ 10 MHz.

Designing the supports to be in nodes for the desired resonance at f_0 meets with limited success. First, if f_0 exceeds 10 MHz, the plate has a thickness below $\sim$200 μm and problems arise in placing the nodal supports accurately enough. Second, the support may or may not damp any particular "spurious mode," depending on whether it is near one of its nodes. Third, various modes can be coupled to the desired one in the material itself or at the boundary thereof so that damping one will also extract energy from the others.

Such coupling effects are ubiquitous. If we write Eq. (24), for instance, as a 9×9 matrix equation between T_m and D_i in one side and S_n and E_j on the other, then any off-diagonal material constant qualifies as a coupling constant. Even an isotropic material expanding in one direction contracts in the lateral ones. Piezoelectric constants serve equally well. Quantitative relations can be obtained by solving Eq. (44) for realistic boundary conditions, an undertaking that strains even today's computational power in view of the three-dimensional treatment needed.

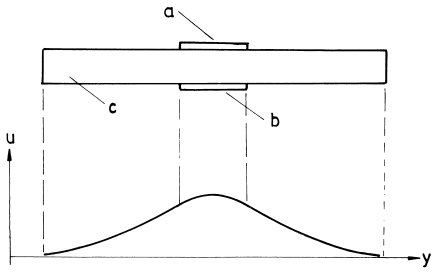


FIGURE 14 Displacement amplitude along a lateral coordinate y of a piezoelectric resonator c with electrodes a, b extending over part of its surfaces.

An important concept has accelerated progress: If a plate or bar is only partly covered with electrodes (Fig. 14), the resonance frequency is lower there than in the uncovered areas. At the lower frequency the multiple reflections do not arrive in phase at the uncovered faces and thus do not sum up to the resonance amplitude. As a result the amplitude decreases outside the covered area and the vibration energy appears "trapped" underneath the electrodes. Since now the outermost regions of the resonator vibrate with reduced amplitude, the resonator can be held there with little damping by the support. At the same time unwanted contour modes (resonance frequencies due to lateral resonances near f_0) can be brought under control. This principle of energy trapping is an essential tool in modern resonator design.

Energy trapping, the confinement of strong vibration to areas on a plate delineated by plating or etching, permits the fabrication of several resonators on a single plate (Fig. 15). Since the amount of mechanical coupling can be varied by the distance of the individual areas, and electrically by interconnecting the individual resonators, monolithic filters can be concieved and built without recourse to external components. They combine an approximately flat passband of several kilohertz at center frequencies of several megahertz with adjacent channel interference suppressed by 80 dB, a feat difficult to achieve by other means. At frequencies lower than a few megahertz the plate dimensions needed to accomodate several resonators become excssively large. Single piezoelectric resonators

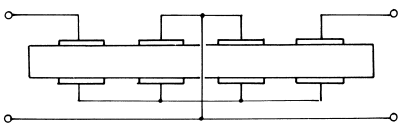


FIGURE 15 Monolithic filter with four elements and mechanical as well as electrical coupling.

interconnected by electrical networks or individual, mechanically coupled resonator bars are then easier to implement. At frequencies above 50 MHz SAW devices can take over.

VIII. BULK WAVE DELAY LINES

Delay lines used simply to delay a signal by a specified delay-time T should have the following attributes:

1. A bandwidth large enough to pass the complete frequency band of the signal with linear phase response, to avoid distortion in the time domain.
2. A minimum insertion loss at midband to avoid impairing the dynamic range of the signal or its signal-to-noise ratio.
3. Freedom from electrical cross-talk and spurious signals due to unwanted reflections.
4. Adequate temperature and time stability to keep data within specified timing tolerances.
5. Simple structure and fabricability, small size, and compatibility with its external circuitry, to keep costs down.

Items 1–3 are interrelated: A delay line with two transducers, each with transducer loss TL in decibels, an absorption loss AL in the delay medium, and beam spread loss BL, has the overall loss

$$\mathrm{DL} = 2\mathrm{TL} + \mathrm{AL} + \mathrm{BL}. \tag{64}$$

The TL can be computed from the equivalent circuit of a transducer, as indicated in Section VI. The frequency response should, according to theory, have linear phase response and an amplitude response of the shape

$$\mathrm{DL} = -20\log\left[\sin^2(xf/2f_0)\right] + \text{const} \tag{65}$$

to provide the shortest possible response to a short unipolar input pulse. With the overall response of Eq. (65), digital pulses can be transmitted through the delay line at the rate f_0 without excessive intersymbol interference (disturbance from the precursor and successor pulse). Such a response can be obtained approximately with a transducer–transmission medium combination for which $z_{b0} \approx 1$, K small, and $z_{t0} \approx 1$, that is, an absorbing backing layer. That, however, produces a rather high TL since little electrical power is converted to sound or extracted at the receiver. Backings lower the conversion efficiency. With K near 0.6 and z_{t0} around 0.6, backings can be avoided and an adequate frequency response with a TL at midband of only a few decibels can be achieved, albeit with higher reflections from the transducers at off-center frequencies.

The most important spurious signal is the triple travel signal appearing with a dealy, 3T. Part of the sound power impinging on the receiving transducer is reflected from it, travels back to the transmitting transducer, is re-reflected there, and returns to the receiver. Backingless transducers reflect strongly at off-center frequencies where TL increases. To suppress this signal by a specified amount, one can deliberately design AL and BL into the transmission path by choosing a dealy medium with an appropriate absorption coefficient and an appropriate transducer diameter. If this loss amounts to, say, 10 dB, the triple travel suppression will be 20 dB below the nominal response. If this is insufficient, a small tilt angle θ can be put between the major planes of the transducers: as the reflection angle doubles after each reflection, the incidence angle for the triple travel path is then 4θ. The angle can thus be chosen to place the triple travel incidence into a null in the directivity of the receiver while introducing only a slight additional tilt loss to the main signal.

Adequate suppression of electrical cross-talk between transmitter and receiver is achieved by standard shielding measures such as grounded metallization on the side walls and suitably designed packaging. Reflections from the side walls of the delay medium impinging on the receiver may also cause spurious signals. These can be suppressed by covering the side walls with sound-absorbing material and making the lateral dimensions of the delay medium so large that reflections arrive with an incidence angle large enough to be suppressed by the directivity of the receiver.

A material absorption coefficient α usually expressed in decibels per meter causes an AL over a path length x of

$$\mathrm{AL} = \alpha\mathrm{L} = (\alpha\lambda)x/\lambda \quad \mathrm{dB}. \tag{66}$$

But, α depends on frequency. The frequency dependence can be modeled once an absorption index $\alpha_0\lambda_0$ is known at frequency f_0 by

$$\alpha\lambda = (\alpha_0\lambda_0)(f/f_0)^{\zeta}; \qquad 0 < \zeta < 1. \tag{67}$$

In any case, the AL increases with frequency.

The BL depends on wavelength in the transmission medium and on the beam diameter, usually that of the effective diameter of the transducer. It can be obtained via Eq. (50). The effective transducer diameter is usually defined by the shape of its back electrode. If this is circular with a radius r for both transducers, BL can be adequately calculated from

$$\begin{aligned} \mathrm{BL/dB} &= x/x_0 \qquad \text{for} \quad 2 < x/x_0 \le 10 \\ &= 20\log(x/x_0\pi) \quad \text{for} \quad x/x_0 > 10, \end{aligned} \tag{68}$$

where

$$x_0 = r^2/\lambda. \tag{69}$$

Since $\lambda = v/f$, BL decreases with frequency. Thus, if AL = BL at midband, their effects on the passband approximately cancel. This can be made to happen by an appropriate choice of transducer diameter.

For many applications the time–frequency product M, sometimes called storage capacity, of a delay line must not change more than ~0.1. Thus, designs with large M made from transmission media with large temperature coefficients of delay may need temperature stabilization. This temperature coefficient contains contributions from thermal expansion and sound velocity. If it is, say, $10^{-4}/^{\circ}$C, a temperature variation of 1°C only can be tolerated. There is also, for most isotropic solids, a batch-to-batch variation of material data up to 0.1%. Thus, delay lines with adequate stability for $M > 1000$ may have to be trimmed individually, which is readily accomplished by lapping back a reflecting surface if one is available. Before the advent of fast and large integrated circuits, memory delay lines were built with bandwidths of 30 MHz and delay times of up to 1 msec, that is, $M = 30{,}000$. They consisted of multifaceted polygons made from fused quartz, an isotropic material made from melted quartz crystals, to accommodate a delay path of several meters without excessive AL. They required high precision of facet angles, had to be run in temperature-controlled ovens, and were extremely costly.

This practice has been superseded by other methods. Today, manufacturers of optical glasses supply "isopaustic" glasses with much smaller temperature coefficients of delay but also much higher absorption unsuitable for building delay lines with M much exceeding 1000. With today's low-loss transducers one would use several simpler delay lines with intermediate integrated circuit amplifiers to obtain delay times of that order.

For use in pulse compression radar systems, bulk wave delay lines have been built with a delay time varying over a range ΔT linearly with a frequency variation over the range Δf. Both transducers carry electrode gratings for diffractive beam steering. Figure 16 shows one of several possible versions. A shorter transmitting transducer carries on its backside an array of parallel lines with varying spacing. Its first diffraction order beam scans over a larger, similarly equipped receiving transducer, the scan angle, and thus the effective delay path length varies with frequency. If the delay is to vary over a range ΔT if the frequency varies between f_1 and f_2, then the design prescription to have the directivities of the transducers track is as follows:

$$\begin{aligned} D &= \Delta T/(f_2 - f_1) \\ N_x &= 1 + \Delta T\left(f_1^2 - f_2^2/4\right)/2(f_2 - f_1) \\ N_y &= N_x + \Delta T(f_1 + f_2)/2 \\ x_1^2 &= y_1^2 = (f_1 v D)^2/2 - v^2 D(N-1) \\ x_n^2 &= x_1^2 + 2Dv^2(n-1); \qquad n = 1, \ldots, N_x \\ y_m^2 &= y_1^2 + 2Dv^2(m-1); \qquad m = 1, \ldots, N \end{aligned} \tag{70}$$

where N_x and N_y are the number of grating lines on the transducers and x_n and y_m are the individual line positions measured from the rectangular corner.

While the performance of these perpendicular diffraction delay lines (PDDLs) was unexcelled, their dimensions were rather large and their manufacture correspondingly expensive, if large $\Delta f\,\Delta T$ values, say of the order of 1000, were needed. Because of the long transmission path lengths, it was necessary to use fused quartz, which has low absorption but a relatively large temperature coefficient of delay. With anisotropic media, sound beams with varying directions are, however, difficult to use because of the complication outlined in Section IV. Today the less bulky SAW devices are available for this application.

Table IV presents some materials used in contemporary bulk delay lines. Where absorbing backings are necessary, sintered metal powder bodies or epoxy resin loaded with metal powder offer themselves since their ρv values can be tailored to measure.

Bulk wave delay lines for frequencies up to 10 MHz compete today with semiconductor memory devices for digital signals. In analog signal and precision timing applications and at frequencies above 100 MHz, they excel. With deposited thin-film transducers and crystalline delay media, frequencies greater than 1 GHz are available.

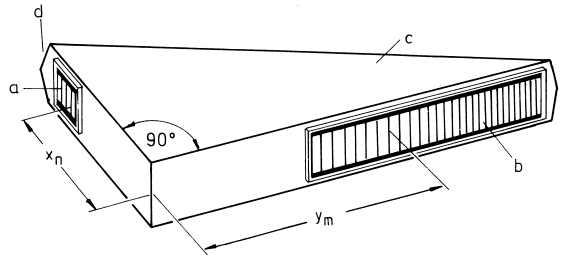


FIGURE 16 Perpendicular diffraction delay line. (a, b) Transducers; (c) delay medium; (d) bevel to suppress spurious reflections. [Adapted from Sittig, E. K., and Coquin, G. A. (1970). "*IEEE Trans. SSAICS-Ultrasonics*," SU-17(1/70), 23–29.]

IX. ACOUSTO-OPTIC LIGHT DEFLECTORS AND MODULATORS

In an acousto-optic light deflector (ALD), a sound beam of wavelength Λ forms a diffraction grating for a light beam of wavelength λ. Accordingly, variation of Λ via the sound frequency varies the diffraction angle θ of each light diffraction order.

The sound beam modulates the components B_{ij} in the light wave equations [Eqs. (47)] via the elasto-optic tensor

TABLE IV Sound Propagation Data for Delay Line Materials

Material	v_L (m sec^{-1})	v_T (m sec^{-1})	ρ (g cm^{-1})	$\alpha_L\lambda_L$ (10^{-3} dB)	$\alpha_T\lambda_T$ (10^{-3} dB)	At f_0 (MHz)
Fused quartz	5960	3760	2.20	1.2	1.13	100
T-40 glass	4310	2580	3.38	—	7.7	100
Aluminum	6420	3040	2.70	0.1	—	1000
Gold	3240	1200	19.3	0.07	0.09	1000
Indium	2300	1440	7.39	18.0	14.5	1000
Epoxy resin	2600	1220	1.10	Very high	Very high	Very high

(p_{ijkl}). The result is a sinusoidally varying phase grating, which moves with the sound velocity v_s along the direction of the wave vector **K** of the sound beam. A snapshot of a light beam with wave vector **k** traversing the sound beam laterally would show the plane wave fronts before traversal corrugated with a phase excursion $\Delta\phi$, which depends on strain amplitude S. Assuming that suitable values for directions and components have already been determined, the tensor subscripts can now be omitted.

Figure 17 shows the schematic arrangement. Let **K** and $\mathbf{k}_i$ be the wave vectors of the sound and incident light beam, respectively, with their magnitudes being

$$K = 2\pi/\Lambda = 2\pi f/v_s; \qquad \mathbf{k} = 2\pi/\lambda. \tag{71}$$

Equation (53) generalized for wave vectors $\mathbf{k}_i, \mathbf{k}_d, \mathbf{K}$ gives the angles θ for the diffraction orders q for the incidence angle θ_i:

$$\mathbf{k}_d = \mathbf{k}_i + q\mathbf{K}; \qquad q = 0, 1, 2, \ldots. \tag{72}$$

For high diffraction efficiency, the incident light power should emerge in one diffraction order only, and thus

$$\eta_q = E_q^2/E_i^2 \tag{73}$$

should be a maximum for just one order, say $q = 1$. The sound power P to obtain a prescribed $\Delta\phi$ is also of interest. with the definitions

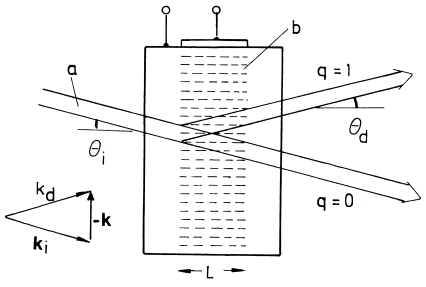


FIGURE 17 Diffraction orders of a "thick" phase grating produced by a traveling sound wave b for a light beam a in an elasto-optic material.

$$\begin{aligned} \sin\theta_B &= K/2k = \lambda/2\Lambda; \quad \psi = \Delta\phi/\cos\theta_i \\ Q &= 2KL\sin\theta_B/\cos\theta_i; \\ \alpha &= \sin\theta_i/(2\sin\theta_B) \end{aligned} \tag{74}$$

θ_B being the Bragg angle, we find that indeed only one diffracted order $q = 1$ can be had, if $Q \gg 1$ is chosen; that is, a "thick" phase grating is used. The diffraction efficiency for the first order is

$$\begin{aligned} \eta_1 &= (\psi/2\sigma)^2 \sin^2\sigma; \\ \sigma &= [Q^2(1-2\alpha)^2 + \psi^2]^{1/2}/2 \end{aligned} \tag{75}$$

and thus can attain unity for

$$\theta_i = \theta_d = \theta_B; \qquad \Delta\phi = \pi\cos\theta_i. \tag{76}$$

A further result is that the output light beam frequency v is shifted to

$$v_1 = v \pm f; \qquad v = v_1/\lambda \tag{77}$$

with the sign depending on whether the light beam traverses the sound beam "upstream" or "downstream." Since $v \approx 10^5 f$, this is a small change for most purposes.

To find the sound power of Eq. (41) to produce the necessary $\Delta\phi$, the directions of **k** and **K** in terms of the crystal coordinates have to be selected. Thereupon subscripts can be dropped. For the selected configuration we define the refractivity of the unstressed material as

$$n_0 = v_{10}/v_1; \qquad v_{10} = 2.998 \times 10^8 \text{ msec}^{-1} \tag{78}$$

and obtain, λ_0 being the light wavelength *in vacuo* and the sound beam cross-section LH.

$$\Delta\phi = \pi\left[2(L/H)PM_2/\lambda_0^2\right]^{1/2} \tag{79}$$

where all material data have been lumped into a figure of merit,

$$M_2 = n_0^6 p^2/\rho v_s^2. \tag{80}$$

It is, therefore, not just p but M_2 that has to be maximized in a selection of the deflector material.

TABLE V Data for Materials Used in Acousto-Optic Devices

Material	Water	Fused quartz	Lead molybdate PbMoO	Paratellurite TeO_2
Transparency range (μm)	0.2–0.9	0.2–4.5	0.4–4	0.35–4
Sound wave type	L	L	L(001)	L (001) S(110)
Light $\nsim$ **E, K**	$\parallel$ $\perp$	$\perp$	$\parallel$	$\perp$ Circularly polarized
v_L (mm μsec^{-1})	1.49	5.96	3.63	4.20
v_T (mm μsec^{-1})	—	—	—	0.616
n_0 at $\lambda_0 = 0.633$ μm	1.33	1.46	2.39	2.26
$M_2(10^{-15}$ s^3 kg$^{-1})$	126	1.56	86	34.5 (L (001)) 793 (S (110))
$M_3(10^{-11}$ msec^2 kg$^{-1})$	25	1.35	29.8	32.8 (L (001)) 110 (S (110))

From a practical point of view the dimensions of a deflector have to be chosen on the basis of the following considerations. The incident light beam of diameter W has a certain diffractive divergence angle β given by Eq. (50) or

$$\beta = R\lambda/W; \qquad 1 < R < 5 \tag{81}$$

with the resolution criterion R being a measure of how much light is tolerated outside that angle. Two adjacent positions of the exit beam are resolvable only if their angular difference exceeds β. The deflector thus has a number of resolvable positions N. With Eq. (81) this is

$$N = \Delta\theta_a/\beta = \Delta f\tau/R, \tag{82}$$

where τ is the access time of a position, that is, the time the sound beam with a new frequency needs to traverse the diameter of the light beam. Modulators, having to switch a light beam, are designed with $N = 1$, the undeflected beam being blocked off by a "stop."

With Δf thus determined, the band center f_0 should be chosen to avoid spurious beams in the desired range of deflection angles due to harmonics of the drive frequency or residual higher diffraction orders. This is the case for the choice

$$f_0 = 1.5\Delta f. \tag{83}$$

Now according to Eq. (79), the diameter of the sound beam in the direction of light propagation, the interaction length L, should be made as large as possible to get by with minimal sound power. However, if the incident wave vector $\mathbf{k}_i$ has fixed direction and the direction of $\mathbf{K}$ has to satisfy Eq. (72) as $\mathbf{k}_d$ varies its direction, the sound beam diameter L must stay small enough to provide a sound beam spread to cover the range over which $\mathbf{K}$ has to vary its direction. Thus, L has to be limited to

$$L = v_s^2/(\lambda f_0 \Delta f). \tag{84}$$

Combining all this by writing

$$\eta_1 = \sin^2(P/P_s), \tag{85}$$

we find that the reference sound power P_s for $\eta_1 = 0.7$ is

$$P_s = 3\lambda_0^3 \cos^2\theta_i N^2 R^2/\tau\pi^2 M_3$$
$$M_3 = M_2 n_0 v_s^2. \tag{86}$$

This value has been chosen, because above that limit, according to Eq. (75), the increase in η_1 with P flattens markedly.

Table V presents some data on deflector materials in current use. The optimized directions of sound and light amplitude vectors and the crystal cuts are indicated by conventional crystallographic notation, and the data pertain to these.

It is quite apparent that deflectors for large N may need sound powers of several watts for high diffraction efficiency. Thus, heating by sound absorption becomes a problem: Because refractivity is temperature dependent, inhomogeneous heating may cause optical wave front distortions, affecting image quality. Since the power requirement is a consequence of the limit on L by Eq. (84), various stratagems have evolved to lift it:

1. Using several transducers on tilted faces along the interaction length to combine their beam angles to a larger total.
2. Using a grating transducer designed to make its first-order sound beam scan over the required angle with frequency.
3. Using birefringent crystals, in which two light waves can propagate for a given direction of $\mathbf{k}$. Then $\mathbf{k}_d$ can be shifted to the other mode, thereby reducing the angular range $\mathbf{K}$ has to scan. A larger L can then be used.

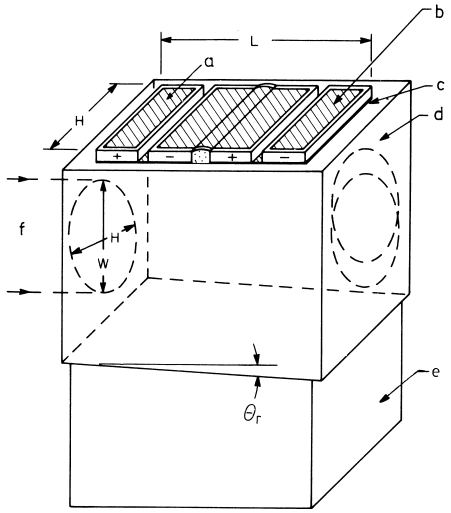


FIGURE 18 An acousto-optic light deflector. (a, b) Terminals of its transducer, which is subdivided into four series-connected segments for ease of impedance match; (c) elasto-optic crystal; (e) absorber of the sound beam, (f) entrance face for light beam. [Adapted from Sittig. E. K. (1972). In "Progress in Optics" (E. Wolf. ed.), Vol. 10. pp. 229–298, North-Holland, Amsterdam.

Figure 18 shows a deflector operating at sound frequencies of ~150 MHz and $N = 500$. The deflection medium is lead molybdate, and the transducers are made as series-connected subdivisions. For fast response a traveling sound wave is necessary since standing waves would have to involve multiple traversals. The sound beam must thus be absorbed after one passage through the deflection medium. The absorber has to be heat sinked to prevent appreciable temperature rise and is therefore made from metal. The interface between deflector medium and absorber is inclined by an angle so that reflections are sufficiently outside the angular range of operation not to cause any spurious beam deflection. The light entrance and exit faces have to be antireflection coated according to optical practice.

ALDs can be made for $N > 1000$. Rotating polygon mirrors and the like can attain higher N but not the speed of ALDs. Electro-optical devices can be faster but cannot at present reach the N of ALDs without becoming very complicated and expensive. Together with integrated and fiber optics, ALDs using surface waves are beginning to emerge.

X. SURFACE ACOUSTIC WAVE DEVICES

SAW devices make use of the surface sound waves mentioned in Section IV as signal carriers. In isotropic solids these waves propagate with the Rayleigh wave velocity v_R of Eq. (43) and the amplitude distribution shown in Fig. 7.

Since the power of the sound beam is concentrated within a layer of about a wavelength's depth below the surface, the power density becomes very high at high frequencies. It produces nonlinear effects at quite moderate power levels, which manifest as power-dependent absorption loss and the appearance of sums and differences of frequencies contained in the signal spectrum. In appropriate applications these effects can be put to good use. In addition, even at small power the absorption is very much dependent on the properties of a thin layer near the surface. With the surface finished to high optical standard, absorption for surface waves hardly exceeds that of transversal waves at equal frequency. Small scratches, roughness, or surface deposits such as condensation droplets and grease films can, however, increase the loss by many orders of magnitude. Thus, not only must the devices be properly encapsulated, but the propagating surface must also remain free of any contamination in process.

Crystals used as transmission media are anisotropic. Only in certain directions relative to the crystallographic coordinate system can surface waves be used and wave vector **k** and power propagation vector **P** be parallel. On the other hand, isopaustic directions are available and, if the media are themselves piezoelectric, transducers may be realized by simple electrode gratings deposited on the surface.

In contrast to bulk wave devices, where the sound path is accessible only at reflecting corners if such are provided, the entire sound path is accessible in SAW devices. Thus, guiding, reflecting, diffracting, and other modifying structures can be inserted into the sound path as can tapping transducers and the like.

A. Surface Wave Transducers

On a piezoelectric substrate a single pair of line electrodes on the surface can serve as a transducer for surface waves if the spacing of the lines equals $\lambda/2$ and the electric field between them can stimulate mechanical motion via some component of the piezoelectric tensor. However, one pair of lines alone has such a wide directivity that it hardly discriminates against bulk waves. It also has a very small electromechanical coupling factor K because part of the electrical field it creates when driven by an electrical source lies outside the substrate. It has a broad frequency response according to $z_{b0} = z_{t0} = 1$ if modeled by an MEC but a rather high transducer loss. The directivity and thus the mode discrimination is improved if more line pairs are electrically connected into a grating structure often referred to as interdigital transducer (Fig. 19a). These are readily seen to correspond to multilayer bulk transducers with many piezoelectrical layers. Figure 20 shows an MEC computation for $K = 0.2$, $z_{b0} = z_{t0} = 1$. As the

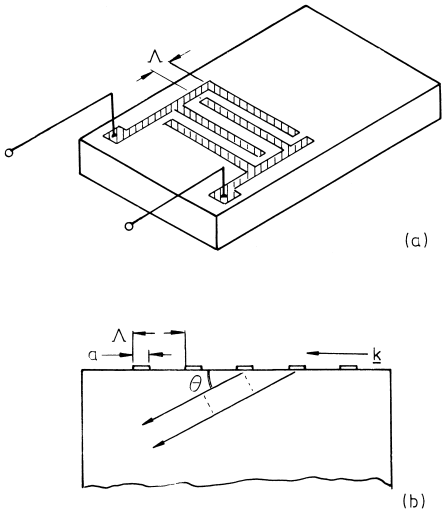


FIGURE 19 (a) Interdigital transducer grating on a piezoelectric substrate; (b) interference grating as reflector for surface waves or mode converter between surface and bulk waves.

number N of line pairs is increased, the transducer loss at midband decreases but its bandwidth narrows. In this case it is advantageous to use parallel inductors to tune the capacitance at midband according to Eq. (60), which is indicated by $x_p = 1$ in Fig. 20. In no case, however, can the transducer loss be below 3 dB since it emits surface waves from both ends of the array, of which only one side is used. It is possible to make these transducers unidirectional by using several interlaced arrays driven from sources so phased that their outputs add in one direction but cancel in the opposite one, but at the cost of larger complexity in processing.

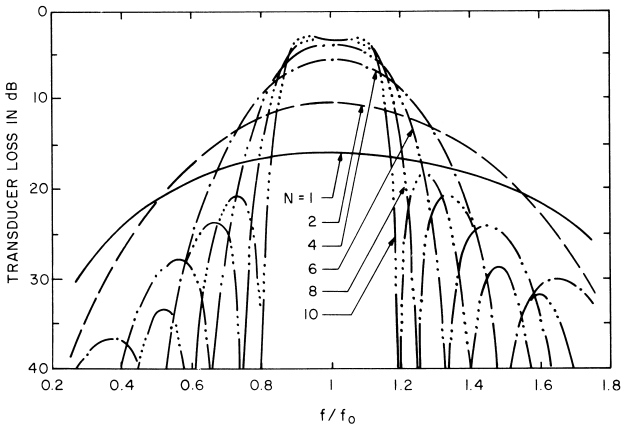


FIGURE 20 Mason's equivalent circuit computation of an interdigital transducer with $K = 0.2$ and parallel tuning inductance for various line pair numbers N; $r_s = 0.565$; $x_p = 1$. [Adapted from Sittig. E. K. (1972). In "Physical Acoustics" (W. P. Mason and R. N. Thurston, eds.), Vol. 9, pp. 221–275, Academic Press, New York.]

It is also evident that such a transducer has a narrow bandwidth and a rather high sidelobe level, that is, secondary response peaks near the main lobe. Improvement can be made if the spacing is varied along the length of the array and likewise the effective length of the "fingers." Since the outputs of the individual line pairs sum at the connecting "bus" when receiving a signal or in the sound beam when transmitting, a very general class of filter can be realized with this technique, the transversal filter in the terminology of filter theory.

B. Interference Filter

One important structure in the transmission path is a diffraction grating consisting of parallel lines of width a, length $b \gg a$, and line spacing Λ, in the form of metal deposited on the surface or shallow grooves etched into the surface (Fig. 19b). The lines may, but need not, be arranged perpendicular to $\mathbf{k}$. For $a \ll \Lambda$, each line acts as a radiator of sound waves with wide-angle downward directivity according to Eq. (50). If the grating is stimulated by a surface wave with wavelength λ_R incident from one end and passing along it, it will launch bulk waves of velocity v into the depth of the substrate at an angle θ given by

$$\cos\theta = (v/v_R)(\lambda_R/\Lambda - 1) \tag{87}$$

withdrawing power from the surface wave beam. Conversely, an incident longitudinal wave or a transversal wave with $\mathbf{u}$ perpendicular to the lines will launch a surface wave out of one end of the grating. This is one way of generating or receiving surface waves. At a frequency for which

$$\Lambda = \lambda_R/2, \tag{88}$$

Eq. (87) yields $\theta = 0$ and the incident surface wave is reflected. The grating thus acts as a frequency-selective reflector, an interference filter in optical terminology. Used in this mode of operation the grating can be modeled as a tandem connection of equivalent circuits according to Fig. 11b, with the lines being represented by appropriate values of z and t, somewhat different from the values for the gaps. If these values are known, the frequency dependence of the reflection factor with z, t, and the total number N of line-gap pairs as parameters can be computed. Line length and thickness can be modified for trimming purposes.

If the line spacing is varied along the length of the grating, a particular frequency contained in the signal is reflected wherever Eq. (88) holds. The reflection thus has a time delay that varies with frequency. In this way a dispersive delay line can be realized that functions much like the PDDL described in Section VIII. A short pulse signal passing along the grating produces reflections from

each line, which together form a longer signal train of pulses with variable spacings. In this way "coded signal sequences" can be realized.

Figure 21 a shows a nondispersive wide-band SAW delay line embodying the aforementioned principles. Different parts of the grating transducers are sensitive to different frequencies, but the travel distance L is the same for all. In Fig. 21b one of the gratings is reversed. This device accordingly has a delay time that varies with frequency. Another embodiment of a dispersive delay line (Fig. 21c) is the reflective array compressor used in chirp radar. Here, two interference gratings provide a reflection path with a distance from the transducers varying with frequency. Figure 21d shows the principle of a SAW resonator with interference reflectors. By arranging for two transducers between the reflectors, a high-Q two-port filter can be realized up to frequencies of 1 GHz.

If in the device of Fig. 21a the space between the gratings is extended to the length of the gratings and covered with a thin-film metal electrode, signals can be entered on both gratings. If signal strength suffices to produce nonlinear sound propagation, the convolution of the two signals can be obtained at the third electrode. To obtain the necessary high-power density at moderate sound power levels, the sound beams may be concentrated by triangular

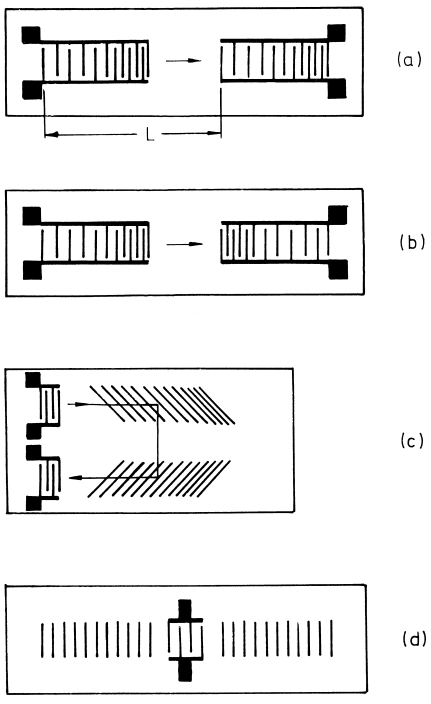


FIGURE 21 Surface acoustic wave devices. (a) Wide-band nondispersive delay line; (b) wide-band dispersive delay line; (c) wide-band linearly dispersive delay line using reflective array interference reflectors; (d) high-Q resonator using interference reflectors.

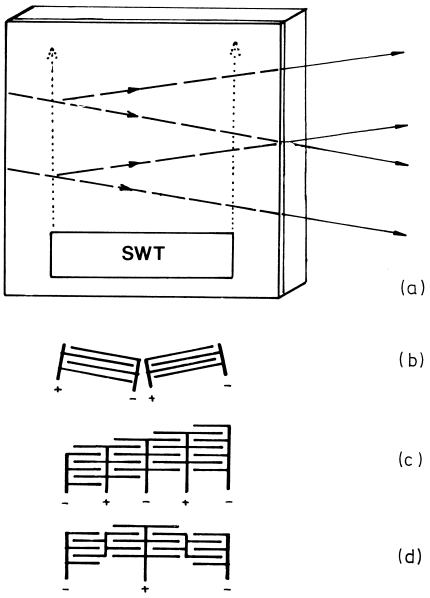


FIGURE 22 (a) Surface acoustic wave light deflector with light guide layer. The surface wave transducer may be that shown in (b), (c), or (d).

platings, with their apices leading the sound beams into a narrow electrode from which the convolution signal is extracted.

Finally, Fig. 22a shows a SAW light deflector that combines some of the aforementioned principles. The light must be guided near the surface by a guiding layer with lower light velocity than the adjacent layers. The diffracting sound beam is launched by an SAW transducer (SWT). As such, a single, wide interdigital one can be used or any of the configurations, (b), (c), or (d), namely, the tilted or scanning transducers, which implement the stratagems (1) and (2) mentioned at the end of Section IX, to overcome the limits on L.

XI. TECHNOLOGY FOR BULK AND SURFACE ACOUSTIC WAVE DEVICES

Modern microelectronics has developed a series of standard fabrication processes that yield inexpensive and reliable integrated circuits. AWD fabrication, to be economically competitive, follows this method of batch processing a large number of units simultaneously wherever possible. At frequencies approaching 1 GHz submicrometer size tolerances are necessary; these require precision fabrication techniques of optics or microelectronics. The following processes are directly applicable:

1. Crystal orientation by means of X-ray diffraction patterns; sawing plates or bars by means of diamond saws

with specified orientation of faces and finished to specified dimensions by grinding, lapping, and polishing.

2. Deposition of thin films or layers of metals and other materials by vacuum evaporation or ion deposition (sputtering).

3. Etching these films or their substrates with chemicals or ion beams to remove parts of the surface under precise control.

4. Cleaning processes to prepare surfaces to be worked on, to ensure proper adhesion of films, or to remove particulate contaminants that might impair surface quality.

5. Photolithography to define lateral structures such as subdivided electrodes. Here, the substrate is coated with a thin film of photoresist, a light-sensitive lacquer with the property of changing its solubility in a developer when exposed to light. Thus, exposure through a mask, a photographic plate carrying suitable patterns, transfers these on developments as windows into the resist film through which an etchant can attack the surface. In this way structures of submicrometer dimensions can be fabricated.

A. Bulk Wave Devices

To fabricate a bulk wave delay line or ALD the following procedures are required:

1. Fabricating the transmission medium to dimension and roughness tolerances of about $\lambda/10$. Since λ is $\sim 3\ \mu$m at 1 GHz, standard optical shop techniques are used. For ALDs the light entrance and exit faces have to be finished to the optical specifications the application demands. Antireflection coatings may be applied to light entrance and exit faces to enhance light transmission.

2. Depositing metal thin films on the faces the transducers are to be affixed to. These serve as ground electrodes. Suitable metals are aluminum or the combination of about 0.1 μm of chromium and 0.5 μm of gold, the former ensuring adequate adhesion of the gold to glass or crystal surfaces.

3. Fabricating the transducer chips (plates) with a thickness the application demands but at least 0.2 mm for ease of handling. Orientation as specified. The side to be bonded to the transmission medium has to be polished and a thin chromium or titanium film applied to it to promote adhesion.

4. Bonding the transducers to the transmission medium, taking care to maintain proper orientation. If for transversal waves the directions of **u** are perpendicular to each other, the receiver will be insensitive to the transmitter's signal.

5. If the resonant thickness l is to be less than 0.2 mm, lapping the transducer thickness down to required dimension, the transmission medium to which the transducer is bonded giving adequate support.

6. Depositing the back electrode by, for example, evaporation through a thin sheet metal mask to define the area. If the electrode is to be subdivided into a grating, this can be done by means of photolithography.

7. Bonding electrical connections to the back electrodes and the ground electrode, mounting, and encapsulation.

The bonding procedure (4) is the most critical operation. An epoxy resin bond, while suitable in principle, has to be made extremely thin ($<0.1\ \mu$m), as was pointed out in Section VI. Bonding can succeed only with meticulous attention to cleanliness, because globules and particulate contaminants in the metal layers and uncured resin act as spacers between transducer and transmission medium. The parts to be bonded have to be joined under high pressure to allow the excess resin to escape. If the starting viscosity is too high or the resin sets too fast, this cannot take place and the bond ends up being too thick. Using a heat-curing resin is no solution, because lateral strains in the bond resulting from thermal contraction after cooldown impair aging stability or, in ALDs, optical image quality.

Cold-welding methods involving soft metals such as indium or tin are less critical. The faces to be joined are coated by an indium film by evaporation and, without breaking vacuum, pressed together, whereupon the indium films cold weld at room temperature, making a satisfactory bond. Since these metals have a higher impedance ρv, thickness is not as critical. If removed from vacuum, the fresh metal surfaces oxidize, and subsequent bonding succeeds only at elevated temperature or under ultrasonic agitation.

At frequencies approaching 1 GHz the absorption loss even in these bonds becomes rather high. The gold thermocompression methods used in semiconductor technology function well but need temperatures of $>250^\circ$C, undesirable for the reasons mentioned. Steps 3–5 can be dispensed with altogether, however, if piezoelectric epitaxial layers are deposited as transducers. Some materials of the hexagonal class $6mm$, such as cadmium sulfide (CdS) and zinc oxide (ZnO), can be evaporated or sputtered on amorphous metal layers as polycrystalline layers with the crystallographic x_3 axis pointing in the direction of the evaporating source. With the x_3 axis "vertical," longitudinal waves can be obtained. With the x_3 axis tilted against the normal by 40°, transversal waves can be obtained. A problem with both materials is that they are semiconductors and can come out of processing with considerable electrical conductivity. A compensating posttreatment at elevated temperature may then be necessary.

B. Surface Acoustic Wave Devices

SAW devices are usually fabricated entirely on piezoelectric substrates, and thus fabrication starts with making the wafers from quartz or lithium niobate crystals in the specified orientation and as disks with 50-mm diameter, 0.5-mm thickness. The side to be processed further is highly polished, and the other side is rough ground to prevent reflection of spurious bulk waves therefrom. All further processing is done on one side only, so that a large batch of devices can be made concurrently, as is done in semiconductor technology.

A thin layer ($<0.1\ \mu$m) of titanium is deposited on the polished face of the wafer to promote adhesion of subsequent depositions. A photoresist film follows, and from a mask all patterns for all devices are photolithographed as windows in the resist. The entire face is coated with a layer of aluminum thinner than the resist coating. If the resist is now dissolved in a suitable solvent, the aluminum remains only in the windows, and etching of the aluminum film becomes unnecessary. If reflective gratings are to be etched into the transmission medium surface, this can be done by ion etching, using the deposited aluminum grating as the mask. A titanium–gold metallization is then evaporated through a thin sheet metal mask onto the areas serving as bonding pads for wires to the outside. Thereafter, the wafer is cut up into chips, which are individually mounted and encapsulated for protection.

If SAW devices are to be made on nonpiezoelectric transmission media, the process sequence is supplemented by a deposition of a thin layer of ZnO before the deposition of the aluminum layer. Thus, individual transducers are formed underneath the grating lines. Their orientation must be suitable for longitudinal mode or transversal mode with **u** parallel to the propagation direction of the surface wave.

This process sequence is deceptively simple in that its tolerances are extremely demanding. Narrow grating spacing, breaks in the electrode lines, and shorts between lines conspire to impair or impede performance. Thus, precise process control and extreme cleanliness are required.

AWD components can be used as fast, volatile, and serial (first in, first out) signal storage devices for analog and digital signals. Due to the fixed relation between travel time and length of travel paths—that is, the sound velocity—they serve as filters, delay lines, and stabilizing elements in frequency generators and time bases unsurpassed by other implementations of similar simplicity. Their fabrication costs compare favorably with silicon integrated circuits where they can be made in comparable quantity, and they functionally complement silicon circuits. These can provide switched filters and delay lines at frequencies below, say, 10 MHz though not with comparable quality. The ubiquity of crystal resonators, wrist watches, and television receivers serve as examples that attest to this role in present-day electronics.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • ACOUSTICS, LINEAR • CRYSTAL GROWTH • CRYSTALLOGRAPHY • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, GENERAL • ULTRASONICS AND ACOUSTICS • WAVE PHENOMENA

BIBLIOGRAPHY

Ballantine, D. S., Jr., White, R. M., Martin, S. J., Ricco, A. J., and Frye, G. C., Zellers, E. T., and Wohltijen, E. T. (1996). "Acoustic Wave Sensors. In "Applications of Moder Acoustics" (R. Stern and M. Levy, Series eds.), Academic Press, New York.

Campbell, C. (1998). "Surface Acoustic Wave Devices for Mobile and Wireless," Academic Press, New York.

Campbell, S. A. (1996). "Science and Engineering of Microelectronic Fabrication," Oxford Univ. Press, London and New York.

Elliott, D. J. (1982). "Integrated Circuit Fabrication Technology," McGraw-Hill, New York.

Gerard, H. M. (1978). *Top. Appl. Phys.* **28**.

Gerber, E. A., and Ballato, A., eds. (1985). "Precision Frequency Control (Vol. 2): Oscillators and Standards," Academic Press, New York.

Gottlieb, M., Ireland, C. L. M., and Ley, J. M. (1983). In "Optical Engineering" (B. J. Thompson, ed.), Vol. 3, Dekker, New York.

Hashimoto, K.-Y. (2000). "Surface Acoustic Wave Devices in Telecommunications: Modelling and Simulation" (Engineering Online Library), Springer-Verlag, Berlin and New York.

"IEEE Standard on Piezoelectricity" (1984). *IEEE Trans. Sonics Ultrason.* **SU-31**.

Mason, W. P., and Thurston, R. N., eds. (1962–1982). "Physical Acoustics," Vols. 1–16, Academic Press, New York.

Mason, W. P., and Thurston, R. N., eds. (1990). "Physical Acoustics Principles and Methods," Vol. 19, W. P. Mason and R. N. Thurston (editors) Academic Press, New York.

Mattiat, O. E., ed. (1971). "Ultrasonic Transducer Materials," Plenum, New York.

Miller, R. L., Bailey, D. S., and Nothnick, C. E. (1992). "Acoustic Charge Transport: Device Technology and Applications" (Microwave Library), Artech House, Boston.

Milsom, R. F., Redwood, M., and Reilly, N. H. C. (1977). In "Surface Wave Filters" (H. Matthews, ed.), pp. 55–108, Wiley, New York.

Nye, J. F. (1976). "Physical Properties of Crystals," 5th ed., Oxford Univ. Press, London and New York.

Nye, J. F. (1976). "Physical Properties of Crystals," Oxford Univ. Press, London and New York.

Sittig, E. K. (1972). In "Progress in Optics" (E. Wolf, ed.), Vol. 10, North-Holland, Amsterdam.

Thompson, M., and Stone, D. C. (1997). "Surface-Launched Acoustic Wave Sensors: Chemical Sensing and Thin-Film Characterization," Wiley, New York.

Thurston, R. N. (1998). In "Physical Acoustics, Ultrasonic Instruments and Devices" (R. N. Thurston and A. D. Pierce, Series eds.), Vol. 23, Academic Press, New York.

Tichy, J., and Gautschi, G. (1980). "Piezoelektrische Messtechnik," Springer-Verlag, Berlin and New York.



Epitaxial Technology for Integrated Circuit Manufacturing

D. M. Hansen
T. F. Kuech
University of Wisconsin

I. Introduction
II. Substrates for Epitaxy
III. Vapor Phase Epitaxy
IV. Molecular Beam Epitaxy
V. Liquid Phase Epitaxy
VI. Conclusion

GLOSSARY

Epitaxy Growth of a single-crystal semiconductor material when the crystal structure of the underlying substrate controls the structure of a growing semiconductor film.

Liquid phase epitaxy (LPE) Epitaxy in which material in the liquid phase is used as the source of the growing film.

Molecular beam epitaxy (MBE) Epitaxy by physical deposition of material under ultrahigh vacuum conditions in which evaporation is used as the source of the growing film.

Substrate Material used as the epitaxial template for the growing film.

III–V semiconductor Semiconductor formed from elements in column III and column V of the periodic table (e.g., GaAs).

Vapor phase epitaxy (VPE) Epitaxy in which chemicals in the vapor phase are used as the source of the growing film.

I. INTRODUCTION

This article first addresses issues for substrates used in epitaxial growth. The substrate is critical to the epitaxial process, providing the template for the deposited layer. Next, the most common classes of epitaxial growth techniques are reviewed. The vapor phase epitaxial growth of silicon materials is by far the most common of these techniques in semiconductor processing. The techniques of liquid phase epitaxy and molecular beam epitaxy are alternatives to vapor phase epitaxy for specialized applications. Silicon (S:) growth for devices and integrated circuits dominates the bulk of the epitaxial processes. The

epitaxial growth of III–V semiconductor materials is critical to many specific classes of non-Si devices such as lasers and light-emitting diodes.

The active regions of modern devices and integrated circuits are typically thin layers of materials with tightly controlled properties. The chemical, physical, and structural properties of these layers as well as their interfaces can largely determine the device characteristics, particularly as the device dimension and layer thickness continue to shrink. Increasingly strict demands have therefore been placed on the manufacturing processes used to create these thin layers.

One process critical to the success of modern semiconductor devices is the deposition of high-quality structural layers on an underlying substrate. The process is termed epitaxial growth when the layer is single crystalline and its structure is controlled by or replicates the crystal structure of the underlying substrate. There are several epitaxial growth techniques and the specific technique employed depends on such factors as the desired film structure, electrical and optical properties, and structural quality of the interfaces between different layers. The goal of the epitaxial process is to deposit a layer with well-controlled structure, composition, and concentration of active impurities, or dopants, to obtain semiconductor materials suitable for device applications.

All epitaxial growth techniques have several common and basic considerations. In all cases, epitaxial growth involves the transport of material from a neighboring phase to the growing layer. Once the atomic constituents are placed on the surface, the atoms must adsorb and diffuse over the surface and find a suitable final site for incorporation. The limiting step, which can change with the technique, determines the specific growth characteristics. In all techniques, epitaxial growth is a nonequilibrium process. Epitaxial growth requires a driving force for material to form an epitaxial layer. This driving force is provided by an intentionally created supersaturation of the source material, or precursor, in a growth ambient. The degree of supersaturation in the system and the phase forming the growth ambient, or medium, distinguish the different epitaxial growth processes.

II. SUBSTRATES FOR EPITAXY

The goal of epitaxial growth is to create a layer of material with uniform and controlled thickness, composition, and electrical, chemical, and optical properties for fabrication of a device or an integrated circuit. The nature of the substrate is critical to achieving these goals since it provides the template for the subsequent epitaxial layer.

A substrate is typically a bulk-grown material oriented such that a low-index crystal plane, for example the (100) or (111) plane, is parallel to the wafer surface. Substrates can be fabricated by bulk crystal growth techniques for commonly used semiconductors such as Si. Czochralski and float-zone processes are well-established techniques capable of producing high-quality bulk Si substrates. When the substrate is the same material as the epitaxial layer, the growth is termed homoepitaxy. The widely used Si-on-Si epitaxy is an example of homoepitaxial growth for devices and integrated circuits. Homoepitaxy is capable of depositing smooth, uniform films and is the simplest form of epitaxial growth.

Heteroepitaxial growth is the deposition of materials that are structurally or chemically different from the substrate. Common examples of heteroepitaxy include growth of $Al_xGa_{1-x}As$ on GaAs, Si_xGe_{1-x} on Si, and GaN on Al_2O_3. Heteroepitaxial growth is more complicated than homoepitaxy due to the differences between the chemical and physical properties of the substrate and those of the film. For example, different materials may have different lattice constants, which is the interatomic spacing in the crystal. A difference in lattice constants can lead to the distortion of the atoms from their natural positions as the atoms in the film bond to and replicate the substrate structure in strict registry. This type of growth, wherein the film and the substrate have the same lattice spacing, is termed pseudomorphic growth. For materials such as $Al_xGa_{1-x}As$ on GaAs, pseudomorphic growth is easily achieved since these materials have similar lattice constants. When the difference in lattice constants is great, the distortion required by this strict registry of the epitaxial layer is large, whch leads to considerable elastic strain. In this case, there is a critical thickness beyond which the layer no longer replicates the substrate, instead relaxing to its natural lattice spacing through the formation of dislocations that can degrade the layer quality. The deposited layers must generally be thinner than the critical thickness for high-quality epitaxial layers to be deposited. Several important heteroepitaxial systems are Si_xGe_{1-x} on Si for electronic devices and GaN on Al_2O_3. GaN is used for blue and ultraviolet (UV) optical devices as well as for high-power electronics.

All growth initially occurs on a substrate surface. At the atomic scale, a substrate surface is not smooth and featureless, as it appears on the macroscopic scale. A microscopic view of the surface reveals a much finer structure, as schematically shown in Fig. 1. Real surfaces consist of structures such as terraces, ledges, and kinks. These features are energetically different, and atoms deposited on surfaces will generally tend to bond to the highest energy site. On the basis of the average number of bonds available

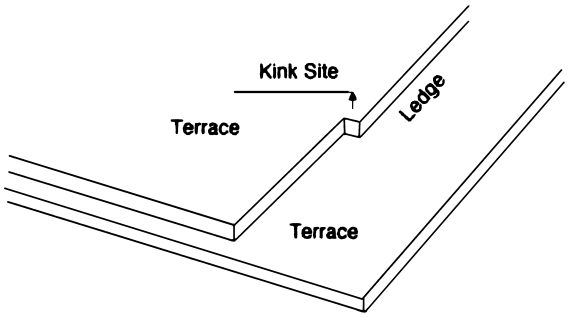


FIGURE 1 Schematic diagram of a substrate surface consisting of terraces, ledges, and kinks. Kink sites are the highest energy, followed by ledges then terraces. [Reprinted from Kuech, T. F., and Tischler, M. A. (1996). Epitaxial growth. *In* "Materials Science and Technology: A Comprehensive Treatment" (K. A. Jackson, ed.), Vol. 16, pp. 107–172, Wiley/VCH, Weinheim.]

for attachment, the highest-energy surface site is the kink, followed by the ledge and then the terrace. The amount of each of these types of structures present at the surface primarily depends on the orientation of the surface. Singular surfaces are those that are exactly oriented to a principal crystallographic plane. Surfaces that are misoriented from a principal plane toward another crystallographic direction are termed vicinal surfaces. Vicinal surfaces tend to have a larger number of ledges and kinks than that of singular surfaces. Most epitaxial growth requires a high surface concentration of such features to ensure a smooth growth surface.

Since growth occurs on a substrate surface, the substrate must be chemically clean in addition to having the appropriate structure. Semiconductor substrates need to be free of organic residues and oxides. *Ex situ* cleaning processes have been developed as a means of preparing the surface for growth. For example, H_2SO_4 and H_2O_2 solutions can remove organic compounds from Si surfaces and HF-based solutions are capable of removing SiO_2 from Si surfaces. Additional *in situ* cleaning processes can also be used to prepare the substrate for epitaxial growth. These processes can include ion bombardment or sputtering of the surface as well as thermal annealing treatments.

In addition to the development of bulk substrates, the development of substrates possessing unique structures and properties has enhanced epitaxial growth technology. In particular, the advancement of silicon-on-insulator (SOI) structures can improve many silicon-based devices. The SOI structure is shown in Fig. 2 and consists of a single-crystal Si layer supported by SiO_2 on a bulk Si substrate. This structure provides a means of electrically isolating devices and can reduce parasitic capacitances within

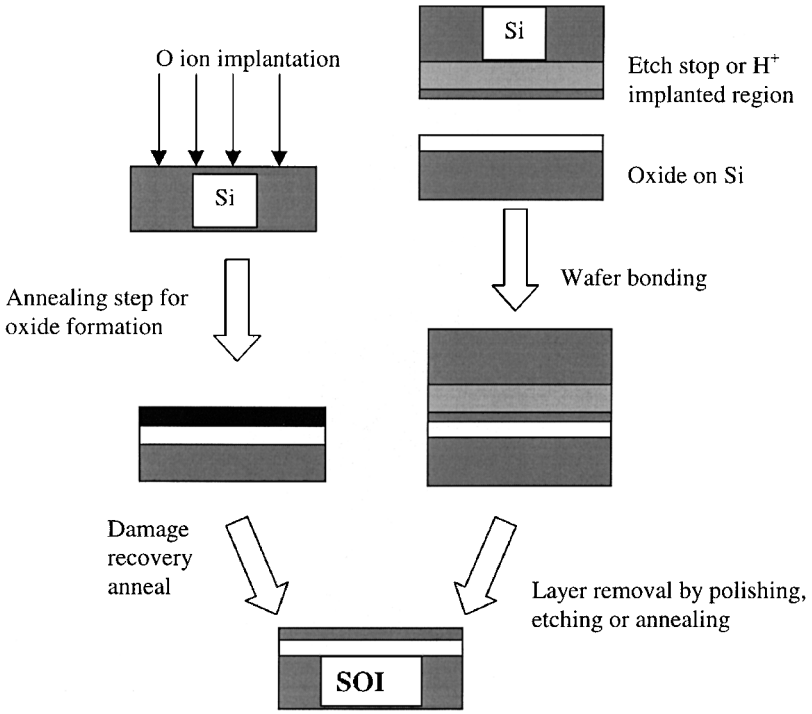


FIGURE 2 Methods for fabricating silicon-on-insulator (SOI) structures. The SIMOX process (left) involves the implantation of oxygen followed by thermal annealing. Wafer bonding (right) involves bonding Si to SiO_2/Si to form a wafer stack with subsequent removal of a majority of one Si wafer by techniques such as mechanical lapping or H^+ implantation and annealing.

integrated circuits. SOI structures can be fabricated through a technique called separation by implantation of oxygen (SIMOX), as shown in Fig. 2 (left). In this technique, O^{2-} ions are implanted into a Si substrate. Thermal annealing leads to oxygen's diffusing and reacting within localized regions of the Si lattice to form the buried SiO_2 layer. The depth and thickness of the buried oxide depends on the fluence and energy of the implanted ions and the thermal annealing procedures. An additional high-temperature annealing cycle is needed to eliminate the damage to the top Si layer caused by the O^{2-} ion implantation.

SOI structures are also fabricated by wafer bonding techniques, also as shown schematically in Fig. 2 (right). In this case, SiO_2 is deposited on Si by any of a number of established methods. The SiO_2 on Si is then bonded as a wafer to another Si wafer. This wafer stack is thinned to leave a Si layer on the SiO_2. The Si can be removed by conventional lapping and polishing methods. An alternative technique requires that the Si be implanted with H^+ ions prior to bonding of the wafer stack. After wafer bonding, thermal annealing of the structure results in splitting of the surface Si layer at the peak of the H^+ implant range, which leaves a layer of Si and forms the SOI structure. The advantage of this technique is that the removed Si wafer can be reused in another processing step.

III. VAPOR PHASE EPITAXY

Vapor phase epitaxy (VPE) is a type of chemical vapor deposition (CVD) and pertains to the deposition of epitaxial material from a gas phase environment. There are many different classifications of VPE techniques, all involving the flow of a gas phase mixture containing reactive precursors over a heated substrate. VPE is a versatile technique capable of depositing controlled thin layers of semiconductor materials with high wafer throughput. Among the disadvantages of this class of techniques are the possible impurities from the reactor gas environment and difficulties in finding suitable gas phase chemistries.

A. VPE Reactor Design

Typical reactor designs used in VPE applications are presented in Fig. 3. These reactors can generally be classified according to whether the gas flow is parallel or perpendicular to the wafer surface. At high pressures and low gas velocities, the gas flow can exhibit recirculation pathways that can bring unwanted gas by-products into the feed stream. A depletion of precursors down the length of the substrate can alter the layer thickness uniformity and is another area of concern in reactor design.

A reactor system consists of two main components: the gas panel and the reaction chamber. The gas panel consists

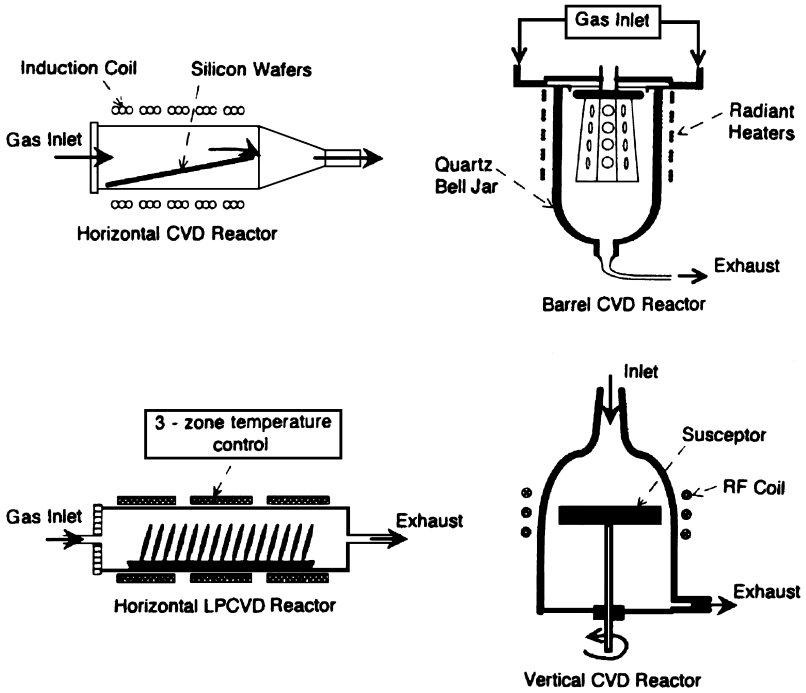


FIGURE 3 Common types of vapor phase epitaxy (VPE) reactor designs. The associated gas panels are not shown. CVD, chemical vapor desposition; LPCVD, low-pressure chemical vapor desposition; RF, radio frequency. [Reprinted from Kuech, T. F., and Tischler, M. A. (1996). Epitaxial growth. *In* "Materials Science and Technology: A Comprehensive Treatment" (K. A. Jackson, ed.), Vol. 16, pp. 107–172, Wiley/VCH, Weinheim.]

of gas lines and controllers that deliver the precursors to the reaction chamber. The gas lines are typically constructed of stainless steel with all-metal seals for line segment connections. The delivery of the precursors to the reactor depends on the phase of the precursor. Gas phase precursors are controllably delivered to the reactor by using a mass flow controller. Solid and liquid precursors require more complicated delivery systems such as the use of a bubbler. In a bubbler, a carrier gas is passed through the solid or liquid precursor and vapor of the precursor material is entrained and delivered to the reactor with the carrier gas. The delivery rate of the precursor is controlled by the temperature and pressure maintained in the bubbler as well as by the carrier gas flow rate. Standard choices of carrier gas include high-purity hydrogen and nitrogen. Computer control of the valves on the gas lines is common in order to automate the valve sequence required for growth of multiple, tightly controlled layers.

The reaction chamber, where the substrate is placed and growth occurs, is generally constructed of high-purity materials such as quartz or stainless steel. Hot-wall reactors locate the chamber inside a furnace such that all components attain the same temperature. This technique is conceptually straightforward, but the high internal surface area can act as a deposition surface, which leads to depletion of reactants from the nutrient stream before they reach the wafer. This could result in growth thickness variations and low growth efficiency. Cold-wall reactors require local heating of the substrate, through placement of the substrate onto a heated susceptor. Radio frequency induction and lamp-based radiative heating of a graphite susceptor within a nonconductive reactor are common heating techniques. A cold-wall reactor can have a lower reactant depletion rate; however, the temperature gradients often seen in these systems, due to the hot wafer surface relative to the cold reactor wall, can lead to complex, thermal-buoyancy-driven flow.

VPE systems can operate between atmospheric pressure and ultrahigh vacuum (UHV). Atmospheric pressure systems do not require additional pumping equipment, which presents some process simplification. UHV reactors, however, require complicated and specialized pumping systems to maintain the low pressure, but they typically alleviate mass transport limitations as the rate-limiting step in growth.

B. Kinetics of the VPE Growth Process

The deposition of material by VPE involves a sequence of steps: diffusion of the precursor to the surface of the growing film, adsorption of the reactant precursor onto the surface, decomposition of the precursor, surface transport, incorporation of the precursor into the growing layer, and finally the removal of reaction by-products away from the surface. The slowest of these steps determines the overall speed of the process as the rate-limiting step. Identification of the rate-limiting step to the deposition allows both insight into the process and identification of key variables that control the uniformity and properties of the deposition. Given the difference in materials and growth chemistries, both surface-reaction-limited and mass-transport-limited reaction conditions are employed in semiconductor growth and manufacturing.

The complete reaction chemistry leading to deposition involves both gas phase and surface reactions. In both cold- and hot-wall reactors the gas phase temperatures near the substrate can lead to significant gas phase reactions. For example, in Si VPE, the Si_2H_6 precursor decomposes to a mixture of SiH_2, SiH_4, and Si_2H_6 in the gas phase, all of which can adsorb onto the surface. Subsequent surface reactions decompose these species, which leads to the incorporation of Si into the epitaxial layer. The growth chemistry even for single precursor systems, therefore, can be complex and is complicated by the coadsorption and reaction of additional precursors used as dopants or other major constituents in the case of alloy growth, and the potentially complex transport phenomena within the VPE reactor.

VPE systems are classified according to whether their gas phase transport is in either a viscous or a molecular flow regime. When the gas phase mean free path, which is the distance between molecular collisions, is small compared with the overall reactor dimension, as tends to occur at higher reactor pressures and lower growth temperatures, the transport is termed viscous flow. Viscous flow is modeled by standard continuum fluid mechanics. The boundary layer model, often used to describe reactor transport, is shown schematically in Fig. 4. For simplification of the transport description, the near-surface gas phase is described as a stagnant layer, the boundary layer, through which reactant species diffuse as a result of a concentration gradient toward the substrate surface. Within this boundary layer model, decreases in the thickness of this stagnant layer can increase the delivery rate of precursors and, correspondingly, the layer growth rate. Decreases in the boundary layer thickness can be accomplished by a variety of means such as increasing the velocity within the reactor.

In the molecular transport regime, the gas phase mean free path is large compared with the dimensions of the reactor. Molecular flow is more likely when a reactor is being operated at a high temperature and a low pressure. This type of transport is modeled by the kinetic theory of gas flow. The molecular flow regime is often used when a surface reaction is the rate-limiting step for epitaxial growth since transport can be very fast.

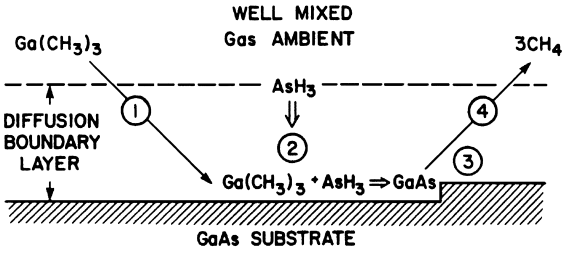


FIGURE 4 Schematic representation of the sequential steps required for deposition of material during VPE. This example is specific to deposition of III–V semiconductors but is readily extended to all chemistries. [Reprinted from Kuech, T. F. (1987). Metal-Organic Vapor Phase Epitaxy of Compound Semiconductors. *Materials Science Reports* **2,** 1–49 with permission from Elsevier Science.]

C. Dopant Incorporation in VPE

The controlled introduction of electrically active impurities, referred to as doping, is required in most semiconductor devices. Doping during VPE growth occurs when a small concentration of an appropriate impurity-containing source is added to the reactor environment. Standard n-type dopants in Si VPE are PH_3 and AsH_3. B_2H_6 is a common precursor for p-type doping of epitaxial Si layers. The dopant incorporation rate is controlled by the precursor partial pressure, reactor temperature and pressure, and rate-limiting step of the reaction chemistry.

Dopant incorporation can complicate the understanding of the overall growth chemistry. For example, on the one hand, the introduction of PH_3 or AsH_3 can reduce the growth rate of Si by VPE when SiH_4 is the Si growth precursor. Blocking of potentially reactive sites for SiH_4 adsorption and Si incorporation by strongly surface-adsorbed hydrides on the surface may explain this observation. On the other hand, the p-type dopant B_2H_6 enhances the Si growth rate at low growth temperatures. The origin of this enhanced growth rate is not clearly understood. In general, the layer growth rate and final dopant profile are affected by both the constituent and dopant chemistry. In general, operating conditions are sought such that the dopant concentration in the epitaxial layer is linear with the partial pressure of the dopant precursor in the reactor.

The ability to controllably dope epitaxial layers also requires that the unintentional incorporation of impurities be minimized. The reactor walls can be a source of unintentional impurities. Adsorbed, impurity-containing molecules can desorb from the walls and lead to subsequent incorporation. Cold-wall reactors can minimize desorption of species. In addition, unintentional impurity incorporation can occur within or between layers of deposited semiconductor material through solid-state diffusion of impurities from neighboring layers. The process of autodoping can also occur due to dopant evaporation from an existing layer and subsequent reincorporation into the growing epitaxial layer. These routes for impurity incorporation are mainly a function of the temperature of the substrate. Maintaining a low reaction temperature while still obtaining the desired layer properties reduces the undesirable incorporation of impurities from the substrate. Finally, the gas phase by-products from the growth reactor can lead to incorporation of unintentional materials into the epitaxial layer. For example, carbon is a common impurity in material deposited by metal-organic vapor phase epitaxy (MOVPE) since the metal-organic precursors, such as trimethylgallium, or $Ga(CH_3)_3$, contain carbon constituents which are removed from the growth front through the formation of CH_4. The CH_3 groups can follow other reaction pathways that do not form CH_4 and lead to carbon incorporation into the epitaxial layer. Adjusting the reactor operating conditions, on the basis of knowledge of the deposition chemistry, prevents these undesired routes of impurity incorporation.

D. Applications of VPE

VPE is widely used for epitaxial growth of layers for integrated circuit applications because it combines the abilities of controlled layer growth with high wafer throughput. Silicon growth on Si substrates by VPE is the most utilized application of epitaxial growth for integrated circuits. The Si growth precursors are commonly silanes and chlorosilanes. These reactors typically operate near ambient pressure at high temperatures (700–800°C), which makes redistribution of existing dopants in the substrate a concern. Silicon VPE under these conditions is used to deposit thick layers of high-quality material for use in complementary metal-oxide semiconductor (CMOS) technologies. Another less common Si growth technique is the deposition of epitaxial Si layers and Si_xGe_{1-x} by UHV-CVD. Silicon UHV-CVD operates at lower temperatures (<700°C) and is used to make abrupt doping and compositional junctions in the epitaxial films. UHV-CVD has been employed to grow Si_xGe_{1-x} alloys on Si substrates that can be used in high-speed electronic devices. The thickness of Si_xGe_{1-x} layers that may be deposited on Si substrates without the formation of defects is limited by the lattice mismatch between the layers.

A dominant form of III–V semiconductor VPE is metal-organic vapor phase epitaxy, or MOVPE. The name is derived from the use of metal-organic compounds such as trimethylgallium (TMGa) or trimethylindium (TMIn). The group V precursors are typically hydrides of the material of interest, such as AsH_3 for the As component and PH_3 for the P component. The growth of III–V semiconductors is complicated by the difference in elemental vapor pressure among the material constituents. For example, Ga and As in GaAs have different vapor pressures in equilibrium with the solid. In general, the group V component, As in this example, has a higher vapor pressure than the group III component, which could lead to its rapid desorption. An excess of the group V component is used within the reactor to compensate for this difference in vapor pressure. MOVPE-based materials are now used to produce optical devices such as lasers, light-emitting diodes, and solar cells.

IV. MOLECULAR BEAM EPITAXY

Molecular beam epitaxy (MBE) is the process of epitaxial growth by physical deposition of elemental sources onto a heated substrate. MBE is performed in a UHV environment at generally lower temperatures than those used in VPE. MBE growth is capable of producing junctions between heteroepitaxial layers and dopant profiles that are near atomically abrupt and hence can be used to make a wide range of optical and electronic devices. Among the disadvantages of the technique is the high capital cost associated with obtaining the UHV environment.

A. MBE Reactor Design

Since MBE is a UHV process, the design of the reactor chamber (Fig. 5) is a key aspect to successful growth. MBE reactors are constructed of stainless steel with all-metal seals. View ports may be added for optical access to the chamber. All the materials used in the construction of the MBE must withstand heating and cooling to 200°C. Heating of the chamber, or "bake out," is used to de-gas the reactor walls and reach UHV base pressures of less than $\sim 10^{-9}$ Pa. As shown in Fig. 5, the reaction chamber is isolated from the environment by one or more intermediary chambers in order to maintain UHV. The load lock is a chamber directly exposed to the air and is used for sample introduction. This chamber is typically evacuated to less than 1×10^{-5} Pa before the substrate is transferred to the deposition chamber or some other analysis or preparation chamber.

The pumping system is an integral part of the UHV environment. Oil-free pumps must be used to prevent carbon contamination. Typical chamber pumping systems include adsorption pumps followed by cryogenic or ion pumps. The pumping systems can be augmented by the placement of liquid-nitrogen-cooled cryo-panels within the growth chamber. These panels condense volatile gas impurities and lead to an improvement in the material purity.

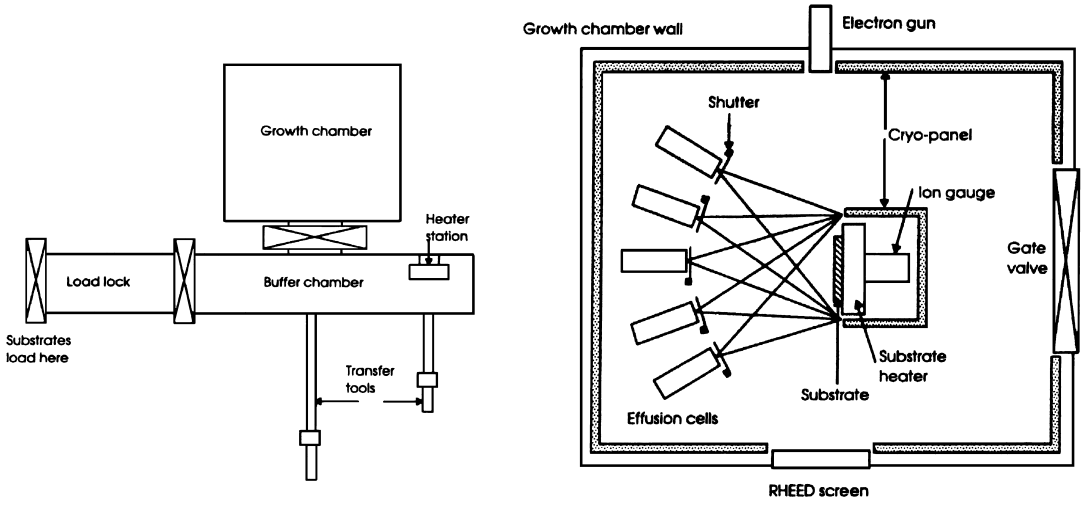


FIGURE 5 Schematic of a molecular beam epitaxy (MBE) system. The figure on the left demonstrates the chambers that may compose an MBE. The figure on the right is a schematic of the reactor chamber, indicating sources in the form of effusion cells and reflection high-energy electron diffraction (RHEED) *in situ* diagnostics. [Reprinted from Kuech, T. F., and Tischler, M. A. (1996). Epitaxial growth. *In* "Materials Science and Technology: A Comprehensive Treatment" (K. A. Jackson, ed.), Vol. 16, pp. 107–172, Wiley/VCH, Weinheim.]

MBE is essentially a cold-wall reactor with the cryopanels used to minimize contamination from the reactor walls. The substrate is heated locally by resting on a susceptor that is either resistively or radiatively heated. The temperature of the substrate is measured and controlled by a thermocouple attached to the susceptor or by optical pyrometry of the substrate surface. Exact temperature control and uniform heating of the substrate, such that the temperature does not vary by more than 10°C, is critical to the uniformity of the deposited layers.

The elemental sources used in MBE are delivered by a beam or a flux directed at the heated substrate. Thermal evaporation of a high-purity elemental source is the principal method of creating the flux. Thermal evaporation occurs by placing the source in an effusion cell, sometimes called a Knudsen cell, or k-cell, designed to maintain accurate and reproducible temperatures. Placement of the effusion cells, shown schematically in Fig. 5, is such that each cell has line-of-sight access to the substrate. Within the cell, the material evaporating or subliming from the source escapes through a small aperture at the cell end which creates a flux or a beam of material directed at the substrate. The flux of the material is largely dependent on the cell temperature. The cell is constructed of a refractory, high-purity material such as pyrolytic BN to maintain the purity of the source and growing layer. In the case of very low vapor pressure sources, electron-beam-based heating and evaporation is used, as in the case of Si MBE growth.

A wide range of diagnostic tools is available to monitor the MBE growth processes. Flux monitoring is used to determine the effusion rate from the sources in order to control the arrival rate of species and accurately control the thickness of the growing layer. These flux monitors are used to calibrate and control the cell temperature, which in turn determines the rate of source evaporation. The UHV environment allows the use of several conventional surface science tools including reflection high-energy electron diffraction (RHEED) and auger electron spectroscopy (AES). These tools are used to monitor the surface crystallographic structure and chemical composition and have aided in the development of a detailed mechanistic description of the MBE process.

B. Kinetics of the MBE Growth Process

In the UHV environment used in MBE growth, the very low pressures allow the source material to be delivered to the substrate as a direct, unperturbed beam. Since there are no gas phase interactions, the dominant kinetic processes occur on the surface. Certain growth processes can be studied during MBE growth by monitoring the intensity of the specularly reflected RHEED diffraction spot. This diffraction spot can undergo intensity oscillations whose period is correlated to the deposition of a single monolayer of material by a two-dimensional growth mode, as shown schematically in Fig. 6. Initially, the surface is smooth

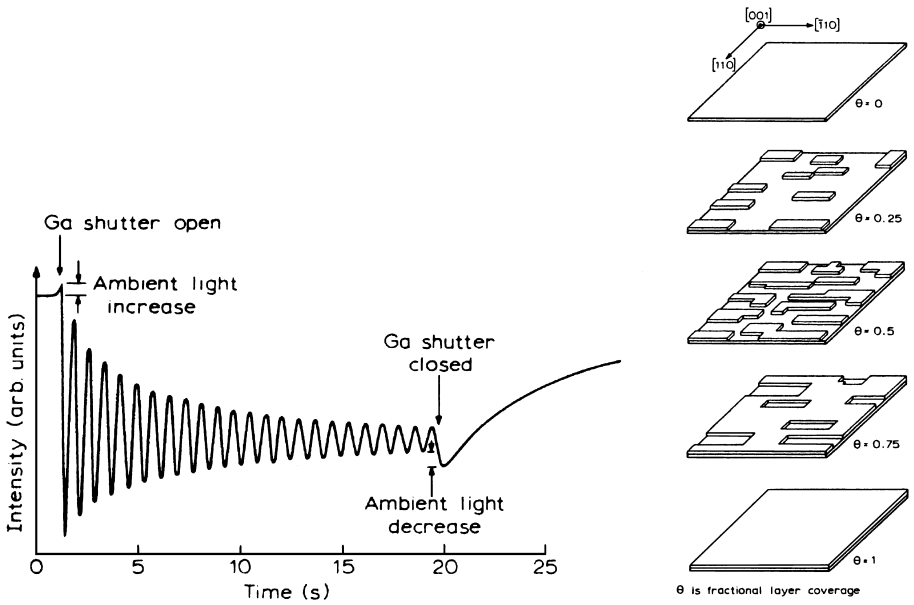


FIGURE 6 The left figure shows intensity oscillations in the specularly reflected RHEED diffraction spot during MBE growth of GaAs. These oscillations are correlated to the deposition of a single atomic layer of material as schematically shown on the right. [Reprinted from Neave, J. H., Joyce, B. A., Dobson, P. J., and Norton, N. (1983). Dynamics of Film Growth of GaAs by MBE from RHEED Observations. *Applied Physics A: Solids and Surfaces* **31,** 1–8 with permission from Springer-Verlag.]

and the specularly reflected RHEED intensity is at a maximum. Two-dimensional nucleation events occur on the surface and these begin to grow as material diffuses on the surface and coalesces. This corresponds to a roughening of the surface and a drop in the intensity of the RHEED spot. When the surface is covered with one-half of a monolayer of material the intensity is at its minimum due to this surface roughness. As the deposition continues to complete the single layer of atoms, the RHEED intensity again increases to its maximum value as the surface becomes smooth. Such oscillations confirm the presence of the two-dimensional growth mode during MBE and this is the typical growth mode used in MBE to produce device-quality material. The RHEED oscillations are eventually damped with time as a result of initiation of the growth of additional layers prior to the completion of the previous layer, which results in an overall surface roughening with time.

The MBE growth rate and materials properties are dependent on the arrival rate of material to the surface. The surface species require sufficient time to diffuse on the surface and incorporate into the epitaxial layer in order to obtain smooth and uniform layer growth. Surface diffusion exponentially depends on temperature and since MBE is a low-temperature process, the overall diffusion rate of surface species can be slow. The arrival rate of material is therefore controlled so that the surface species have time to diffuse and incorporate before being trapped in place by subsequently deposited material. If the flux of material arriving at the surface is too high, the material will accumulate and island growth will result, which will destroy the abrupt interfaces that can be achieved by MBE.

C. Dopant Incorporation in MBE

Dopant introduction into the growing layer during MBE is achieved by codeposition of dopant atoms, again through evaporation from a source cell. In Si MBE the common dopants are Sb for *n*-type doping and either Ga or B for *p*-type doping. During GaAs growth by MBE, the most common *n*-type dopant is Si while *p*-type doping is typically achieved with Be.

The steps to dopant incorporation are similar to those for incorporation of the constituent material. Dopants adsorb onto the surface and diffuse until they incorporate into the growing film. The presence of the dopant, however, can alter or even hinder the growth of the epitaxial layer. Dopant atoms on the surface can impede the lateral diffusion of atoms, which can alter the layer-by-layer growth mode of the epitaxial film. At very high dopant levels, breakdown of the two-dimensional growth behavior sometimes occurs, which degrades the film's surface morphology and utility in subsequent device formation.

Unintentional impurity incorporation in MBE primarily results from the common components making up the residual or background gases in the MBE chamber, such as H_2O, CO_2, CO, and O_2. The level of these species can be reduced with proper reactor design and sound operating procedures that include system bake outs. Other sources of impurities in the epitaxial layer arise from the materials of construction used in the effusion cells. As the effusion cells are heated, these materials can also evaporate at a low rate and deposit onto the growing surface. Refractory, high-purity materials, such as Ta or BN, are used to reduce this source of unintentional impurities. Dopant redistribution from existing layers can be minimal during MBE due to the low growth temperatures.

D. Applications of MBE

MBE has been used to deposit a wide range of materials, including Si, Ge, III–V semiconductors, II–VI semiconductors, and metals and oxides. In general, MBE is an excellent technique for research applications as well as for specialized devices for which abrupt junctions and thin layers are required but a high wafer volume is not necessarily needed.

GaAs growth is the most widely studied MBE growth system. GaAs MBE uses Ga and As elemental sources that are placed in effusion cells and form molecular beams of Ga, As_2, and As_4 that impinge on the surface. Growth temperatures are typically between 600 and 700°C, with a corresponding growth rate of $\sim$1 μm/h. The specific growth temperature and rates used are dependent on the device to be made from the epitaxial layer. For example, growth at higher temperatures reduces oxygen incorporation into the layer, which results in greater efficiency in optical devices. MBE-grown electronic devices can use high-mobility two-dimensional electron gases (2DEGs) formed at the junction of a suitable combination of a high and low band-gap material (e.g., GaAs/$Al_xGa_{1-x}As$). Field effect transistors based on these 2DEGs, as well as heterojunction bipolar transistors, are finding application in communication and microwave applications.

V. LIQUID PHASE EPITAXY

Liquid phase epitaxy (LPE) is based on the deposition of material onto a substrate from a supersaturated solution. LPE is a near-equilibrium growth process that is most commonly used for III–V semiconductors. Some of the advantages of this technique are the simple and low-cost equipment, the high utilization efficiency of precursors, and the high layer purity. Among the principal disadvantages is the inability to easily deposit thin layers or to obtain sharp junctions.

A. LPE Reactor Design

LPE growth occurs by placing a supersaturated liquid solution containing the precursors near a suitable substrate. For example, GaAs growth solutions consist of liquid Ga melt saturated with dissolved As. Therefore, unlike MBE and VPE, LPE is performed in a group III rich environment. The method by which the supersaturated solution makes contact with the substrate distinguishes the various forms of LPE growth. The three most common techniques used to grow layers by LPE are tipping, dipping, and sliding. In tipping, growth begins when a crucible containing the growth solution is tipped until the solution wets or covers the substrate. Returning the crucible to the original position so that the solution no longer wets the substrate stops the growth. Tipping is easily performed but is not readily extended to multilayer depositions. In a related fashion, the substrate can be dipped into the growth solution to initiate growth and removed from the solution to terminate growth. Dipping has been used to grow multilayer structures, but layer thickness can often be difficult to control.

The slider technique is the most successful type of LPE and can be used for growing multilayer structures. A typical reactor used in the slider technique is shown in Fig. 7. Layer growth occurs by positioning the substrate held in the base under one of the growth solution bins. Each bin can contain a different growth solution and multiple-layer growth is performed by sliding the substrate under the different bins. In addition to the slider and the base, caps can be placed on the bins to reduce evaporation of volatile components in the solution. Graphite is used to construct the apparatus because it can be high purity, it can be machined relatively easily, and carbon has a low solubility in the growth solutions. Also, the growth solutions do not wet the graphite surface and growth is arrested simply by sliding the substrate from under the bin of solution, which allows the solution to be wiped from the growth surface. Contamination of neighboring solutions due to the small amount of material transfer is therefore reduced due to this nonwetting behavior.

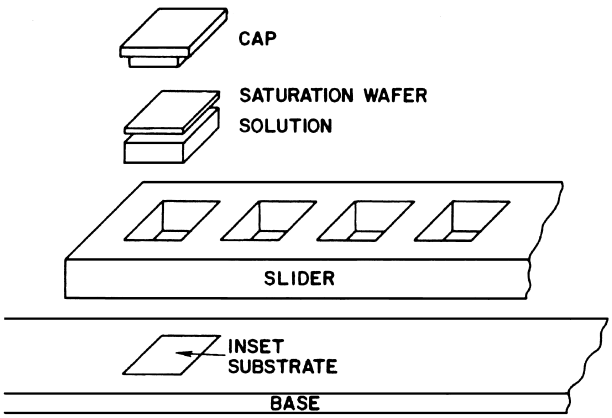


FIGURE 7 Schematic diagram of the slider-type reactor used to deposit semiconductor layers by liquid phase epitaxy (LPE). [Reprinted from Small, M. B., Giess E. A., and Ghez, R. (1994). Liquid phase epitaxy. *In* "Thin Films Epitaxy 3A: Basic Techniques" (D. T. J. Hurle, ed.), Vol. 3, pp. 107–172, Elsevier Sci., Amsterdam.]

The LPE slider apparatus is placed in a high-purity quartz tube that is purged with high-purity hydrogen in order to provide a clean environment. The tube is placed inside a multiple-zone furnace used to control the temperature of the process. The stability and uniformity of the temperature profile inside the furnace is critical to successful LPE growth. Small changes in temperature can lead to dissolution of the substrate due to the near-equilibrium nature of LPE. Heat pipes and other means of enhancing the temperature uniformity are often used.

The LPE growth solutions can be prepared by two methods. In one method the solution is created by weighing out high-purity components precisely and heating the solution to form a single liquid phase. The solution is supercooled and the resulting supersaturation provides the driving force for deposition. As the growth proceeds, this supersaturation continuously decreases and eventually limits the amount of deposition. Another method to prepare a solution is to deliberately and continuously supply an excess of the minor component. This can be done in the form of an additional wafer that floats on top of the solution during epitaxial growth. This two-phase solution will ensure the continued supersaturation of the melt during growth. The two-phase melt can be employed not only for single and binary semiconductors, but also for ternary and quaternary alloys; however, for these more complicated alloy materials the choice of the additional wafer requires careful consideration.

The temperature sequence of the LPE reactor is a key aspect to the growth behavior. The first process step is to saturate the melt solution by heating it to a temperature dictated by examination of the phase diagram of the system. Further reduction in temperature creates a supersaturation within the solution that provides the driving force for deposition. The continual reduction in temperature controls the rate and amount of deposition. Both step changes and continuous cooling are employed to yield materials of controlled thickness and composition. Continuous cooling maintains the supersaturation of the melt and can be used to grow thick layers (e.g., $\geq 100\ \mu$m).

B. Kinetics of the LPE Growth Process

The basic steps for LPE film growth are similar to those discussed with regard to VPE. LPE is a near-equilibrium deposition process and tends to rely less on the surface

reaction and surface transport than VPE and MBE do. As a result, the transport of material through the solution to the substrate is the rate-limiting step. LPE layer growth can be understood by models of material transport through the solution to the growth surface where it is consumed. One-dimensional transport-limited growth models have been developed and have been continually investigated and improved, which has furthered the understanding of the growth process.

The growth rate for LPE can be large, on the order of 0.1–1.0 μm/min. Although LPE is capable of growing high-purity material with low capital cost, VPE and MBE have dominated the growth of thin layers. A planar morphology over a large-area substrate can be difficult with LPE. A high surface mobility during growth results in large lateral growth rates. This, combined with difficulty in maintaining a uniform temperature distribution over large areas, can result in local changes in the growth rate and a rough morphology. Also, the growth rate and surface morphology are strong functions of the substrate orientation in LPE that can lead to surface faceting and again a rough surface morphology.

C. Dopant Incorporation in LPE

Intentional dopants are incorporated into the growing layer by adding the desired material to the growth solution. For the $Al_xGa_{1-x}As$ material system, several elements can be used, with the typical *n*-type dopants being S, Se, and Te and the *p*-type being Zn and Cd. The group IV elements Si, Ge, and Sn can be used as either donors or acceptors in this material system, with the behavior determined by the growth temperature and the composition of the melt and epitaxial layer.

Unintentional dopants in LPE can also arise from the reactor, as in the case of VPE and MBE. High-purity graphite is the material of choice for slider construction because it may help to minimize such impurity incorporation and because of the nonwetting of the graphite by the melt. Unintentional impurities in the melt are also a concern for LPE material since they can be incorporated into the layer. A common impurity in LPE of III–V semiconductors is O_2. This impurity is generally a concern in growth of Al containing alloys by any technique since the strong Al–O bond leads to the formation of Al_2O_3, a stable material. These types of impurities can be removed by careful preparation and purification by baking the solution in a high-purity gas environment.

Etch back of the substrate into the solution prior to layer deposition is another source of unintentional impurities. Etch back is a process by which the substrate initially dissolves into the melt, which causes unintentional materials to be added. This process can be intentionally used to dissolve the top region of the substrate to remove surface impurities and perhaps handling damage. When the epitaxial growth begins, the dopants and constituents dissolved from the substrate surface may be reincorporated as unintentional impurities. Carefully designed growth procedures such that the supersaturation in the melt strongly favors deposition over dissolution of the substrate can minimize unintentional etch back.

D. Applications of LPE

LPE can be an inexpensive technique with large deposition rates of high-purity layers that is, in principle, not very complex in its design. LPE has had success in depositing III–V semiconductors for applications in optical devices such as lasers, light-emitting diodes, infrared detectors, and solar cells, particularly in the $Al_xGa_{1-x}As$ material system. LPE materials have excellent optical properties in this system since Al in the melt solution reacts with residual oxygen and removes it from the growing material. Also, the growth in a group III rich environment reduces group III crystal defects that are associated with reduced optical efficiency.

Some drawbacks of LPE have been touched upon in the previous section, such as difficulties in producing very thin epitaxial layers with planar morphology. In addition, not all compositions of ternary and quaternary alloys can be deposited by LPE due to immiscibility regions in the composition range of some melts. Since LPE is near equilibrium, the miscibility gap of some alloys leads to phase separation, which precludes this technique as a viable means of depositing a homogeneous alloy material. The Ga-In-As-P quaternary alloy, material useful in visible optical devices, is one system that demonstrates regions of immiscibility and is difficult or impossible to produce by LPE.

VI. CONCLUSION

The goal of epitaxial growth is to deposit thin layers of material with well-defined chemical, physical, and optical properties. The distinguishing aspects of the growth techniques discussed in this article are principally the phase of the growth precursors and the degree of supersaturation driving the growth. These differences between the epitaxial growth techniques lead to different growth behaviors. Consequently, each technique has found application for specific devices and integrated circuits. VPE is the most

widely used epitaxial growth technique; however, VPE, MBE, and LPE have all been used to form device materials and to improve integrated circuit and device performance in the semiconductor industry.

SEE ALSO THE FOLLOWING ARTICLES

CRYSTAL GROWTH • INTEGRATED CIRCUIT MANUFACTURE • MOLECULAR BEAM EPITAXY, SEMICONDUCTORS • SUPERLATTICES • THIN FILM TRANSISTORS • ULTRA LARGE SCALE INTEGRATION (ULSI) DESIGN

BIBLIOGRAPHY

Grovenor, C. R. (1989). "Microelectronic Materials," Hilger, Bristol.
Hess, D., and Jensen, K. F. (1989). "Microelectronics Processing," Am. Chem. Soc., Washington, DC.
Hurle, D. T. J., ed. (1994). "Handbook of Crystal Growth 3: Thin Films and Epitaxy," Elsevier Sci., Amsterdam.
Kasper, E., and Bean, J. C., eds. (1988). "Silicon–Molecular Beam Epitaxy," CRC Press, Boca Raton, FL.
Lee, H. (1990). "Fundamentals of Microelectronics Processing," McGraw-Hill, New York.
Vossen, J. L., and Kern, W. "Thin Film Processes II," Academic Press, San Diego.



Field-Effect Transistors

A. Blicher*
I. H. Kalish
David Sarnoff Research Center

K. F. Brennan
Georgia Institute of Technology

I. Fabrication Techniques
II. The Field-Effect
III. Junction Field-Effect Transistor
IV. Metal-Insulator-Semiconductor Field-Effect Transistor

GLOSSARY

Abrupt semiconductor junction Junction with a step-like doping density transition from the p- to the n-type semiconductor.

Built-in potential Potential difference existing in a pn junction in equilibrium.

Channel Path for the majority carrier current flow.

Drain Sink for the majority charge carriers.

Flat-band voltage Metal oxide semiconductor transistor gate voltage required to counterbalance the work-function difference between the gate metal and the semiconductor and the effect of all electric charges in the gate dielectric. The higher the flat-band voltage, the higher is the device gate turn-on (threshold) voltage.

Gate Control electrode whose field modulates the magnitude of the current of the field-effect transistor.

Insulated-gate field-effect transistor (IGFET) Field-effect transistor with the conductive gate insulated from the semiconductor.

Junction field-effect transistor (JFET) Field-effect transistor with a junction-type gate.

Junction depletion region The $p - n$ junction space charge formed at the interface of the p- and n-type semiconductors.

Metal oxide semiconductor field-effect transistor (MOSFET) An IGFET with the metal gate insulated with an oxide from the semiconductor surface; a four-terminal device consisting of the source, drain, gate, and substrate (body).

Pentode Electron tube that consists of five electrodes and whose anode current stops increasing after the anode voltage reaches some critical (saturation) value.

Source Electrode that supplies the majority carriers.

Threshold voltage Turn-on voltage that must be applied to the MOSFET gate in order to induce a strongly conducting channel.

Triode Electron tube consisting of three electrodes. Its

*Formerly with RCA Laboratories

anode current increases with the anode voltage and stops growing only if there is not enough electrons emitted by the cathode.

A FIELD-EFFECT TRANSISTOR (FET) is a semiconductor device in which the magnitude of the majority carrier current depends on the modulation of the conductance by an electric field. A potential applied to a control electrode (gate) produces the field that controls the current flow from the source electrode to the drain electrode. Because of the very high gate input impedance, the input current is very small. Two types of FETs are of significance: JFET with a junction-type gate, and IGFET with an insulated gate. FETs offer several advantages over bipolar devices in many applications due to their high input impedance, their good thermal stability, and the relative ease of integration of many devices on the same semiconductor chip. The last-named feature has been the basis for a rapid expansion in the development of VLSI (very large scale integration).

I. FABRICATION TECHNIQUES

The basic techniques for fabricating FETs include semiconductor crystal oxidation, photolithography, ion implantation, diffusion, epitaxial crystal growth, etching of patterns and surfaces, and metallization.

In many cases *pn* junctions are formed by the diffusion of *n*-type and/or *p*-type impurities. Diffusion involves two steps: gaseous or impurity coating deposition and drive-in diffusion, which moves the impurity atoms below the semiconductor surface. The process takes place in furnaces with a neutral or oxidizing atmosphere and at elevated temperatures. Diffusion usually requires temperatures above 1000°C.

In order to restrict the deposition of impurities to well-defined areas, the semiconductor surface is first oxidized to form a barrier to the impurities. Then, the crystal is covered with a photoresist film, a material sensitive to ultraviolet light. By the use of appropriate masking (photomasks) the photosensitive film is exposed, baked, and chemically etched away in those areas that undergo diffusion. In the case of very small dimensions, ultraviolet light is replaced by X-ray or electron beam exposure.

To fabricate thin, lightly, and uniformly doped layers above a highly doped layer of the same or different polarity, it is advantageous to use epitaxial growth in place of diffusion. The semiconductor substrate (e.g., a silicon wafer) is placed in a chamber on a carbon susceptor, which is raised by induction heating to a very high temperature (>1000°C). A gas, such as silane (SiH_4), silicon tetrachloride ($SiCl_4$), or trichlorosilane ($SiHCl_3$), is passed over the silicon wafer. In the case of SiH_4, at temperatures above 1000°C the silane decomposes irreversibly into hydrogen and silicon, the latter of which is deposited as a film on the silicon wafer surface.

Another widely used and more precise technique for the deposition of very thin layers is known as ion implantation. The doping impurity is first ionized in high vacuum and then accelerated by high electric field before reaching the semiconductor substrate. After the removal of the substrate from the vacuum chamber the impurity is diffused and heat-treated (annealed) in order to repair the mechanical damage to the crystal caused by the highly energetic ions. The advantage of ion implantation over thermal deposition and diffusion is due to the introduction of the dopant at low temperature, at specific places, and with a very exact impurity dosage. Masking can be accomplished by using silicon dioxide (SiO_2), silicon nitride (SiN_4), metals, or photoresists.

In molecular beam epitaxy both the semiconductor atoms and doping atoms are deposited on heated substrates in an ultrahigh vacuum in order to very slowly grow single crystals. The precision of the deposited film thickness is achieved by either blocking or opening a mechanical shutter placed in the path of the molecular beam. The shutter is controlled by a preprogrammed computer. Layers less than 1 nm thick are achieved by means of this technique. Molecular beam epitaxy is used in the fabrication of modulation-doped FETs.

During device processing, it is necessary to remove the oxide in the areas undergoing diffusion or to open some areas to make electrical contacts. Cleaning of semiconductor surfaces is of great importance, since to a large extent it determines the quality of the device. All these functions can be performed by dry chemical etching, which can be accomplished by gas plasma discharges in a closed chamber. Whereas wet etching tends to undercut photoresist or silicon dioxide films, dry etching, being anisotropic, avoids this undesirable feature.

Ohmic (i.e., very low resistance) electrical contacts to the semiconductor are made by using very carefully selected metals. The most commonly used metals for silicon devices are aluminum–silicon and gold–silicon eutectics that melt at comparatively low temperatures. These metals are first deposited and then alloyed into the semiconductor, forming a good electrical contact with either the highly dopedn n^+ or the highly doped p^+ silicon.

II. THE FIELD-EFFECT

A. Definition of the Field-Effect

Unlike a metal or an insulator, the electrical conductivity of a semiconductor can be readily engineered. The

conductivity of a semiconductor can be controlled in several different ways through the direct modification of the conductivity by the introduction of impurities called dopants into the host semiconductor lattice; through the generation of carriers by optical excitation; through electrical injection; and by the field-effect. In all of these cases, the conductivity can be altered by many orders of magnitude and many useful electrical devices can be made. The field-effect is arguably the most ubiquitous of these techniques since it underlies the workings of the transistor devices used in virtually all integrated circuits.

The electrical conductivity of a semiconductor can be expressed as the product of the charge density, n, charge, q, and carrier mobility, μ, as

$$\sigma = qn\mu \tag{1}$$

Altering the charge density, n, can control the conductivity of the semiconductor. The field-effect provides a very useful and reversible means of engineering the conductivity of a semiconductor layer. The field-effect can be defined as the modulation of the conductivity of an underlying semiconductor layer by the application of a voltage to a surface electrode, typically called the gate. The application of the gate bias alters the charge concentration within the underlying semiconductor. Depending upon the magnitude and sign of the applied gate voltage, charge can be attracted or repulsed from the gate. As a result, the conductivity of the underlying semiconductor layer can be increased, decreased, or undergo a change in carrier type, i.e., change from n-type (majority carrier electrons) to p-type (majority carrier holes) or from p- to n-type. The semiconductor layer can become highly conductive or resistive thereby altering its current-carrying properties. When the gate is biased such that the semiconductor layer is conducting, a current can flow through the layer producing a conducting channel. Likewise, when the gate is biased such that the semiconductor layer is insulating, the current can be constricted. By controlling the voltage state of the gate, the current flow in the semiconductor layer can be altered.

B. Types of Gating Action

Different types of structures exist that exploit the field-effect. There are three main types. These are the junction field-effect, metal-semiconductor field-effect, and the metal-insulator-semiconductor field-effect structures. In each structure the basic operating principles are the same, i.e., the conductivity of the underlying semiconductor is modulated by the application of a gate bias. The only significant difference is the technique by which the field-effect is produced.

The junction field-effect for an n-channel device utilizes two $p^+ - n$ junctions in opposition. In an asymmetrically doped $p^+ - n$ junction the depletion region is mainly formed within the lighter doped n-side of the junction. Application of a reverse bias increases the depletion region width. If the two junctions are placed spatially in opposition, as shown in Fig 1, the two depletion regions extend toward one another constricting the conducting channel formed between them. At sufficiently high reverse bias, the two depletion regions touch and the channel is closed. Hence, the conducting properties of the semiconductor are modulated by the reverse biased diodes that form the gates of the device. Devices that utilize $p - n$ junctions to induce the field effect are called junction field-effect transistors, JFETs.

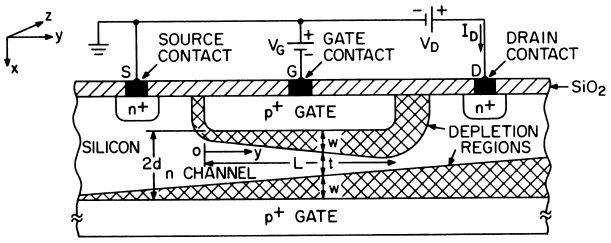


FIGURE 1 Schematic cross section of a JFET.

The field-effect can be induced using a Schottky barrier diode in place of the $p^+ - n$ diodes in JFETs. In this case, a Schottky barrier is formed on top of a thin semiconductor layer. The thin semiconductor layer is in turn grown on top of a semi-insulating material. A conducting channel is formed between the depletion region formed by the Schottky barrier and the semi-insulating material. Reverse biasing of the Schottky barrier further constricts the channel by increasing the depletion region width. In this way, the reverse biased Schottky barrier controls the conducting properties of the layer and can open and close the conducting channel. Devices that include Schottky barriers to induce the field-effect are called metal-semiconductor field-effect transistors, MESFETs.

The last type of field-effect structure discussed here relies on a metal-insulator-semiconductor, MIS, structure. In this structure, an MIS is formed on top of a semiconductor material. Depending upon the bias condition of the MIS, the underlying semiconductor layer can be accumulated, depleted, or inverted. When a voltage of opposite sign to the majority carriers is applied to the metal gate, majority carriers are attracted to the semiconductor-insulator interface. The structure is said to be in accumulation. If a voltage of the same sign as the majority carriers is applied to the metal gate, then the majority carriers are repelled from the interface forming a depletion region. Under these conditions the structure is said to be in depletion and the semiconductor becomes highly resistive. If the voltage is further increased in magnitude, minority carriers of opposite sign to the voltage are

attracted to the insulator-semiconductor interface and the structure is said to be inverted. Devices that operate with an MIS gate are called MISFETs. The most common type of MISFET is made with silicon as the semiconductor and silicon dioxide as the insulator. These structures are called MOSFETs.

III. JUNCTION FIELD-EFFECT TRANSISTOR

A. Basic Structure

For simplicity, our initial discussion of the JFET is limited to the n-channel device operating by drift of electrons. The results apply as well to the p-channel devices, depending on the drift of holes, after the appropriate change of semiconductor and bias polarities.

An n-type semiconductor layer is deposited by epitaxy on a p^+-semiconductor substrate. Subsequently another p^+ region is deposited on top of the n layer (Fig. 1). Two p^+n gate junctions are thus formed. They are electrically connected (not shown in the figure) and serve as control electrodes (gates). A reverse (negative) potential is applied between the gate and the grounded source electrode S. A positive potential V_D is applied between the drain D and the source. The reverse gate potential creates a depletion region W, thick in the lightly doped n region and very thin (not shown in the figure) in the p^+ region. The variable gate voltage, by changing the depletion region thickness, modulates the magnitude of the current in the channel. The channel thickness $t(y)$ is the difference between the undepleted channel thickness $2d$ and the total depletion region thickness $2W(y)$,

$$t(y) = 2[d - W(y)] \tag{2}$$

1. Modes of Operation and Symbols

Most JFETs operate in the depletion mode; that is, they conduct current (are on) with zero voltage applied to the gate. The normally off (enhancement-type) JFETs do not conduct current for zero gate bias and are suitable for very low power applications. The enhancement mode of operation is made possible by using a very lightly doped channel region so that the built-in junction potential is sufficient to deplete the channel region entirely. A forward voltage (positive for the p^+n junction) greater than the threshold voltage V_T reduces the depletion region thickness and initiates the current flow.

The symbols representing p- and n-channel JFETs in the depletion and enhancement modes of operation are shown in Fig. 2.

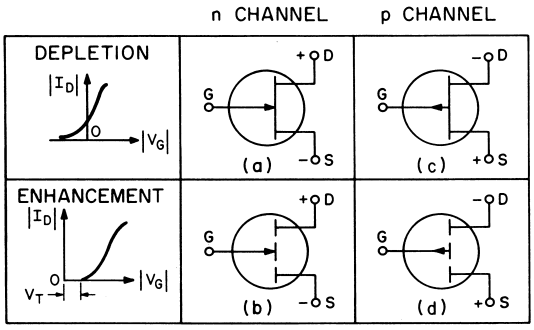


FIGURE 2 Symbols for depletion (normally on) and enhancement (normally off) n- and p-channel JFETs.

B. JFET Static Characteristics

The static (output) characteristics represent the relationship between the drain current I_D and drain-to-source voltage V_D at various values of the gate-to-source voltage V_G (Fig. 3). The study of these characteristics is simplified by considering the case of a uniformly doped channel region and abrupt (steplike) gate junctions.

If the channel length L (Fig. 1) is greater than the maximum channel thickness $2d$, it is reasonable to assume that the channel thickness $2(d - W)$ will vary slowly with distance y and that the field in the channel along y is very small compared with the transverse field in the depletion region. This model is known as the gradual channel approximation (Shockley). For nonuniform doping the result is appreciably different only for very short channels ($L < 2\,\mu$m). The first-order theory also assumes that carrier velocity is field independent.

1. Linear and Saturation Regions

For a uniformly doped channel region the magnitude of the resistive incremental voltage drop along the I_D current flow (Fig. 1) is

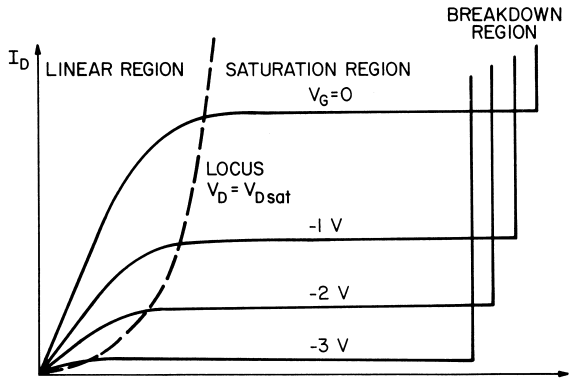


FIGURE 3 n-Channel JFET current–voltage static characteristics.

$$dV = I_D\,dR = \frac{I_D\,dy}{2[d - W(y)]Zq\mu_n N_D} \tag{3}$$

where dR is the incremental channel resistance, μ_n the electron mobility, q the magnitude of the electronic charge, N_D the ionized donor density of the n channel, and Z the channel width in the z direction. At any point y along the channel, the depletion region thickness $W(y)$ is given by

$$W(y) = \left\{\frac{2\varepsilon_s}{qN_D}[V_i + V_G + V(y)]\right\}^{1/2} \tag{4}$$

where ε_s is the permittivity of the semiconductor, V_i the built-in gate junction potential, V_G the magnitude of the gate-to-source voltage, and $V(y)$ the potential in the channel at y. Substitution of Eq. 4 in Eq. 3 and integration between $V = 0$ at the source ($y = 0$) and V_D at the drain ($y = L$) yields the drain current magnitude:

$$I_D = G_0\left\{V_D - \frac{2}{3}\frac{1}{V_p^{1/2}} \times \left[(V_D + V_i + V_G)^{3/2} - (V_i + V_G)^{3/2}\right]\right\} \tag{5}$$

with

$$G_0 = 2qdZ\mu_n N_D/L \tag{6}$$

Here G_0 is the undepleted channel conductance, and V_p the magnitude[1] of the gate-to-source pinch-off voltage at which the gate depletion regions meet for $V_D = 0$. From $p - n$ junction theory,

$$V_p = qd^2 N_D/2\varepsilon_s \tag{7}$$

For small V_D values such that $V_D \ll (V_i - V_G)$, Eq. (5) reduces to

$$I_D = G_0\left[1 - \frac{(V_i + V_G)^{1/2}}{V_p^{1/2}}\right]V_D \tag{8}$$

which shows the linear dependence of the drain current on the drain voltage. That is, the device behaves like a voltage-controlled resistor for small drain voltages. As would be expected from the assumption of an abrupt gate junction, I_D varies as a square root of the gate voltage V_G.

With the drain voltage applied, the gate depletion region thickness is the largest at the drain end of the channel and the smallest at the source end due to the voltage drop along the channel.

At sufficiently high positive drain voltage V_D the gate depletion regions become large enough to start merging at the drain side, the channel is pinched-off and the current flows in part through the depletion region. Beyond this point any further increase in the drain voltage increases the drain current only insignificantly: The drain current saturates. Equation (5) for I_D is applicable only up to this point. The drain voltage $V_D = V_{Dsat}$ at which the saturation occurs for long channel devices is given by

$$V_{Dsat} = V_p - (V_i + V_G) \tag{9}$$

the saturation voltage is smaller for larger absolute values of V_G. For $V_D > V_{Dsat}$ the excess drain voltage is dropped across the depletion region.

Substitution of Eq. (9) for V_D in Eq. (5) gives I_{Dsat}, the magnitude of the drain current in saturation as

$$I_{Dsat} = I_p\left[1 - 3\left(\frac{V_i + V_G}{V_p}\right) + 2\left(\frac{V_i + V_G}{V_p}\right)^{3/2}\right] \tag{10}$$

where I_p, the pinch-off current, is

$$I_p \equiv q^2\mu_n Z N_D^2 d^3/3\varepsilon_s L \tag{11}$$

Therefore, to the first approximation I_{Dsat} is independent of the drain voltage.

For drain voltages in excess of V_{Dsat} the depletion region moves somewhat toward the source, reducing the channel length and thus slightly increasing the drain current so that the current-voltage characteristics in actual devices show a small but finite slope in the current saturation region. The shorter the channel length, the more pronounced is the effect.

In the third region of the current–voltage characteristics the drain current increases extremely rapidly because of the avalanche breakdown of the gate junction that occurs when the drain-to-source voltage is

$$V_D = V_B - V_G \tag{12}$$

where the absolute voltage values are used; V_B is the gate junction breakdown.

For devices with short channels (small $L/2d$ ratios) the equations relating the drain current to the drain and gate voltages become rather inaccurate. One of the reasons is that the electron mobility μ_n is not constant as assumed but is a function of the electric field. When the field reaches the magnitude of 3 to 5×10^4 V cm^{-1} the electron velocity in silicon crystal reaches its maximum possible saturation value $v_s \simeq 10^7$ cm sec^{-1}. In a short channel, electric fields may exceed the critical field value. In this situation, keeping all biases unchanged, the drain current is reduced by a factor of $1 + (\mu_0 V_D/v_s L)$, with μ_0 being the low-field electron mobility.

The current–voltage characteristics of Fig. 3 are described as pentodelike. In JFETs with very short channels that are depleted even in the absence of the external gate potential, the potential barrier at the source prevents the drain current from flowing at low positive drain voltages. A high enough drain voltage lowers the barrier height

[1] V_p includes the built-in potential V_i.

and allows current to flow. In such devices and for specific geometries the drain current may not saturate and the current-voltage characteristics are triodelike.

C. Small-Signal Parameters

At the output, the dc drain current I_D is a function of both the dc source-drain voltage V_D and the dc gate voltage V_G. The total drain current, the sum of the small-signal drain current, i_D and the dc drain current I_D, is then a function of the total drain and gate voltages. The small-signal drain current can then be written as,

$$i_D = I_D(V_D + v_D, V_G + v_G) - I_D(V_D, V_G) \tag{13}$$

Expanding the total drain current in Eq. (13) to first-order obtains,

$$I_D(V_D + v_D, V_G + v_G) = I_D(V_D, V_G) + \left.\frac{\partial I_D}{\partial V_D}\right|_{V_G} v_D + \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D} v_G \tag{14}$$

Substituting Eq. (14) into Eq. (13) obtains,

$$i_D = \left.\frac{\partial I_D}{\partial V_D}\right|_{V_G} v_D + \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D} v_G \tag{15}$$

Using the following definitions,

$$g_D \equiv \left.\frac{\partial I_D}{\partial V_D}\right|_{V_G} \qquad g_m \equiv \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D} \tag{16}$$

the small-signal drain current becomes,

$$i_D = g_D v_D + g_m v_G \tag{17}$$

The quantity, g_D is called the drain or channel conductance and g_m is called the transconductance. Using these parameters, a simplifed small-signal (incremental) equivalent circuit for the common source configuration and valid only for low frequencies can be constructed. This circuit is shown in Fig. 4.

The channel conductance, g_D, in the linear region of the characteristic can be determined by differentiating Eq. 8.

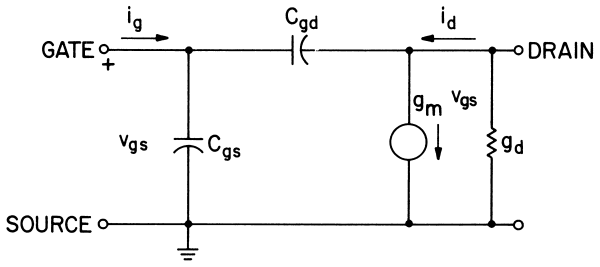


FIGURE 4 JFET incremental low-frequency equivalent circuit.

The channel conductance is given then as

$$g_D \equiv \left.\frac{\partial I_D}{\partial V_D}\right|_{V_G=\text{const}} = G_0\left[1 - \frac{(V_i + V_G)^{1/2}}{V_p^{1/2}}\right] \tag{18}$$

which is valid for drain voltages $V_D \ll (V_i - V_G)$.

In an ideal JFET the incremental conductance in the saturation region is zero, but in actual devices it has some small value for the reasons mentioned in Section II.B.1.

The effectiveness of the gate voltage control over the drain current is measured by the trans-conductance,

$$g_m \equiv \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D} \tag{19}$$

For the linear region, g_m can be obtained by differentiating Eq. (8) for I_D with respect to V_G keeping V_D constant; g_m reaches its maximum in the saturation region. Differentiating I_{Dsat} [Eq. (10)] with respect to V_G, one finds that g_{msat} is equal to g_d; that is, it is equal to the conductance in the linear region.

In a practical device the conductance g_d is smaller than that given by Eq. (18) because of the existence of small resistances in series with the source and drain. The source resistance also lowers the value of g_m, as given by Eq. (19).

The gate capacitance C_G, although of a distributed nature, usually can be accurately represented by the sum of two capacitances C_{GS} and C_{GD}. The first is the capacitance between the gate and the source, and the second is the capacitance between the gate and the drain as shown in Fig. 4:

$$C_G = C_{GS} + C_{GD} \tag{20}$$

These are intrinsic device capacitances. At high frequencies it is necessary to take extrinsic parasitic capacitances into account as well. In very short channel transistors the parasitic capacitive and inductive parameters may be most important.

The cutoff frequency f_T is the frequency at which the input current becomes equal to the output current with the output short-circuited (Fig. 4). It follows from this definition that

$$f_T = g_m/2\pi C_G \tag{21}$$

It can be shown that in saturation the maximum possible value for f_T is

$$f_{Tmax} = \frac{q\mu_n N_D d^2}{4\pi\varepsilon_s L^2} \tag{22}$$

The cutoff frequency is obviously higher for small channel lengths and for n-channel devices in silicon because μ_n, the electron mobility, is much higher than the hole mobility μ_p.

In devices with channels shorter than $2\,\mu$m and operating in the saturated region of the current–voltage characteristics, electrons mainly flow in that channel

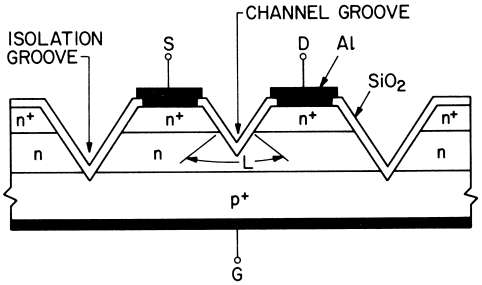


FIGURE 5 Schematic representation of the V-groove channel JFET. [After Mok, T. D., and Salama, C. A. T. (1978). *IEEE Transactions on Electron Devices*, ED-25, No. 3. Copyright 1978. IEEE.]

region, where the electric field is high and the electron drift velocity reaches its saturation value $v_s = 10^7$ cm sec^{-1}. As a result, the saturated drain current, I_{Dsat}, the transconductance, $g_{m\,\text{sat}}$, and the cutoff frequency, f_T, all become directly proportional to the saturated electron velocity v_s.

D. Silicon JFET Structures

A lateral JFET can be fabricated by double-diffusion techniques, making the process compatible with bipolar transistors integrated on the same silicon chip. The gate and channel regions are diffused or implanted from the top surface into an epitaxial silicon layer deposited on a silicon substrate.

A structure known as a V-groove channel JFET (Fig. 5) is obtained by preferential (anisotropic) etching of a groove in the (100)-oriented n-type epitaxial silicon layer deposited on the p^+ substrate. This $p^+ - n$ junction acts as a single gate. The top n^+ layers are diffused or ion-implanted to achieve good ohmic contact to subsequent metallization. In addition, etched isolation grooves are necessary to achieve electrical isolation between individual JFETs or other devices on the same silicon chip.

As the reverse bias on the p^+n junction increases, the depletion region moves up, cutting into the electron current path. Small channel lengths and high packing density on the silicon chip are achievable with this technique.

Still another JFET structure, the static induction transistor (SIT), is analogous to a vacuum tube (Fig. 6). p^+-Doped regions (gates) are embedded in the n-type silicon, forming a grid interposed between the source and drain electrodes. Depletion regions formed between the gates modulate the electron current flow from the source to drain. This multichannel device can be constructed to exhibit triodelike characteristics.

E. Metal Schottky-Barrier Field-Effect Transistor (MESFET)

The JFET $p^+ - n$ junction gate can be replaced by a Schottky barrier formed by metallizing the n-type semiconductor surface. MESFETs can be fabricated in silicon and other semiconductors. Here, we describe a GaAs device that is particularly useful in very high speed computer circuits and in microwave frequencies above 4 GHz.

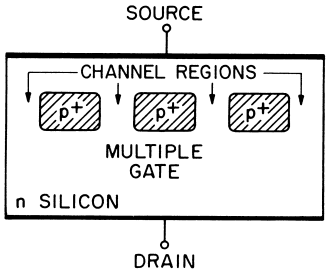


FIGURE 6 Static induction transistor. (Schematic representation.)

An n-type GaAs layer is deposited on a semi-insulating GaAs substrate (doped, e.g., with chromium) to make it electrically nonconducting (Fig. 7a). The substrate insulates the active device electrically, but not thermally, from the heat sink. GaAs is well suited for microwave operation because its low-field electron mobility is five times higher than that of silicon, and its peak electron drift velocity is twice as large as silicon. At fields lower than 3×10^3 V cm^{-1} the drift velocity reaches its peak value $v_p = 2 \times 10^7$ cm sec^{-1} and then decays to 50% of this value as the field increases to $\sim 2 \times 10^4$ V cm^{-1}. As Fig. 7b indicates the drift velocity peaks at y_1, close to the channel center, then drops to $v_s = 10^7$ cm sec^{-1} at y_2 under the drain edge of the gate, where the field is the highest. Since the electrons' velocity decreases, there is an accumulation of electrons to preserve current continuity. As the channel thickness starts increasing, the field decreases and the electron velocity increases, eventually reaching

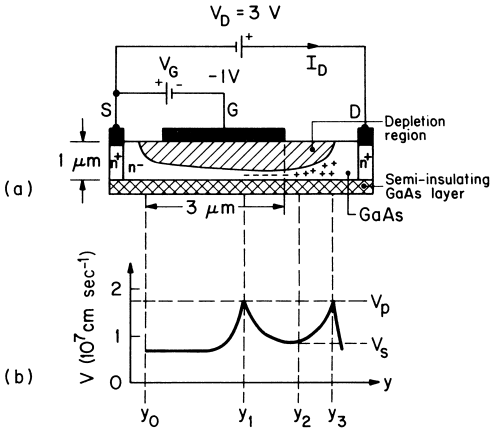


FIGURE 7 (a) Schematic cross section of a MESFET; (b) electron velocity as a function of positive y. [After Liechti, A. (1976). *IEEE Transactions on Microwave Theory and Techniques, MTT-24,* No. 6. Copyright 1976, IEEE.]

the saturation velocity, v_s. This time, in order to maintain the current continuity a depletion of electrons takes place. This amounts to the formation of positive charge. Thus a dipole layer forms underneath the gate edge. The presence of this layer affects unfavorably such device properties as the drain breakdown voltage and forward voltage drop and power dissipation.

For a gate shorter than 1 μm the electron transit time from source to drain may be shorter than the electric relaxation time of the semiconductor. This results in an electron velocity that exceeds several times the *steady-state* peak velocity v_p and in a considerably improved device frequency response. In these devices only the extraneous parasitic reactances and resistances restrict the high-frequency performance.

F. Modulation-Doped Field-Effect Transistor

The modulation-doped field-effect transistor (MODFET) is a modified MESFET device that seeks to optimize speed and current-carrying capability. In a MESFET the dopants that provide the necessary charge carriers and the source-drain current flow both reside in the top, thin semiconductor layer. The close proximity of the ionized dopant atoms and the charge carriers leads to a strong Coulomb interaction that in turn produces strong impurity scattering. As a consequence of the high impurity scattering rate, the carrier mobility is reduced resulting in relatively slow device performance. Though reducing the doping concentration would have the beneficial result of enhancing the mobility, it would also reduce the conductivity and hence the current carrying capacity of the device. Therefore, a trade-off exists in a standard MESFET structure.

MODFETs utilize modulation doping to increase the free carrier concentration without a concomitant decrease in the carrier mobility. When a highly doped wide band gap semiconductor material, such as AlGaAs, is used to form a heterostructure with an intrinsic narrower gap semiconductor, such as GaAs, carriers will move from the highly doped layer into the intrinsic layer to equilibrate the Fermi level. In this way, free charge carriers are introduced into the GaAs layer without adding dopant impurities. The spatial separation of the charge carriers in the GaAs from the ionized dopants in the AlGaAs results in a reduced Coulomb interaction and consequently a much lower ionized impurity scattering rate and higher mobility.

A typical MODFET (Fig. 8) consists of an undoped GaAs layer, 1 μm thick or less, deposited on the top of a semi-insulating GaAs layer. A very thin (a few nanometers thick) undoped, insulating AlGaAs film is grown by molecular beam epitaxy on top of the undoped GaAs layer. Then a thicker (~50-nm) n-type doped AlGaAs layer is epitaxially grown on the undoped layer. A metal gate is finally deposited on this film, forming a Schottky barrier. The source and drain are both electrically connected to the opposite sides of the n AlGaAs film and the two-dimensional gas region.

The two-dimensional electron gas concentration can be changed by varying the gate voltage; this in turn modulates the magnitude of the drain current. In the MESFET, on the other hand, the thickness of the channel is modulated and not the charge concentration.

The purpose of the undoped intrinsic AlGaAs layer is to avoid Coulomb scattering of electrons at the heterojunction interface and thus to increase the electron mobility. Depletion- or enhancement-type FETs can be

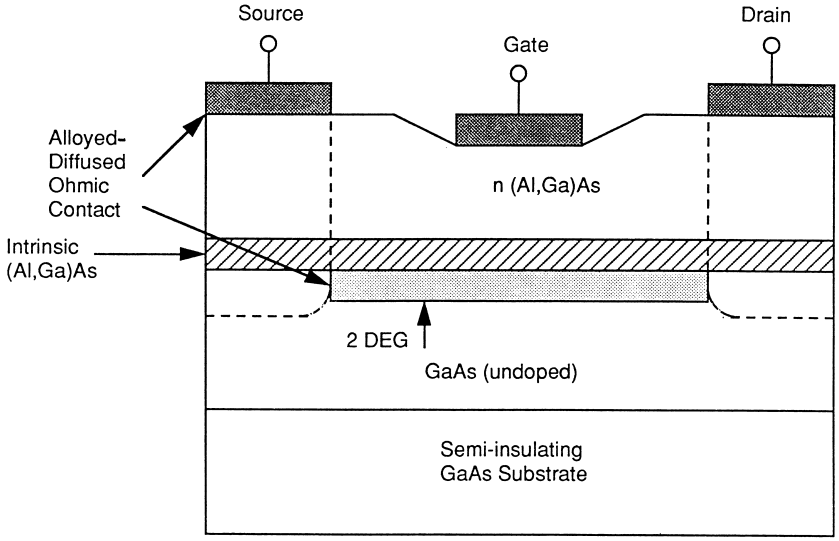


FIGURE 8 Schematic (not to scale) cross section of MODFET [After T. J. Drummond, W. T. Masselink, and H. Morkoç (1986). *Proceedings of the IEEE*, Vol. 74, No. 6, Copyright 1986, IEEE].

made depending on the appropriate choice of the undoped AlGaAs layer thickness and gate polarity. In the case of the depletion-type device the unbiased Schottky-barrier depletion region reaches only the heterojunction interface. For the enhancement-type device the depletion region also encompasses the two-dimensional electron gas.

The MODFET with some structural variations is also known as a high electron mobility transistor (HEMT), two-dimensional electron gas FET (TEGFET), or selectively doped heterojunction transistor (SDHT).

GaAs and AlGaAs have almost perfectly matched crystal lattices, so there is no strain introduced at their interface. In some other compound semiconductor structures such as InGaAs/GaAs used to experimentally fabricate MODFETs, lattices are considerably mismatched. Such MODFETs are known as pseudomorphic (pseudo-isomorphic) or strained quantum well transistors (SQWFETs). A 10–20 nm thick, *narrow* energy bandgap InGaAs layer is sandwiched between two layers of relatively thick (200–1000 nm) *wider* energy bandgap GaAs. The atomic spacing in the very thin film of InGaAs adjusts itself to the atomic spacing of GaAs; thus, the creation of crystal dislocations and deleterious interface states is avoided.

The saturated electron velocity in the InGaAs crystal, at liquid nitrogen temperature (77 K) is 20% higher than that in GaAs. This results in MODFETs with superior, high-frequency response and higher switching speeds than are possible with the AlGaAs/GaAs system. These devices are known as pseudomorphic HEMTs or PHEMTs.

Normally, electrical conduction in the valence band takes place by the transport of slow "heavy" holes. The introduction of the compressive strain into the InGaAs crystal modifies the valence band structure so that fast "light" holes with small effective masses become available for electrical conduction. This makes the fabrication of complementary *p*-type MODFETs with speeds comparable to their *n*-type counter-parts possible.

G. JFET Applications

Silicon JFETs are particularly useful in those applications that require high input impedance and low leakage currents and often are integrated with bipolar transistor circuits on the same silicon chip. Process compatibility between the two types of devices is often easier to achieve with the *p*-channel devices, Junction field-effect transistors operated in the linear region of characteristics are frequently used as voltage-controlled resistors.

By using capacitors combined with JFETs operating in the linear region of characteristics, it is simple to fabricate voltage-controlled phase shifters. Other applications include voltage-controlled attenuators, variable-gate amplifiers, voltage tuned-filter circuits, and operational amplifiers.

Junction field-effect transistors switches show excellent on-to-off current ratios and are capable of switching either very small currents, of the order of microamperes, or very large currents of the order of tens of amperes, as in field-effect power transistors.

GaAs- and InP-based MESFETs and MODFETs are primarily utilized in high-frequency device applications that arise mainly in wireless telecommunications. The key drivers in wireless RF device design are efficiency, power, and frequency performance at high linearity. Of the competing device technologies, GaAs PHEMTs offer the highest f_t at the lowest noise figure. Commercial MODFET structures used in power amplifiers have been reported with operating frequencies up to 40 GHz. Most commercial PHEMTs reported operate between 30 and 40 GHz, but some have been reported with operating frequencies up to 100 GHz.

IV. METAL-INSULATOR-SEMICONDUCTOR FIELD-EFFECT TRANSISTOR

The control electrode of a metal-insulator-semiconductor field-effect transistor MISFET is separated from the channel region by an insulator. The most important MISFET is the metal oxide semiconductor transistor (MOSFET). Its most widely used version utilizes silicon as the semiconductor and silicon dioxide (SiO_2) as the gate insulator. Virtually all semiconductor microprocessors and memory chips are fabricated using silicon-based MOSFETs. They are relatively easy to fabricate, can be readily miniaturized and integrated, and are highly reliable and reproducible. The combination of these meritorious features has conspired to make silicon MOSFETs by far the most attractive devices for computing applications. At this writing, commercially available integrated circuits employing MOSFETs with 0.13-μm gate lengths are becoming available. At these and smaller device dimensions gigabit scale integration will become commercially possible.

The *n*-channel MOSFET (Fig. 9) consists of the *p*-type bulk (substrate) and two n^+ regions: the source *S* and the drain *D*. The metal gate is insulated from the bulk by a thin SiO_2 layer. The source is kept at ground potential for the purpose of this discussion, and the drain is positively biased with respect to the source (V_D). In the absence of the positive gate potential V_G the source and drain pn^+ junctions allow only leakage current to flow. For a sufficiently high positive V_G, an *n*-type inversion layer (channel) is formed electrically, connecting the source to the drain and permitting electrons to flow. This channel current can be

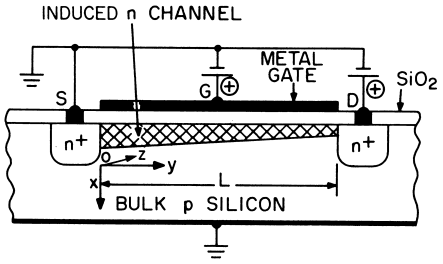


FIGURE 9 Schematic representation of an *n*-chennel MOSFET.

controlled by varying the gate voltage. The device is said to be operating in the enhancement mode.

By changing the n- and p-region doping types and making V_G and V_D negative, one obtains a p-channel MOSFET.

The depletion mode MOSFET is provided with a conducting channel during its fabrication by an ion implantation or diffusion process. Such a MOSFET is on (conducts current) in the absence of gate bias. An application of a negative gate voltage V_G will modulate the n-channel device current flow. Similarly, positive V_G is used for the p-channel depletion mode MOSFET. The symbols for the four possible MOSFET versions are shown in Fig. 10.

To understand the MOSFET operation it is essential to understand the behavior of the silicon surface. This can be described with the aid of a metal oxide semiconductor (MOS) capacitor structure.

A. MOS Capacitor

The MOS capacitor described (see also Fig. 11a) consists of the metal electrode (gate) insulated from the p-type silicon substrate by a thin SiO_2 layer. If the metal's work function $q\psi_m$ (q is the electron charge magnitude) is smaller than the work function $q\psi_{Si}$ of silicon, then as soon as the two materials are brought to the opposing surfaces of the oxide layer, some electrons are transferred from the gate

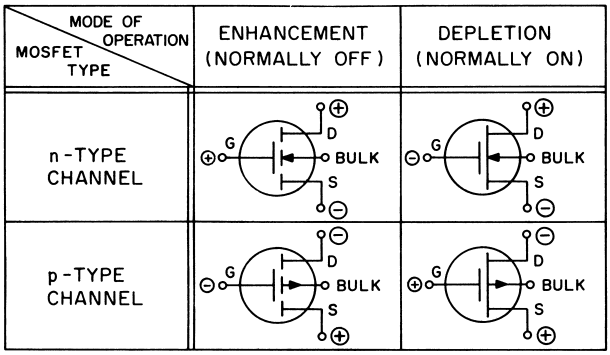


FIGURE 10 MOSFET symbols.

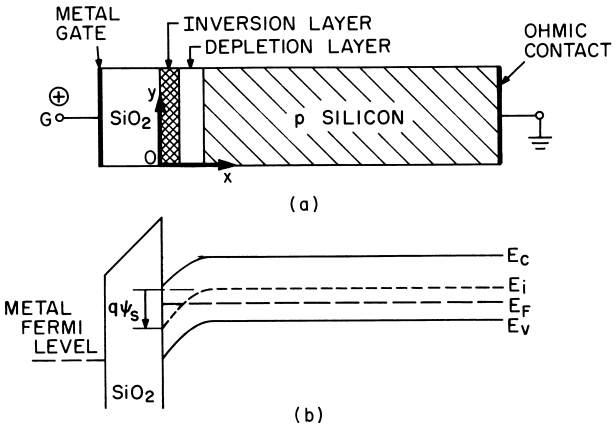


FIGURE 11 MOS capacitor. (a) Schematic cross section; (b) energy band diagram for strong inversion.

to the substrate by the imperfect insulation external to the capacitor. Thus the gate becomes positively charged and therefore repels holes in the silicon at the Si–SiO_2 interface. The p-type silicon at the surface becomes negatively charged as the number of holes is now smaller than expected from the silicon doping density. Consequently, at the surface the silicon energy bands bend so that the silicon intrinsic energy level E_i comes closer to the silicon Fermi level; the Fermi levels of the metal and silicon align, and the device comes to equilibrium (not shown in the figure). The total band bending is measured by the ψ_S (Fig. 11b), sometimes called surface potential; ψ_S is positive when pointing down. The application of an external potential to the gate equal to the potential difference $\psi_m - \psi_{Si} = \psi_{ms}$ (as modified by the presence of the SiO_2 layer) will return the bands to their initial flat-band condition. The gate voltage $V_{FB1} = \psi_{ms}$ required to do so is known as the flat-band voltage. It is dependent on the kind of metal used and the polarity and density of the semiconductor doping. For an aluminum gate and p-type silicon doped to 10^{15} acceptors cm^{-3}, $V_{FB1} \simeq -0.9$ V.

As mentioned in Section II, depending upon the external gate voltage, the MOS structure can be biased into accumulation, depletion or inversion. The metal oxide–silicon system encountered in practice typically contains several sources of impurity charge located at various sites. These charges can affect the bias condition of the semiconductor layer. For example, impurity charge in the oxide can produce inversion even in the absence of an external bias. At $\sim$2 nm from the Si–SiO_2 interface a fixed positive charge exists in the oxide. This charge, Q_{SS}, results from the imperfect oxidation of silicon ions. It is a function of crystal orientation and is smallest for the (100) crystal orientation. By the use of very careful processing and annealing, the number of charges (Q_{SS}/q) can be reduced to less than

10^{10} cm^{-2}. Other charges, such as mobile positive ions in SiO_2 due to the presence of sodium or potassium ions, and charged surface states, are kept to a minimum by various techniques and have been omitted from this discussion for the sake of simplicity.

The capacitance C_{ox} *per unit area* of the capacitor formed by the gate and the semiconductor in a highly conductive state is given by

$$C_{ox} = \varepsilon_s / d_{ox} \tag{23}$$

where ε_s is the semiconductor permittivity and d_{ox} the oxide thickness. The total flat-band voltage V_{FB} that accounts both for the work-function difference and Q_{SS}-induced potential is therefore given by

$$V_{FB} = \psi_{ms} - (Q_{SS}/C_{ox}) \tag{24}$$

For $V_G > V_{FB}$, holes from the p-type silicon surface are repelled, and a surface depletion region consisting of negative acceptors is formed. Hence, the energy bands bend *downward* (Fig. 11b). Further increase of the gate potential results in the attraction of electrons to the silicon surface, inverting it from p- to n-type. In this condition the energy bands bend enough to make the intrinsic energy level E_I cross the Fermi level (Fig. 11b). The density of electrons n induced at the surface by the gate voltage can be calculated from

$$n = \frac{n_i^2}{N_A} \exp \frac{q\psi_S}{kT} \tag{25}$$

where n_i is the silicon intrinsic concentration ($n_i^2 \simeq 2 \times 10^{20}$ cm^{-3}), and N_A—the density of the fully ionized acceptors in the substrate.

Strong inversion takes place when the electron concentration in the inversion layer of the surface is equal to the hole density in the p bulk. This takes place for band bending $\psi_S = 2\psi_F = 2E_F/q$ (E_F is Fermi energy level). The n^--type inversion layer forms an abrupt, practically one sided junction with the p-type bulk. The depletion region thickness of this junction is

$$W_{Dmax} = \left(\frac{4\varepsilon_S \psi_F}{qN_A} \right)^{1/2} \tag{26}$$

For $\psi_S > 2\psi_F$, W_{Dmax} remains almost unchanged, since the inversion layer acts as an electrostatic shield preventing the electric field from penetrating the bulk further.

A sufficiently high negative gate voltage repels electrons and attracts holes to the silicon surface, creating an accumulation layer of holes and making the silicon surface more p-type. The energy bands bend *upward* for this surface condition, and the Fermi level comes closer to the valence band: the band bending surface potential ψ_S becomes *negative*. The concentration of holes at the surface of the accumulation layer is given by

$$p = N_A \exp(-q\psi_S/\mathrm{kT}) \tag{27}$$

1. Threshold Voltage

The turn-on gate voltage for strong inversion is known as the threshold voltage V_T. It must be at least equal to the sum of the flat-band voltage V_{FB}, the band bending surface potential ψ_S, and the voltage dropped across the depletion region,

$$V_T = V_{FB} + 2\psi_F - (Q_{Dmax}/C_{ox}) \tag{28}$$

where

$$Q_{Dmax} = -qN_A W_{Dmax} \tag{29}$$

is the total charge per unit area in the depletion region of thickness W_{Dmax}. Equation (28) still applies in the case where, for ease in fabricating small structures, the metal gate is replaced by a heavily doped p^+ or n^+ polycrystalline silicon layer. Appropriate changes in the value of V_{FB} reflect the change in gate material.

B. Current–voltage Relations in the MOSFET

In the MOSFET, the current flowing in the channel under strong inversion (Fig. 9) causes a resistive voltage drop so that the band bending ψ_S varies with the distance y from the origin at the source as

$$\psi_S(y) = 2\psi_F + V(y) \tag{30}$$

Here, $V(y)$ is the reverse bias between the elemental section of the channel and bulk at y (Fig. 9). In view of Eq. (30) the depletion region thickness as a function of position y is given by

$$W'_{Dmax} = \left\{ \frac{2\varepsilon_s[2\psi_F + V(y)]}{qN_A} \right\}^{1/2} \tag{31}$$

Accordingly, the depletion region charge per unit area is given by

$$Q'_{Dmax}(y) = -\{2\varepsilon_s q N_A[2\psi_F + V(y)]\}^{1/2} \tag{32}$$

and the threshold voltage [Eq. (32)] becomes

$$V'_T = V_{FB} + 2\psi_F + V(y) - (Q'_{Dmax}/C_{ox}) \tag{33}$$

The total charge per unit area $Q_S(y)$ at the silicon surface is the sum of the channel charge per unit area $Q_{ch}(y)$ and the depletion layer charge Q'_{Dmax} per unit area. Hence,

$$Q_{ch}(y) = Q_S(y) - Q'_{Dmax}(y) \tag{34}$$

(For an n-channel MOSFET all these charges are negative.)

The total charge per unit area, Q_s, is equal and opposite in sign to the charge per unit area on the gate, Q_G, that is,

$$Q_S(y) = -Q_G(y) = -C_{ox} V_{ox} \tag{35}$$

Since,

$$V_{ox} = V_G - V_{FB} - 2\phi_F - V(y) \tag{36}$$

$$Q_{ch} = -C_{ox}[V_G - V_{FB} - 2\phi_F - V(y)] - Q'_{Dmax} \tag{37}$$

Therefore, by Ohm's law the drain current is given by

$$I_D = -Z\mu_n Q_{ch}(y)\frac{dV(y)}{dy} \tag{38}$$

where Z is the channel width in the z-direction (Fig. 9). The electron mobility μ_n is assumed here to be constant and not affected by the electric field $(-dV/dy)$ magnitude. Integration of Eq. (38) with limits 0 and V_D for $V(y)$, and 0 and L for y, respectively, gives the magnitude of the drain current as

$$I_D = \frac{Z\mu_n C_{ox}}{L}\left\{\left(V_G - V_{FB} - 2\psi_F - \frac{V_D}{2}\right)V_D - \frac{2}{3}\frac{(2\varepsilon_s q N_A)^{1/2}}{C_{ox}}\left[(V_D + 2\psi_F)^{3/2} - (2\psi_F)^{3/2}\right]\right\} \tag{39}$$

More exact theories include the effect of the channel region nonuniform doping (if any) and of current carrier diffusion in the channel resulting from the carrier concentration gradients. The concentration is the highest at the source.

1. Linear Region of Characteristics

For small drain voltages such that $V_D < 2\psi_F$ and $V_D < (V_G - V_{FB} - 2\psi_F)$, Eq. (39) reduces to

$$I_D = G_0(V_G - V_T)V_D \tag{40}$$

where $G_0 = Z\mu_n C_{ox}/L$ and V_T is the equilibrium threshold voltage [Eq. (28)]. Analogous equations can be derived for the p-channel MOSFET. Equation (40) demonstrates the linearity of the current–voltage static characteristics for low drain voltages (Fig. 12). The linear region of the current–voltage characteristics is also known as the triode region.

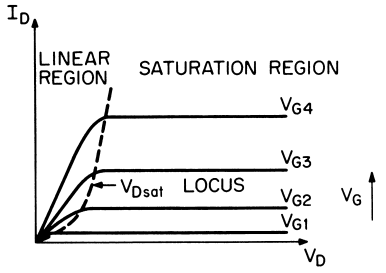


FIGURE 12 Current–voltage characteristics of an n-channel MOSFET.

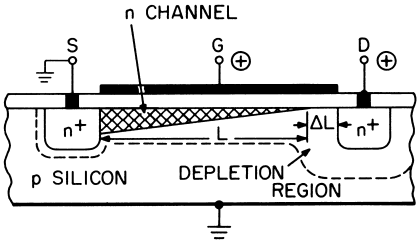


FIGURE 13 n-Channel MOSFET operating in saturation: ΔL represents the thickness of the drain depletion region.

2. Saturation Region of Characteristics

Equation (39) for the drain current shows that I_D first increases linearly with the drain voltage V_D. Then its rate of increase slows down. To a first approximation, at $V_D = V_{Dsat}$ the current reaches its saturated value I_{Dsat} (Fig. 12). Beyond this point, Eq. (39) for I_D loses its validity. At V_{Dsat} no electrons are induced at the channel drain end ($y = L$): The channel is pinched off, and there is no more gate-induced charge for current conduction. This happens because the drain voltage neutralizes the gate voltage.[2] Further electron flow occurs in the drain depletion region (Fig. 13), whose length ΔL increases as V_D increases above V_{Dsat}.

The channel charge, $Q_{ch}(y)$ vanishes near the drain, at $y = L$. Therfore, Q_{ch} is equal to zero in Eq. (37). Calling $V(y = L)$ V_{Dsat} Eq. 37 can be solved for V_{Dsat} to give,

$$V_{Dsat} = V_G - \left(V_{FB} + 2\phi_F - \frac{Q'_{Dmax}}{C_{ox}}\right) \tag{41}$$

Equation (41) can be rewritten using the expression for V'_T given by Eq. (33) recognizing that $V(y)$ is equal to zero under pinch-off conditions. Q_{Dmax} is then equal to Q'_{Dmax} and $V'_T = V_T$. Equation (41) becomes

$$V_{Dsat} = V_G - V_T \tag{42}$$

In silicon devices with gates that are shorter than 2 μm, the longitudinal electric field in the channel, in many cases, exceeds the critical field at which the electron velocity reaches its saturation value v_s and causes the drain current to saturate under the gate before the channel pinch-off takes place. In this situation Eqs. (41, 42) loose their validity.

The magnitude of the drain current in saturation I_{Dsat} can be obtained from Eq. (39) by substituting Eq. (41) for V_{Dsat}. For moderately or lightly doped substrates and for oxides that are thin compared with the surface depletion

[2]In actuality there is, of course, a small remaining charge at the edge of the depletion region in order to make the current flow possible.

region thickness, I_{Dsat} for both p- and n-channel devices can be approximated by the quadratic relationship[3]

$$I_{Dsat} = \tfrac{1}{2} G_0 (V_G - V_T)^2 \quad (43)$$

Equation (43) demonstrates the saturation drain current independence of the drain voltage for an idealized device. In actual devices the drain conductance dI_{Dsat}/dV_{Dsat} is not equal to zero. As the drain voltage is increased above V_{Dsat} the pinch-off point moves slightly toward the source because of the increased drain depletion region thickness (Fig. 13). This reduces the channel length slightly, thus increasing the current. This effect is much more pronounced in short-channel devices where the channel length is of the order of few micrometers or less.

Another complication arises for drain voltages higher than V_{Dsat} because the drain field starts dominating the gate field at the channel end, and the current spreads away from the surface so that the simple theory based on the gradual channel approximation loses some of its validity.

3. Subthreshold Behavior

When the gate voltage is below the threshold voltage in an enhancement mode device, the semiconductor layer is no longer in strong inversion. This region of operation of the device is called cutoff. In cutoff, the relationships derived above for the drain current are no longer valid. Though the drain current under these conditions is not very large, it is, nevertheless, not necessarily equal to zero either. The current that flows between the source and the drain below threshold is called the subthreshold current. The subthreshold current is extremely important in digital circuit design. In digital circuits a MOSFET is switched into cutoff in order to turn the drain current off. Ideally, no drain current should flow in cutoff. In general a small subthreshold current does flow. A nonzero subthreshold current within the device still draws power even when it is switched off. This results in unnecessary power consumption reducing battery lifetimes in portable electronic applications and heating of the chip. The subthreshold current is dominated by diffusion due to the relatively large channel resistance in cutoff.

4. Body Bias Effect

In some circuits using MOSFETs it is necessary to introduce a reverse bias V_{BS} between the body and source. This voltage appearing between the channel and body increases the channel–body depletion region thickness and therefore the total depleted charge Q_{Dmax}. As a result, the threshold voltage V_T is increased by an increment ΔV_T that is roughly proportional to $|V_{BS}|^{1/2}$ for body biases of a few volts.

[3]More accurate analysis shows that for heavily doped substrates the one-half factor is replaced by a factor smaller than one-half and dependent on the substrate concentration.

C. Short-Channel MOSFET Operation

The quantitative expressions derived above for the current-voltage characteristic for a MOSFET are based on the long-channel approximation. In this approximation, it is assumed that the longitudinal electric field is negligible with respect to the transverse field. However, as the channel length decreases, the source and drain-depletion layer thicknesses become comparable with the channel length and encroach on the area under the gate. As a result, the field becomes effectively two-dimensional. For current state-of-the-art devices, the feature lengths are sufficiently small that the long-channel approximation is no longer valid. Reduction of the gate length to ultrasubmicron dimensions influences device behavior in several important ways.

There are many different effects that appear at short-channel lengths that greatly alter device performance. These effects arise from three primary sources: (1) the electric-field profile becomes two-dimensional; (2) the electric field strength in the channel becomes very high; and (3) a decrease in the physical separation between the source and drain. The two-dimensionality of the potential profile causes a reduction in the threshold voltage of the device as well as a reduction of the mobility by gate-induced surface fields. The presence of a high electric field strength in the channel leads to velocity saturation, impact ionization near the drain, gate oxide charging, and parasitic bipolar effects. The decrease in the physical separation of the source and drain results in punchthrough and channel-length modulation. The effect of some of these mechanisms on device behavior can be summarized as follows.

1. Threshold Voltage Variation

As the channel length decreases a smaller surface depletion region charge is controlled by the gate potential (Fig. 14a). This results in a smaller value of the threshold voltage, V_T, from that of a comparable long-channel device. A channel with an extremely small width, Z, however, may show a net increase in the threshold voltage because of the lateral spreading of the depletion region in the z-direction due to the presence of fringing fields (Fig. 14b).

The threshold voltage can be intentionally altered by implanting p- or n-type dopants through the gate-oxide layer. Very precise quantities of impurities can be added in this manner and the threshold voltage can be carefully altered.

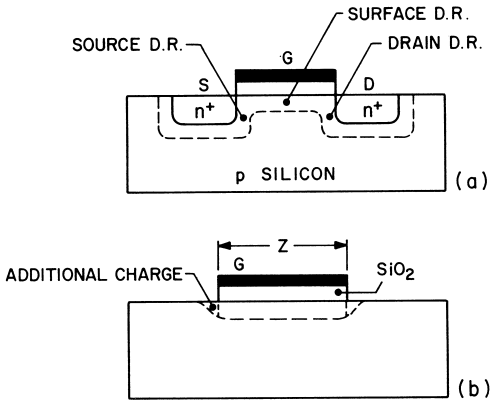


FIGURE 14 (a) Overlapping of the source and drain depletion regions (D.R.) with the surface depletion region: (b) cross section along the *z* axis showing additional charges induced by the fringe fields.

2. Surface Drift Mobility

In the expressions for the current-voltage characteristic it is assumed that the carrier mobility is constant, independent of the channel depth in the x-direction and of the electric field strength. The mobility at the Si–SiO_2 interface, often called the surface mobility, is typically smaller than that within bulk silicon, due to the presence of interface roughness scattering and surface charge scattering that reduce the mobility. However, at small gate lengths, the carrier mobility is further lowered by the action of the transverse fields. The high gate field more strongly attracts the charge carriers to the imperfect surface reducing the mobility.

3. Saturation Drift Velocity

If the voltage is not appropriately scaled as the channel length decreases, the electric field within the channel increases substantially. In silicon, the electron and hole drift velocities saturate at applied electric field strengths in excess of $\sim$100 kV/cm. In devices with very short channel lengths, the electrons move at the saturated velocity over most of their current path. Hence, the saturation of the drain current may occur at a significantly lower voltage before the expected drain pinch-off voltage is reached. The velocity saturated current in the limit as the channel length, L, goes to zero is

$$I_{\text{Dsat}} = ZC_{\text{ox}}(V_{\text{G}} - V_{\text{T}})v_{\text{s}} \tag{44}$$

4. Gate-Oxide Charging

The very high source-drain electric field heats the carriers to very high kinetic energy near the drain. In an n-channel device, the electron charge carriers can reach sufficient kinetic energy such that they can transfer from the silicon into the oxide. The actual amount of energy that the electrons need to absorb is quite high. Nevertheless, quantum-mechanical tunneling and barrier-lowering effects can reduce the potential barrier of the oxide. Carrier injection leads to gate oxide charging which in turn alters the threshold voltage of the device. The injected charge also makes V_{T} unstable, degrades the transconductance, and causes excessive gate current. The threshold voltage shift occurs over time. With continued exposure to high electric fields, the injected electron concentration within the oxide increases, gradually changing the threshold voltage. Therefore, gate-oxide charging is a long-term degradation mechanism that is particularly difficult to compensate.

5. Drain-Source Voltage Breakdown and Parasitic Bipolar Effect

The MOSFET drain-to-source breakdown voltage for short gate devices is considerably different from the avalanche breakdown voltage V_{B} of the drain–bulk junction because of the presence of an $n^+ - p - n^+$ parasitic transistor in parallel with the n-channel MOSFET. As the source current flows into the avalanching drain region, a strong carrier multiplication takes place. The generated electrons flow to the positively biased drain electrode, whereas holes go to the substrate, ground, and also directly to the source. The voltage drop caused by this hole current in the bulk forward biases the source junction, which becomes the emitter of the parasitic bipolar transistor. The drain-to-source breakdown voltage becomes that of an *npn* transistor with open base, BV_{CEO}. This voltage is dependent on the drain–bulk junction breakdown, the gate length, and the electron diffusion length, in the substrate. It is lower than V_{B} for short gate lengths and long diffusion lengths.

The existence of the parasitic $n - p - n$ transistor makes the MOSFET susceptible to the destructive thermal instability known as second breakdown. During an inductive circuit turn-off, the drain voltage may become very high while a large drain current is flowing. The high voltage causes the drain to go into avalanche breakdown. The flow of avalanche-generated holes toward the substrate and source develops a very high emitter (source) bias, establishing a positive feedback mechanism that may lead to device destruction.

6. Punchthrough

As the channel length is decreased, the physical separation of the source and drain regions necessarily decreases. Punchthrough arises when the channel length is very small. Under this condition, the source and drain-depletion regions can touch, resulting in a large increase in the drain current.

D. Incremental (Dynamic) MOSFET Parameters

As in the case for the JFET the small signal drain current can be expressed in terms of the channel conductance and the transconductance using Eq. (17).

The incremental channel conductance in the linear region, $\partial I_D/\partial V_D|_{V_G} = g_D$, is obtained by differentiating Eq. (40), which yields

$$g_D = G_0(V_G - V_T) \tag{45}$$

where g_D is linearly dependent on the gate voltage. Its reciprocal is the channel resistance in the linear portion of the current–voltage characteristics (Fig. 12).

The transconductance in the linear region is obtained by differentiating Eq. (40) with respect to V_G with $V_D = \text{const}$

$$g_m = G_0 V_D \tag{46}$$

g_m reaches its maximum value in the saturation region. The differentiation of I_{Dsat} [Eq. (43)] with respect to V_G (keeping V_D constant) results in

$$g_{msat} = G_0(V_G - V_T) = G_0 V_{Dsat} \tag{47a}$$

If the saturation current I_{Dsat} is limited by v_s (Eq. 47a) then the transconductance becomes equal to (44)

$$g_{msat} = ZC_{ox}v_s \tag{47b}$$

Equations (44)–(47) are valid for idealized devices.

1. Incremental Equivalent Circuit and Cutoff Frequency

The simplified small-signal equivalent circuit of a MOSFET is essentially the same as that of a JFET (Fig. 4). All the extraneous parasitic capacitances and resistances are omitted, and only the intrinsic device is considered. Because of these omissions the equivalent circuit is valid only at low frequencies.

As in the JFET case the cutoff frequency f_T is given by

$$f_T = g_{msat}/2\pi C_G \tag{48}$$

For the MOSFET in saturation, the drain-to-gate capacitance C_{GD} is very small because of the presence of the wide drain junction depletion region and can be neglected as compared with C_{GS}. Consequently the input capacitance C_G is expressed as

$$C_G \simeq C_{GS} \simeq LZC_{ox} \tag{49}$$

Combining Eqs. (46)–(48), we get

$$2\pi f_T = \mu_n V_{Dsat}/L^2 = 1/T \tag{50}$$

The angular cutoff frequency $\omega_T = 2\pi f_T$ is the reciprocal of the charge carrier transit time in the channel. Hence, one can see the importance of short channel lengths for high speed or high frequency of operation. In well-designed high-speed devices the parasitic elements should be reduced to a minimum since they determine the high-frequency response.

For very short channels (of the order of 1 μm) the charge transit may take place entirely in the drain region, where the electron velocity reaches its maximum saturated value v_s. In those cases the angular frequency cutoff is given by

$$\omega_T = v_s/L \tag{51}$$

E. MOSFET Applications

MOSFETs are used in numerous discrete and integrated analog circuits as small-signal and power amplifiers, operational amplifiers, phased-locked loops, and so on. Their most important application, however, is in the digital logic and memory circuits of VLSI devices, where they operate as binary elements. Power dissipation can limit the number of MOSFET devices that are practical in a given circuit. The use of complementary MOS (CMOS) devices in digital circuits represents a way of achieving very low power dissipation. The approach uses n-channel MOSFET in series with a p-channel MOSFET (Fig. 15) integrated on the same semiconductor chip. The high input voltage pulse $V_{in} = V_{DD}$ turns the n-channel transistor T_1 on and the p-channel transistor T_2 off. The low input voltage $V_{in} = 0$ turns T_1 off and T_2 on. The two transistors are never on or off simultaneously. Only a very small leakage current, the subthreshold current, may flow in steady state. Thus, in this quiescent state, no appreciable power is dissipated. The only power dissipation that occurs is that which takes place during the switching transients. This can become significant as the switching frequency increases.

1. MOSFET in Very Large Scale and Ultra-Large-Scale Integration

Continuous miniaturization of MOSFETs is necessary to maintain progress in improving integrated circuit

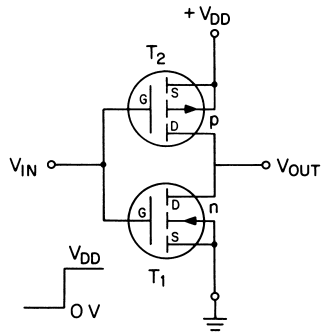


FIGURE 15 Basic CMOS circuit.

performance at lower cost. As shown in the preceeding sections simple shrinking of the channel length (below 2 μm) has a number of undesirable effects on the MOSFET performance, such as lowering of the threshold voltage; degradation of the drain-to-source breakdown voltage; the possibility of punch-through between the drain and source; and mobile carrier injection into the oxide, leading to device instability. Moreover, the subthreshold current, which adds to the device leakage, increases as the device becomes shorter. Ultra-large-scale integration (ULSI), that approaching gigabit capacity, requires devices with submicron length channels. In the design of such devices it is important to have a simple method for reducing all the device dimensions without detrimental effects.

The most successful approach to designing short-channel MOSFETs is scaling. Starting with a long-channel device that delivers satisfactory performance, a miniaturized version is made by scaling the important features such that similar operation is maintained. In this way, short-channel effects can be minimized in miniaturized devices.

There are several approaches to scaling. One method is constant-field scaling in which the device dimensions and voltages are scaled by the same factor such that the electric field remains unchanged. However, some physical quantities do not scale in the same way. Generally, the two current components, drift and diffusion, do not scale identically. Though the diffusion current is relatively unimportant at long-channel lengths, it plays a strong role in short-channel devices. Therefore, for increasingly smaller devices, constant field scaling starts to fail. Other scaling methods have replaced constant field scaling, most notably empirical techniques. From these techniques, it can be demonstrated that miniaturized devices must have thinner gate oxides and either lower applied voltages or heavier doping concentrations in order to minimize short-channel effects.

The scaling process cannot be continued indefinitely, however, since it requires ever-decreasing gate oxide thicknesses. At some point, the gate oxide can no longer be thinned since it cannot be scaled to zero. Realistically, it is unlikely that the gate oxide can be thinned to less than 1 nm, at least if SiO_2 oxides are retained. Similarly, there are constraints on the maximum doping concentrations available. In addition, at smaller dimensions new quantum mechanical effects appear. Still other limits are imposed by the circuit and system requirements.

F. MOSFET Structures

A variety of modified basic MOSFET structures have been proposed to achieve high switching speed, improved current and voltage performance, and high packing density.

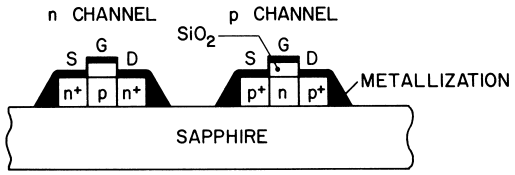


FIGURE 16 Silicon-on-sapphire with *n*- and *p*-channel MOSFETs.

1. Silicon-On-Sapphire Transistor

The silicon-on-sapphire transistor (SOS) is the most common of the various silicon on insulator (SOI) transistor structures that have been investigated. Figure 16 shows how the idea is implemented. A very thin silicon film is deposited by epitaxy on a highly polished sapphire crystal substrate. The film is etched to form *p*- and *n*-type small islands, which serve for the processing of the *n*- and *p*-channel MOSFETs, respectively. The sapphire single crystal is an excellent electric insulator and a good thermal conductor. Its primary purpose is to minimize extraneous parasitic capacitances that limit the conventional MOSFETs speed of response, and to provide nonjunction isolation between devices.

Ultra thin ($\sim$0.1 μm) silicon films result in devices with low sensitivity to transient radiation effects. Perhaps the most important advantage of SOI devices is that short channel effects can be better controlled than in conventional devices.

2. Metal Nitride–Oxide Silicon Transistor

This device belongs to the class of metal insulator–oxide silicon (MIOS) transistors and is used mainly as a nonvolatile element of read-only memories; it can store data even after power is removed.

The metal nitride–oxide silicon transistor (MNOS) Fig. 17 differs from the *n*-channel MOSFET in that the gate insulator is a composite of silicon dioxide and silicon nitride (Si_3N_4). The oxide layer is very thin (2–5 nm); the nitride layer, on the other hand, is comparatively thick (30–50 nm).

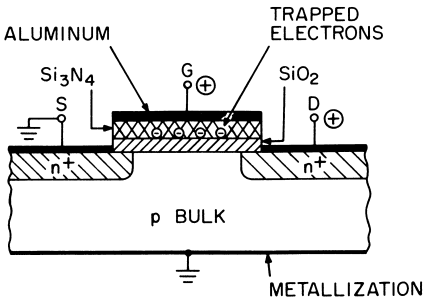


FIGURE 17 Schematic cross section of an MNOS transistor.

On the application of a positive gate voltage V_G the electrons from the bulk move by quantum mechanical tunneling through the thin oxide to the oxide–nitride interface, where they are trapped in localized states. The trapped negative charge adds up to the surface depletion region negative charge Q_{Dmax}, thereby increasing the threshold voltage magnitude. A negative gate voltage pushes the trapped electrons back into the silicon bulk thus lowering the threshold voltage. Since the MNOS transistor may be in one of these two states, it can serve as a binary element in a semiconductor memory. Switching from one state to the other may be effected millions of times before the device shows significant degradation. Metal nitride–oxide silicon transistors can be fabricated with either an n-or a p-type channel.

3. Floating-Gate Avalanche-Injection Transistor

The floating-gate avalanche-injection transistor (FAMOS) is another nonvolatalle memory element. A floating (not electrically connected) polycrystalline gate is embedded in a SiO_2 layer, which is thick enough to prevent any quantum mechanical tunneling (Fig. 18). In order to charge the floating gate the drain voltage V_D is made high enough to cause the avalanche break-down of the drain junction. Some of the electrons generated and accelerated in the drain depletion region become energetic enough to reach the oxide, to penetrate it, and to drift in the field existing there to the floating gate, where they become trapped. The gate becomes negatively charged and causes the surface inversion of the n-type substrate, turning the device (a p-channel transistor is considered) on. The charge can be maintained for an indefinitely long time. Its erasure, which causes the device to revert to the off condition, is accomplished by irradiation by ultraviolet light or X-rays. Electrical erasure is made possible by the addition of another polysilicon gate deposited on the SiO_2 insulator above the floating gate. Such a structure is known as a stacked-gate avalanche-injection MOS (SAMOS).

4. Double-Diffused MOSFET

The double-diffused MOSFET (DMOS) is schematically represented in Fig. 19. A short and well-controlled channel

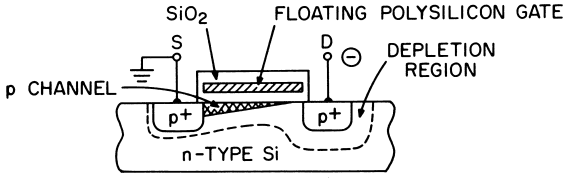


FIGURE 18 FAMOS transistor (schematic Representation).

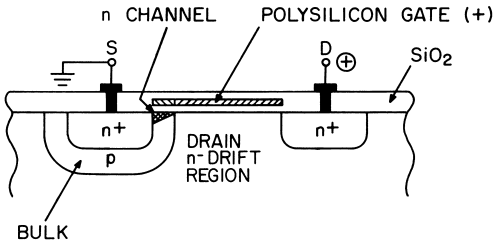


FIGURE 19 Schematic cross section of a horizontal DMOS.

length is obtained by the subsequent diffusions of the p-type bulk region into the n-type silicon crystal and the n-type source into the p-type bulk. Both diffusions take place through the same oxide opening. Similar structures can be obtained by double implantation (DIMOS).

The purpose of the relatively long n^-drift is to allow the drain depletion region to extend mostly into it. In this way the drain avalanche breakdown voltage can be greatly increased. Figure 19 shows the structure of a horizontal type of DMOS. For high-power devices, a similar structure is used, but often with a vertical n^- region.

5. Dual Gate MOSFETs

In addition to the usual gate formed on the top of the device, there can be another gate within the device. These devices are called dual gate MOSFETs. Both gates control the channel potential (Fig. 20). The two gates act to shield the source region of the channel from the action of the drain, providing a high degree of isolation of the source and drain. As a result, short-channel effects are reduced within the device.

6. SiGe MOSFETs

A strained Si layer can be formed using an Si/SiGe heterostructure. The electron mobility within strained Si is significantly higher than that of bulk silicon, and about a factor of five higher than that in Si inversion layers. The enhanced mobility within the strained silicon layer can be used to increase the carrier mobility within a MOSFET device. A higher mobility in turn leads to a higher transconductance and thus a higher frequency of operation of the device.

G. Power MOSFETs

Power MOSFETs are intended to switch or amplify currents of the order of tens of amperes and to support high circuit voltages. To do this the device should exhibit a very small on-resistance and a small voltage drop when the drain current is at its peak value. This is necessary

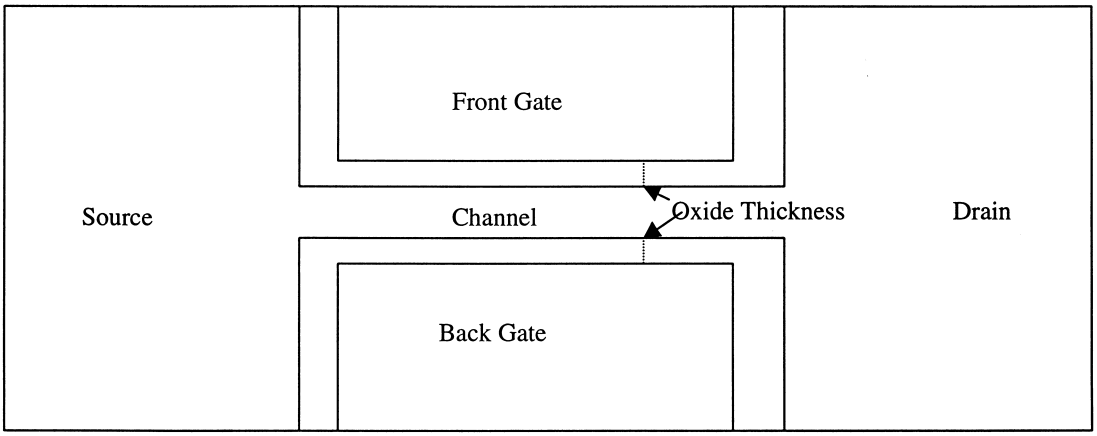


FIGURE 20 Sketch of the cross section of a dual gate MOSFET device. Notice that the channel is controlled by both the front and back gates.

to minimize power losses and keep the junction temperature below the maximum allowable limit of 150°C. The requirement of high drain breakdown voltage is in conflict with the low resistance requirement because it necessitates the use of high-resistivity silicon. For the same silicon crystal area, power MOSFETs are inferior to bipolar transistors from the point of view of on-resistance. On the other hand, they are superior as fast switches for such applications as switching-mode power supplies. As high input impedance devices they requrie less current drive, and their current gain does not degrade at high current levels. The MOSFET on-resistance has a positive temperature coefficient and therefore has a stabilizing effect, making the device resistant to "thermal runaway" instability. The so-called second breakdown, which severely limits the bipolar transistor safe operating area, is not as great a problem in the MOSFET.

Early power devices utilized a structure with a V-shaped gate obtained by anisotropic silicon etching in the manner described for the JFET in Section III.D. The device structure includes a vertical n^- drift region to support high drain voltages. Unfortunately, the V-shaped gate presented a breakdown problem so these devices have been almost entirely superseded by the DMOS (Section III.E.4) with a vertical drain.

The structure of a power MOSFET includes thousands of small MOSFETs integrated on the same silicon chip and operating in parallel to obtain a low on-resistance and therefore high current handling capability. Figure 21 shows the cross section of two n-channel enhancement-type devices forming two distinct cells, each consisting of a hexagonally shaped (as viewed from the top) n^+ source and a similarly shaped polysilicon gate embedded in the oxide. An inversion layer is formed around the hexagonal source; the electrons from the n channel flow into the n^- drift region, and then to the n^+ substrate and drain contact. The vertical drain design permits excellent surface utilization and facilitates device heat sinking.

The source metallization shorts the source perimeter to the heavily doped p^+ surface of the p^- region, providing a shunting resistance across the emitter (source) and the base (p bulk) of the parasitic transistor, reducing its current gain and the probability of second breakdown. The packing density of these structures can be as high as 10^5–10^6 cells per square centimeter with channels 1 μm long; total channel widths can be measured in tens of meters.

1. Conductivity-Modulated Field-Effect Transistor

The current capability of the power MOSFET is considerably extended in the conductivity modulated field-effect transistor (COMFET), which is also known as the insulated-gate transistor (IGT). In this device the n^+ drain region of the power MOSFET (Fig. 21) is replaced by a very thin (2-nm) n^+ film deposited on the p^+

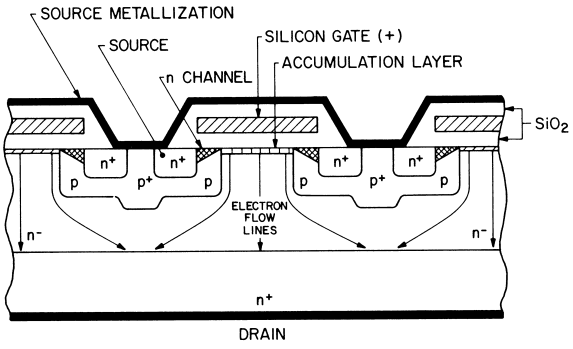


FIGURE 21 Schematic cross section of a multi-cell power MOSFET (vertical DMOS).

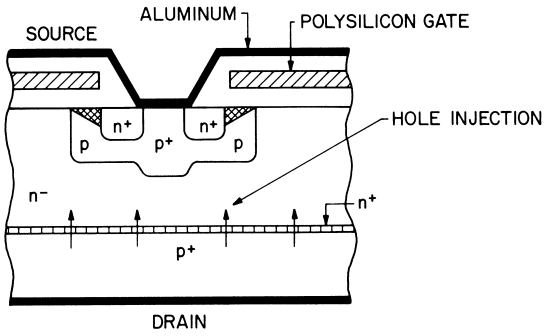


FIGURE 22 Schematic cross section of a COMFET.

substrate (Fig. 22). This creates in essence a four-layered ($n^+ - p - n^- - p^+$) thyristor-like structure activated by a MOSFET. When the device is turned on by the application of the positive gate voltage V_G greater than the threshold voltage V_T, electrons start flowing to the drain region, biasing the drain $n^- - p^+$ junction in the forward direction. The drain (the anode of the thyristor) consequently injects holes into the n^- drift region. To maintain charge neutrality, electrons enter the n^- region to balance the positive charge of holes, thereby increasing the n^- -region conductivity and decreasing the device on-resistance by a factor of 10 as compared with that of the MOSFET of the channel width Z.

The purpose of the p^+ region in each cell's center combined with the metallization shorting the n^+ and p^+ regions is to reduce the n^+ emitter into the p base (bulk) and the current gain α_1 of the $n^+ - p^+ - n^-$ transistor. The 2-nm-thick, heavily doped n^+ layer on top of the p^+ drain region lowers the emitter efficiency of the $p^+ - n^+ - n^-$ junction, and the therefore the current gain α_2 of the $p^+ - n - p^+$ transistor. Keeping the sum of the two current gains $\alpha_1 + \alpha_2$ below unity prevents latching of the $n - p - n - p$ device at high current levels.

SEE ALSO THE FOLLOWING ARTICLES

BIPOLAR TRANSISTORS • INTEGRATED CIRCUIT MANUFACTURE • OPTICAL AMPLIFIERS (SEMICONDUCTOR) • SEMICONDUCTOR ALLOYS • THIN FILM TRANSISTORS

BIBLIOGRAPHY

Blicher, A. (1981). "Field-Effect and Bipolar Power Transistor Physics," Academic Press, New York.

Blicher, A. (1982). "Physics of semiconductor power devices," *Rep. Prog. Phys.* **45**(4). The Institute of Physics, London, England.

Brennan, K. F. (1999). "The Physics of Semiconductors with Applications to Optoelectronic Devices," Cambridge University Press, Cambridge.

Brews, J. R. (1981). "Physics of the MOS transistors," *In* "Silicon Integrated Circuits," (D. Kahng, ed.), Academic Press, New York.

Milnes, A. G. (1980). "Semiconductor Devices and Integrated Electronics," Van Nostrand Reinhold Company, New York.

Nicollian, E. H., and Brews, J. R. (1982). "MOS (Metal–Oxide–Semiconductor) Physics and Technology" Wiley, New York.

Richman, P. (1973). "MOS Field-Effect Transistors and Integrated Circuits," Wiley, New York.

Sze, S. M. (1981). "Physics of Semiconductor Devices, " 2nd ed., Wiley (Interscience), New York.

Tsividis, Y. (1987). "Operation and Modeling of the MOS Transistor," McGraw-Hill, New York.

Zambuto, M. (1989). "Semiconductor Devices," McGraw-Hill, New York.



Integrated Circuit Manufacture

Peter C. Sukanek
University of Mississippi

James G. Ryan
IBM Microelectronics

Akihisa Sekiguchi
IBM Microelectronics

William R. Wilcox
Clarkson University

I. Introduction
II. Production of Semiconductor Grade Silicon: Sand to Wafers
III. Oxidation
IV. Lithography
V. Etching
VI. Diffusion
VII. Ion Implantation and Annealing
VIII. Chemical Vapor Deposition
IX. Epitaxy
X. Metallization and Contacts
XI. Planarization
XII. Packaging
XIII. Reliability

GLOSSARY

Acceptor Impurity in a semiconductor that accepts electrons to become ionized and produce holes. When the acceptor concentration exceeds the donor concentration, the material is called *p*-type.

Amorphous solid Solid with no long-range order; a glass.

Carrier lifetime Mean time between the formation of excess charge carriers (holes or electrons) and their recombination.

CMP Chemical mechanical polishing or planarization, a technique for obtaining flat substrates which combines chemical reaction and mechanical abrasion.

Crystal Solid with a long-range, regular, repeating, three-dimensional order.

Damascene Processing method, usually applied to metals, in which a film is deposited on a patterned substrate. Excess film is removed, often using CMP.

Donor Impurity that donates conduction electrons to the semiconductor to become ionized, producing *n*-type semiconductor when present in excess.

Dopant Impurity intentionally added to a semiconductor to increase its electrical conductivity by producing excess holes (*p*-type) or electrons (*n*-type). For silicon, typical dopants are boron (producing holes) and phosphorus, arsenic, or antimony (producing electrons).

Epitaxy Two crystals in intimate atomic contact with a preferred crystallographic orientation between them.

Interstitial An atom in a crystal at a site other than a lattice site.

Parasitic capacitance Additional capacitance in a transistor caused by interactions with the semiconductor substrate. This reduces the switching speed of the device.

***p-n* junction** Boundary between *p*- and *n*-type regions in a semiconductor.

Substitutional Impurity atom in a crystal residing on a lattice location.

Superlattice Crystal lattice in which the composition changes periodically from one atomic plane to the next.

Via Hole cut in an insulator to permit contact of one level of metal to another.

INTEGRATED CIRCUITS (ICs) are electronic devices consisting of transistors, diodes, capacitors, and resistors made and interconnected on a single solid substrate material. Integrated circuits are used primarily for logic operations and for memory. In ultralarge or giga scale integration (ULSI or GSI), hundreds of millions to billions of these electronic devices are interconnected on a chip of solid typically measuring a centimeter on each side. While the vast majority of ICs is made from single-crystal silicon, a small percentage is made from silicon—germanium mixtures, and compounds from groups III and V such as gallium arsenide, as well as compounds from groups II and VI. Research is underway on the use of polycrystalline materials. The methods currently used to fabricate or manufacture IC chips from single crystal silicon are outlined. However, most of the techniques are applicable to compound semiconductors as well, although the details are different. Integrated circuit fabrication is a very competitive business, and the technology is evolving rapidly. The common thread is the trend to smaller dimensions, necessitating lower processing temperatures, shorter processing times, new materials, new techniques, and scrupulous attention to cleanliness, in terms of both particulates and impurities in process chemicals and the atmosphere.

I. INTRODUCTION

Fabrication of IC chips from single crystal silicon wafers is a multistep process requiring care, cleanliness, and precision at each step. In the crystal, *p-n* junctions, capacitors, and resistors are produced by the controlled introduction of impurities by ion implantation, epitaxial growth, or diffusion. Electrodes, insulators, contact regions, and conductors are formed or deposited in or on the crystal. These steps are not performed simultaneously; they must be repeated several times in the various "levels" of the production sequence.

As a very simple example of the fabrication of a single device in a silicon crystal, consider the steps required to make the *n*-type, enhancement mode, metal oxide semiconductor field effect transistor (MOSFET) illustrated in Fig. 1. This device is normally "off"; no current can flow between the source and the drain. When a positive potential is supplied to the metal gate, electrons are attracted to and holes are repelled from the silicon-silicon dioxide interface. When this gate potential is sufficiently large, the electron concentration exceeds the hole concentration, and so a layer of the substrate near the surface is inverted to *n*-type. Current can then flow between the source and the drain if a potential difference is applied between them. For a given drain-source voltage, the larger the gate potential, the larger the current flow until a saturation point is reached.

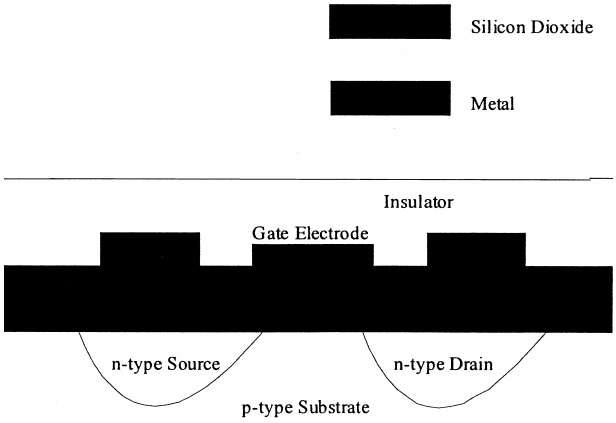


FIGURE 1 Schematic diagram of an enhancement mode MOSFET. Dimensions not to scale.

The fabrication sequence for this MOSFET is illustrated in Fig. 2. First, the surface is oxidized to produce a protective layer of silicon dioxide. An organic, photosensitive material called a photoresist is deposited on the oxide. The solubility of this organic material is altered by exposure to radiation. The surface is then irradiated in a particular pattern to expose the resist. After the soluble portion of the resist is removed in a developer, the resulting pattern

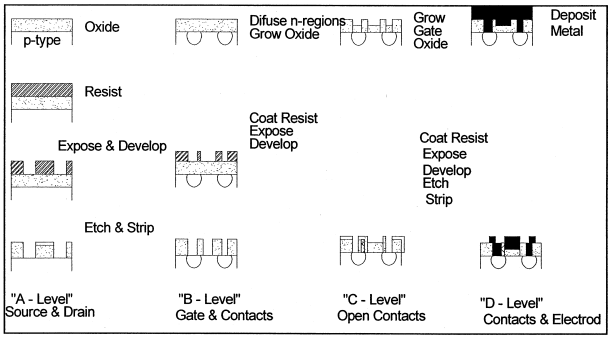


FIGURE 2 Fabrication sequence for the MOSFET shown in Fig. 1. A real process involves the production of tens to hundreds of millions of devices on a chip at one time, with tens or hundreds of chips on a wafer.

is transferred to the underlying oxide layer in an etching operation. The organic material on top of the oxide must resist the etchant. The result is the A level, and its purpose is to define the source and the drain regions in the semiconductor. The exposed silicon is doped with an *n*-type material such as phosphorus in a diffusion or ion implantation process. Another layer of oxide is then grown on the silicon. In the second, or B, level the gate and contact regions are defined. Again, a photoresist layer is deposited on the oxide, irradiated, developed, and the exposed oxide etched away. A thin, dense, high purity oxide is formed in the exposed regions. At level C the contact regions are exposed. A layer of metal, usually aluminum or an aluminum mixture, is deposited. A final mask level serves to define the source and the drain contacts as well as the gate electrode. A passivating layer of insulating material is deposited to protect the device and isolate it from other metal lines.

This simple example required four levels; modern integrated circuits are processed through 15 to 20 levels and require several hundred separate processing steps. While silicon dioxide is still the material used as gate dielectric, other, high permittivity materials, such as barium strontium titanate are being investigated for reduced dimension devices. Aluminum gate electrodes are typically not used. Rather, doped polycrystalline silicon (polysilicon or poly) or metal silicides have replaced them. In most applications, two or more insulating layers of different materials are employed. Significant processing occurs above the surface of the semiconductor material. Multiple metal lines are employed, one stacked above the other, to save space on the semiconductor surface. Special precautions must be taken to isolate one device from an adjacent one as they are placed closer together. The smallest image size or "groundrule" is currently smaller than a quarter of a micrometer, a size approximately one hundredth the diameter of a human hair. Consequently, the ability to overlay one level on a previous one must be controlled to much less than 0.1 μm.

It is important to realize that the same operations described above are occurring almost simultaneously all over the silicon wafer. Literally millions of these field effect transistors (FETs) are made at locations determined by the irradiation pattern of the resist material. In addition, a number of different devices (diodes, *p*-type MOSFETs, capacitors, etc.) are fabricated in each integrated circuit. Some processing steps for these devices might coincide with those for the FETs, while other devices would require additional levels. Several dozen to several hundred integrated circuit chips, in most cases all of them identical, are simultaneously made on a wafer of single crystal silicon.

In the following sections, each of these various processing steps is examined in more detail.

II. PRODUCTION OF SEMICONDUCTOR GRADE SILICON: SAND TO WAFERS

Elemental silicon is produced by heating a powdered mixture of carbon and silica (SiO_2 from quartz sand or crystals) in a carbon arc furnace. By-products include CO and SiC. The silicon powder produced is metallurgical grade (MG), and is used in large quantities for alloying with iron in steel production. Metallurgical grade silicon contains many impurities and must be purified before it can be used for IC fabrication.

The most common method of producing semiconductor-grade (SG) silicon from MG silicon is known as the Siemens process. In a fluidized bed at about 1100°C, MG silicon powder is reacted with HCl gas to produce trichlorosilane ($SiHCl_3$), with other chlorinated silanes as by-products. The $SiHCl_3$ is purified by distillation and then reacted with H_2 over an electrically heated silicon rod to deposit SG silicon. The purified silanes are also used for deposition of single crystal and polycrystalline silicon films in IC fabrication, as described in Section VIII.

An alternate technology for the production of SG silicon is the Komatsu silane process. Here, MG silicon powder is reacted with H_2 and silicon tetrachloride ($SiCl_4$) in a fluidized bed to produce SiH_2Cl_2 and other silanes. The gas mixture is catalytically reacted to $SiCl_4$ and SiH_4. After distillation, the $SiCl_4$ is recycled to the initial reactor and the impurities are discarded. The purified SiH_4 is pyrolized in a fluidized bed to give SG silicon powder. The H_2 by-product is also recycled. The impurity content of the SG silicon from the silane process is somewhat less than that from the standard trichlorosilane process, with boron and phosphorus concentrations on the order of 0.01 and 0.06 ppb, respectively.

About 80% of the SG silicon is converted into single crystal ingots by the Czochralski (CZ) technique (sometimes called "crystal pulling" or the "Teal-Little method"). Some SG silicon is formed into solid rods and used for floating zone melting crystal growth (float zoning).

In CZ growth, a silicon melt is formed in a fused silica (quartz) crucible, as shown in Fig. 3. Surrounding the crucible are typically a graphite crucible support, a graphite resistance heater, thermal insulation, and a water-cooled stainless steel container. A long, thin seed is cut from a previous crystal. Its end is dipped into the melt and melted back. The heater temperature is reduced to cause growth, and the seed is slowly withdrawn, with rotation. The crystal diameter is at first decreased to a few millimeters to reduce thermal stresses and allow any dislocations present to grow out. Then thermal conditions are adjusted to bring the crystal out to its desired diameter. The entire operation is performed in an inert atmosphere of argon or helium at reduced pressure.

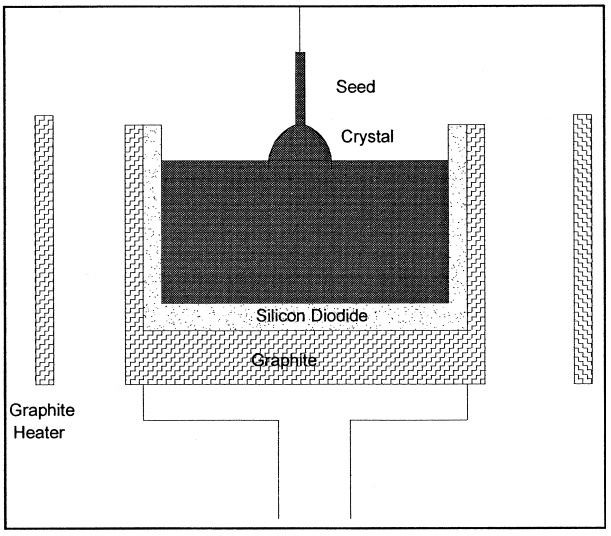


FIGURE 3 Schematic diagram of the CZ crystal growth apparatus.

The largest sized wafer in production use today is 200 mm. The next size, 300 mm, is slowly being introduced. Wafers as large as 400 mm, completely dislocation free, have been grown. Crystals of both (100) and (111) orientation are available.

Higher purity crystals are available from float-zoned silicon, although the maximum commercial diameter is 125 mm. In the float-zone process, the bottom end of a polysilicon rod is brought into contact with an O-shaped radio frequency coil with an internal diameter much less than the rod (referred to as a "needle-eye" heater coil). The rf field induces eddy currents in the silicon, which melts, flows down through the "needle-eye" onto a seed crystal below the coil, and solidifies. The heater traverses the entire rod. Since the melt is not in contact with a crucible, the impurity content of the wafers is determined solely by that of the starting material. Such high purity is required by silicon infrared detectors and by high-voltage power rectifiers and thyristors.

After growth, the cylindrical crystal is centerless-ground to remove irregularities and to reach the desired diameter. Flats to identify orientation and doping and sometimes alignment grooves for lithography are also ground onto the sides of the ingots. The ingots are then sliced "bologna-style" into wafers. For this, an inside-diameter saw, in which cutting is performed on the inside surface of the ring-shaped blade, or a wire saw is used. The wafers are chemical-mechanically lapped to a very flat, damage-free surface by use of colloidal silica particles in a gentle etchant liquid such as hypochlorite in water. After cleaning to remove both organics and inorganics, the wafers are shipped to the semiconductor fabrication plant. The final cost for a 200 mm wafer is on the order of $100 and depends on doping level, impurity content specifications, surface treatment, etc.

The silicon crystal is in the form of the diamond crystal, with the unit cell shown in Fig. 4. This interpenetrating face centered cubic structure is characteristic of other semiconductors, such as germanium or gallium arsenide. There are several important defects that occur in these crystals and are manifest as point, line, area, or volume imperfections in the crystal. Point defects are always present. Even if there were no impurities in the silicon, atoms would still move in the lattice creating vacancies (empty spots on the lattice) and interstitials (atoms residing in-between lattice positions). Impurity atoms, whether or not intentionally introduced, can also substitute for silicon on the lattice ("substitionals") or reside in interstitial locations. Impurities, since they are not the same size as the silicon, will introduce strain in the lattice.

Line defects or dislocations are deleterious for several reasons. They reduce the strength of a material, making it easy for wafers to bow or warp under the thermal stress caused by inserting and removing them from hot furnaces in oxidation, diffusion, and hot-wall chemical vapor deposition. Warpage beyond a small amount can make lithographic operations difficult. In severe cases, the wafers can even shatter. Diffusion of impurities tends to be faster down dislocation cores, causing irregular *p-n* junctions, which ruin device performance. Because there is a compressive stress on one side of the dislocation and tension on the other side, impurities are attracted to these regions. Transition metal impurities, for example, can occupy the dislocation core after heating and cooling. Such decorated dislocations are harmful to performance if they occur within active regions of the device. On the other hand, if the dislocations are on the back side of the wafer, far from the devices, they serve to attract ("getter") harmful impurities and increase the minority carrier lifetime.

Twins, stacking faults, and grain boundaries are area defects in the crystal. The boundary region between

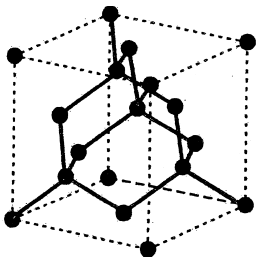


FIGURE 4 Unit cell of the diamond crystal. [From Dillon, R. O. (1997). *In* "Synthesis and Properties of Advanced Materials," (C. McHargue, *et al.*, eds.), Kluwer, New York.]

several different growing crystalline grains is called a grain boundary. Since, by their very nature, they only occur in polycrystalline materials, grain boundaries do not exist in single crystals. Stacking faults and twins, on the other hand, do exist. These can be created during growth or processing. Silicon crystals are three dimensional, with a particular sequence of atoms built up one atop another. If extra planes are added or removed from the crystal, this regular sequence is disrupted. A stacking fault involves a single extra (or omitted) plane, whereas twins involve many planes. At the twin boundary, there is a mirror image misorientation of the lattice. Stacking faults in the crystal are generated because the equilibrium concentration of interstitial silicon atoms decreases rapidly as the temperature falls following growth. If dislocations are present, they absorb excess interstitials. In the absence of dislocations, the interstitials cluster together to form tiny (111) disks. The stacking faults form in a swirl pattern across the wafer and are revealed by certain crystallographic etches.

Volume defects involve local regions of different phases. These might be regions of amorphous silicon, or regions of a precipitated element or compound such as oxygen. Oxygen is present in large quantities in CZ silicon because the silica crucible dissolves in the melt. Again, the solubility of oxygen in the crystal decreases rapidly as the temperature falls following growth. Small oxygen concentrations and short times at high temperature yield oxygen clusters, which strengthen the silicon and reduce warpage problems because of solution hardening. High oxygen concentrations and long annealing times yield oxide precipitates, which punch out dislocations and increase warpage. Warpage is a minimum at oxygen concentrations of 20–25 ppm. With proper heat treatment, the oxygen atoms diffuse out from the surface where devices are to be formed and precipitate only in the bulk of the silicon. These precipitates act as internal getters for harmful impurities and increase the excess-carrier lifetime. It is common practice to supply wafers which have been processed in this manner.

Carbon is also present in silicon. It comes from contamination of the SG silicon itself, from organic contaminants in the furnace and its components, from reaction of the silica crucible with its graphite support to produce CO and CO_2, and from reaction of the SiO evaporating from the melt with the hot graphite which also produces CO and CO_2. Carbon enhances oxide precipitation and swirl formation. When the concentration in the melt becomes large, SiC particles may form and cause the silicon to develop dislocations and new grains. In CZ growth, the carbon content is controlled to the ppm range by operating at a partial vacuum, 4–750 torr ($500–10^5$ Pa).

The impurity concentration in CZ crystals oscillates along their length. These striations occur because the growth rate oscillates as the crystal is rotated and pulled in a nonuniform temperature field. More importantly, there are large convection patterns in the silicon melt during CZ growth caused by the crystal and crucible rotations and by buoyancy. This convective flow leads to nonuniformities in both the axial and the radial directions. Recent work has shown that they may be greatly reduced by applying an external magnetic filed to the crucible. Because the melt is conductive, the external field increases the effective viscosity of the fluid, decreasing convection. Magnetic melt stabilization also reduces the amount of oxygen impurities since the reduction in convection reduces the dissolution rate of the crucible.

Compound semiconductors, such as gallium arsenide (GaAs) and cadmium telluride (CdTe), are often grown by the horizontal Bridgman (HB) technique. A polycrystalline material is placed inside a horizontal tube or boat with a seed crystal at one end. The boat is sealed to prevent evaporation of one of the components. (The vapor pressure of As over GaAs is 1 atm.) An external heater slowly traverses and melts the poly feed, which then crystallizes onto the seed. Crystals produced in this way are D-shaped and are contaminated by the boat material.

Round, single crystal wafers of III-V semiconductors have been made by the high-pressure, liquid-encapsulated CZ method. Gallium and arsenic are heated in a boron nitride crucible with a tightly fitting boric oxide (B_2O_3) cover. The cover melts at about 400°C, reducing As evaporation. Crystal growth proceeds in the same way as described above for silicon, with an As overpressure of 20 atm. Much higher purity crystals are obtained by this method compared with the Bridgman method.

III. OXIDATION

Oxides and other dielectrics serve many functions in IC fabrication and in the completed IC. They can serve as a barrier or mask to diffusion, ion implantation, or epitaxial growth. They can be used as a gate dielectric for a field effect transistor or help protect the completed device from contamination by the environment. Such dielectric layers can be formed either by reaction of the solid substrate with the surrounding gas(es) (oxygen, for example) or by chemical vapor deposition (CVD) of films on the substrate. CVD, both thermal and plasma enhanced, is discussed in Section VIII.

Oxidation can be accomplished in several ways. One method uses a batch furnace type reactor where more than 100 wafers can be loaded at a time. An alternate, single wafer reactor also exists. Here, a wafer is heated in an oxygen atmosphere using an array of high-intensity lamps.

The processing time is measured in seconds, rather than hours as in the case of the furnace. A related method, rapid thermal annealing, is discussed in Section VII.

Several types of insulating films may be grown on a silicon surface, including silicon dioxide (SiO_2), silicon nitride (Si_3N_4), and silicon oxynitirde. Silicon is oxidized either in nominally dry oxygen or in oxygen containing a large amount of water vapor. Growth occurs at the oxide-silicon interface. That is, the silicon atoms in the oxide are immobile while oxidizing species diffuse through the oxide to react at the silicon surface. One reason silicon is the major semiconductor in use today is the ease with which a stable, adherent oxide is formed. For silicon, the properties of "thermal oxide" formed by oxidation are superior to deposited oxide. Silicon nitride can be grown by exposing silicon to ammonia at approximately 1100°C, but the reaction is self-limiting at a few nanometers. Oxynitirde growth is achieved by exposing SiO_2 to ammonia at approximately 1100°C. The growth of silicon nitride and silicon oxynitride by direct nitridation can also be achieved.

Both the diffusion and the surface reaction influence the oxidation rate, with the surface reaction being rate-controlling at low temperatures and for short times. Oxidation rate increases when either the diffusivity of the oxidant or the surface reaction rate (or both) increase. Process conditions causing an increase in both diffusion and surface reaction include the use of wet oxygen rather than dry oxygen, the presence of HCl in the oxidizing ambient, increasing the pressure of the oxidant, and heavy substrate doping. The surface reaction is also sensitive to the crystallographic orientation of the substrate. The oxide growth rate is higher for crystal planes with higher silicon atom density [e.g., (111) > (100)].

The dielectric properties of oxide grown in dry oxygen are superior to that grown in the presence of water vapor, presumably because of the presence of OH groups and a corresponding reduction in the number of covalently-bonded bridging oxygen atoms. When the water content of the gas is suddenly changed, the oxidation rate and the oxide properties change within minutes to those appropriate to the new water content.

After oxidation, the silicon atoms plus the oxygen atoms occupy more than twice the volume occupied by the same number of silicon atoms before oxidation. That is, oxidation moves the silicon atoms farther apart. This has several important consequences. First of all, the expansion can create stresses in the oxide if the surface of the crystal is not planar because of prior patterning. Second, if a portion of the surface is covered, for example, by a polycrystalline silicon or silicide gate with oxide under it, the gate is forced up at the edges because oxidation only occurs there. (Oxidizing species can diffuse only for short distances under the gate.) Third, to create the additional space between silicon atoms in the oxide, silicon atoms are injected as interstitials into the underlying silicon crystal and vacancies in the lattice are consumed at the surface.

Another defect that can result from oxidation and other high-temperature processes is warpage. Warpage is caused by thermal stress during insertion or withdrawal of the wafers from a hot furnace and by formation of films on only one side of the wafer. The warpage problem increases with increased insertion/removal rate, increased wafer diameter, decreased wafer thickness, increased furnace temperature, and increased amount of oxygen precipitates in the silicon. In severe cases, the thermal stresses are sufficient to fracture the wafer.

The fluxes of interstitials into and vacancies out of the silicon during oxidation in turn manifest themselves in several ways. Because the equilibrium point defect (i.e., vacancies and interstitials) concentration increases with doping of both acceptors and donors, the interfacial reaction rate for oxidation increases with impurity doping. The interstitial flux into the silicon also causes an increase in diffusion rate for many impurities. If sufficiently large, the interstitial flux causes stacking faults to form and to grow. At the lower oxidation temperatures becoming common today, however, the interstitial flux can be weak enough that stacking faults emit interstitials and shrink.

Oxides may be characterized by many different methods, but optical and electrical characterization are the most popular. Visual characterization (film color, patterns) can serve as a coarse method to determine film thickness and uniformity. Another commonly used method is ellipsometry. An ellipsometer measures the reflection of monochromatic polarized light from a transparent film such as SiO_2 to determine film thickness and index of refraction. Electrical characterization is used to investigate charges existing inside the oxide and at its interfaces with the semiconductor and metal electrodes. These charges cause bending of the valence and conduction bands in the adjacent semiconductor, altering device performance. Mobile charges cause erratic and drifting device characteristics, leading both to yield problems and to premature failure.

The primary source of mobile charges in silicon oxide is sodium ion, although lithium and potassium ions also may contribute. Sodium is ubiquitous, being present in ordinary developers, solvents, and materials, and spread widely by human beings unless great care is exercised. Avoiding metal contamination prior to, during, and after oxidation requires continual vigilance. Operationally, this vigilance is reflected in wafer handling, as well as semiconductor processing equipment design and maintenance practices.

IV. LITHOGRAPHY

In lithography, a pattern is formed in a radiation-sensitive material called a resist. This pattern is then transferred to the underlying substrate in an etching or implantation operation. The substrate may be an insulating layer, a metal, or the semiconductor itself. This section discusses resist materials, processing, and exposure technologies.

A resist is a material whose solubility changes upon exposure to radiation. Consequently, resists may be classified by the kind of material (organic or inorganic), the solubility change (more or less soluble), and the wavelength or source of radiation used (ultraviolet, electron beam, X-ray, or ion beam). Two important requirements of the resist are that it has a high sensitivity to radiation, so that it can be quickly exposed allowing large wafer throughput and that it has a large contrast; that is, the difference in solubility between the exposed and the unexposed areas should be large. Of course, it should "resist" the etchant used when the pattern is transferred to the substrate.

A *positive* resist becomes *more* soluble after exposure; a *negative* resist becomes *less* soluble. Figure 5 illustrates this behavior. The differences in solubility are brought about by radiation-induced chemical changes in the resist during the exposure. In a positive organic resist, the radiation can either lower the molecular weight of the material (if the radiation is sufficiently energetic) or bring about a reaction which turns a dissolution inhibitor into a dissolution accelerator. In a negative organic resist, the radiation leads to an increase in the molecular weight of the material, or can form a dissolution inhibitor in the material. The size of an image is determined by the wavelength of the radiation used to produce it. Hence, as the minimum feature size shrinks, the wavelengths become shorter, and the resist must be sensitive at shorter wavelengths. Each new generation of lithography tool requires a new resist formulation.

Many of the short wavelength ($\lambda \sim$248–193 nm) resists in use today use "chemical amplification." The incident photons result in a very few chemical reactions. These reaction products, however, are catalysts to further reactions which occur when the wafers are baked at temperatures in the neighborhood of 100°C. Thus one reaction product formed in exposure leads to thousands of reactions during the bake. It is these final products which have solubility characteristics different from the original material. Depending on the formulation and the developer used, the resist may be positive or negative. These resists might include a variety of chemical species: a "film former," which is responsible for the formation of a uniform, thin film as well as providing the "resist" function; a photoactive compound, which absorbs the incident photons and initiates the reaction; and a dissolution inhibitor, which undergoes the catalytic reaction. While some systems might have three separate chemical components, others might contain only two, with the same species serving two roles (e.g., both film former and dissolution inhibitor).

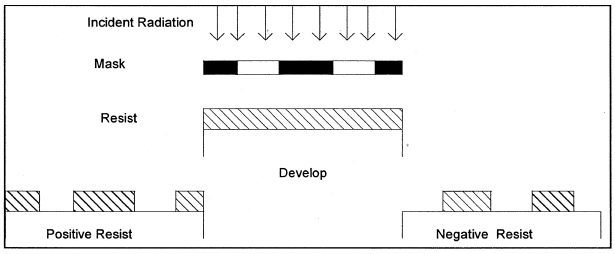


FIGURE 5 The difference between positive and negative resist on exposure to radiation.

Resists formulated for short wavelength, high-intensity exposure with X-rays or electron beams are solutions of high-molecular-weight polymer. The incident radiation is absorbed directly by the polymer, which then undergoes a scission or a cross-linking reaction. Scissions produce low-molecular-weight products which are easily removed in a solvent, or, if their vapor pressure is high enough, simply evaporate. Cross-linking increases the molecular weight and reduces the solubility. In either case it is desirable to start with a polymer of high molecular weight and narrow molecular weight distribution to insure high sensitivity and uniformity of response.

The processes involved with coating, exposing, developing, and removing a resist from a substrate are complex and are critical in many cases, especially when the images to be printed in the resist are 1 μm or less. Upon initial receipt, the wafer is first cleaned in a sulfuric-nitric acid or sulfuric-peroxide bath and then exposed to an oxygen plasma or ultraviolet light source. Such rigorous cleaning is usually impossible for wafers that have already undergone some processing. The surface is treated with an adhesion promoter such as hexamethyldisilazane (HMDS) to prevent the resist from lifting during lithographic processing. Thin, uniform coatings of resist are obtained by spin coating. Here the liquid resist is poured onto the wafer which is then spun at 2000–6000 rpm. The film thickness, which is on the order of 1 μm, increases with fluid viscosity and solvent evaporation rate and decreases with the spin speed. The wafers are baked at a temperature in the range of 70–100°C to evaporate some of the solvents remaining after the spin and to increase resist adhesion. After the bake, the wafers are exposed and, for chemically amplified resist, baked again before they are developed. The sensitivity and contrast of a resist are functions of the resist type and chemical composition, as well as the processing conditions, such as the developer temperature and develop time. In the case of the first chemically amplified resists,

it was found that trace (ppm or ppb) amounts of materials in the air could react with the exposure products between the exposure and the second bake, and significantly influence the final image. The exposure environment must be carefully controlled.

At this point the wafer is ready for the next processing step, which usually involves transferring the resist pattern to the substrate. In some processes where the wafer is exposed to relatively high temperatures or a severe environment, such as in metal etching or ion implantation, the wafers are first "hard" baked to temperatures in the range of 130–200°C. At these temperatures, typical positive resists undergo a number of chemical reactions which increase their molecular weight and decrease their solubility. At the same time, the increased temperature can reduce the viscosity of the resist before the reactions take place. Some deformation of the resist pattern, called resist "reflow," can occur.

In the exposure step the resist surface is illuminated with patterned radiation. In many cases, a broad beam of the radiation passes through a physical mask, illuminating some areas of the wafer and not others. (The processes required to produce a mask are discussed below.) The various types of exposure can be distinguished based on the distance between the wafer and the mask. In contact printing, the mask and resist surfaces are in intimate contact. Near contact or proximity printing is characterized by a mask-resist separation on the order of one to ten wavelengths of light. In projection printing there is a large separation between the two. Diffraction, or the ability to "concentrate" light at a point, is often the limiting factor in obtaining fine patterns. If a sufficiently narrow beam of radiation can be produced, the desired pattern can be written directly on the wafer surface. In this case the "mask" is a set of positioning instructions for the beam which have been stored in a computer. Such direct write systems are possible with electron or ion beams only.

Projection printing is the method most commonly used in the production of ICs where the minimum image size is 3 μm or less. There are two major types of systems available: reflection printing systems which employ mirrors to direct the light, and refraction systems which use lenses. In both types of exposure devices, the feature size is directly proportional to the wavelength of light used and inversely proportional to the numerical aperture (NA) of the optical system. For many resists, the proportionality factor is about 0.8. The numerical aperture is a measure of the light-gathering power of the system. When the system is in a gaseous medium, the NA is equal to the reciprocal of twice the f-number of the optics. The throughput of the system is usually fairly large, on the order of one hundred 150-mm wafers per hour. Modern optical systems allow the mask size to be reduced on the wafer surface. Commercially available printers have mask-to-chip size ratios in the range of 1:1 to 10:1. In modern refraction and reflection systems, the mask image is stepped across the wafer surface, and can be refocused at each step to accommodate wafer bowing (warpage). These devices are often called "wafer steppers" or "step and repeat" tools. The most common source of light is a high-pressure mercury arc lamp. However, the newest steppers employ an excimer laser operating at 248 or 193 nm in the deep ultraviolet.

The minimum size of the images or critical dimension that can be produced in the resist depends on a number of parameters in the exposure step. In addition to the wavelength of light and the NA of the optics, there are the intensity of the radiation, the wavelength spread, the spatial divergence of the beam, the nature of the substrate, the amount of scattering from the underside of the mask, the size and shape of the images in the mask, and the proximity of the images to each other. All of these factors are coupled with the resist properties and processing conditions discussed above.

The required critical dimension shrinks with each generation of IC. Table I summarizes approximate critical dimensions required for different generation dynamic random access memory (DRAM) chips. Chip size is given in megabits (M), the number of binary digits the chip can store. An indication of the exposure wavelength and NA of the exposure tool are also included, together with the depth of focus (DOF).

The depth of focus is a qualitative measure of how far from the focal point the wafer plane must be in order to obtain an acceptable image. As the distance from the focal plane changes, the image becomes more and more blurry. The depth of focus is directly proportional to the wavelength and inversely proportional to the square of the NA. As the feature size shrinks, the DOF decreases. Depending on the resist thickness, it may be impossible to have a "perfect" image on both the top and the bottom surface of the resist. The problem is amplified if the resist is deposited on substrates which are not flat but have some topography from previous processing. There have been several approaches to planarizing substrates. One is to deposit an intermediate layer, usually an organic

TABLE I Critical Dimensions for Different Generation DRAM Chips

DRAM	CD (μm)	Wavelength (nm)	NA	DOF (μm)
4 M	0.7	365	0.42	2.07
16 M	0.5	365	0.60	1.41
64 M	0.35	248	0.42	1.09
256 M	0.25	248	0.60	0.54
1G	0.13	193	0.60	0.32

material, which is planarized by surface tension forces. The time required to planarize decreases as the viscosity of the material decreases, as its thickness increases, and as the distance between substrate features also decreases. Because of this last factor, artificial, nonfunctional features are sometimes placed on the wafer referred to as a planarization block mask. Another technique which is becoming widespread is chemical mechanical planarization. This has several benefits and is discussed further in Section XI.

The wavelength region typically encountered with ultraviolet radiation is approximately 450–200 nm. A 20,000-eV electron beam, on the other hand, has an effective wavelength of only 0.06 nm. Diffraction effects are negligible. While it is possible to achieve much finer line widths with e-beams, scattering in the resist precludes reaching the diffraction limit. Both projection and direct-write e-beam systems are used. The former is in many ways similar to the refraction projection system discussed above. The "lenses" are magnets which direct and focus the beam. In direct-write systems, the electron beam, which can be either round or shaped, is used much as a pencil to draw out the pattern on the resist. While conceptually simple, this type of system requires large computer support to store the pattern and direct the beam. In addition, it is slow. However, it is a common method of making the photomasks used for optical projection printing.

Diffraction and scattering effects are eliminated when an X-ray source is used. In addition, dust and other particles which may be present are transparent to the radiation. Unfortunately, resist materials are also somewhat transparent, decreasing their sensitivity. Obtaining high-contrast masks is another area of concern. A gold pattern on a boron nitride substrate is one type which has been employed. It is also difficult to obtain small-size, high-intensity sources. Synchrotron radiation from electron storage rings has been investigated for this application. A related, short wavelength method is extreme ultraviolet (EUV) lithography. This method, currently under development, is based on the energy emitted from a plasma. Extreme ultraviolet wavelengths are on the order of 10 nm.

A final source of radiation for lithography is ion beams. As with electron beams, these can be used either in a projection system or in a direct-write mode. Spot sizes on the order of 0.1 μm and smaller have been obtained, and there is little scattering in the resist film to cause an increase in the linewidth. In addition, organic resist materials are usually an order of magnitude more sensitive to ion beams than to electron beams. A final significant advantage of focused ion beams is that they may ultimately allow "resist-less" processing at some levels. That is, the ions can be directly implanted into the substrate only in those areas where they are desired. Commercial application of such exposure tools is still several years in the future. However, ion beams are used for mask repair.

With the exception of the direct-write systems possible with electron and ion beams, a physical mask is required for all types of lithography. The mask must be made of a substrate which is transparent to the radiation, and a pattern of some opaque material is formed on this substrate. Since the quality of the resist images are determined by the quality of the mask, a mask set for a product represents a significant investment of time and money.

For ultraviolet lithography, the mask substrate is high-quality quartz glass. The opaque coating can be chromium or silicon dioxide. The first step in the mask-making process is often the creation of a "reticle," which is the desired pattern at 5–20 times the final size. In the past when image sizes were large, the reticle was made by photographically reducing a mechanically drawn plot of the pattern. As the size of images has decreased, the technology involved with creating a reticle has become more complex. The method most often used currently is a direct-write e-beam system which draws the pattern in the resist at about ten times the final chip size. This pattern is then etched into the opaque material underneath. The final 1X photomask is created in a step-and-repeat operation using the reticle as the mask. If the wafers are to be exposed in a reduction stepper, the mask can be made directly by the e-beam.

Recent advances in mask making have focused on adjusting the mask properties to sharpen the images created. Optical proximity correction (OPC) modifies the mask pattern to improve the wafer image. Features are added to the mask which are too small to print, but which, by interfering with transmitted light from other regions, modify the image. A related technology is the use of phase-shifting masks. Here, not only the transparency of the substrate, but also its ability to shift the phase of the light passing through it, is utilized. Near the image boundaries, a "phase shifter," which may simply mean a region of glass of a different thickness, is added. Light emanating from this region has a different phase than the light from nearby transparent regions. By carefully controlling this phase shift, the light rays can be made to interfere in such a way as to sharpen the image of smaller features.

Mask-making for e-beam, X-ray, and ion beam lithography requires different processes, each with different problems. Since electron beams interact strongly with matter, the transparent parts of the mask must be free of all material. Thus the masks are usually suspended metal foils with holes cut in them. For ion beams, (110) silicon has been used as the mask substrate since fairly large channels exist in this crystalline direction.

As the critical dimension decreases in size, the demand placed on lithography grows and the search for the "next

generation" process continues. Optical lithography has been able to meet the demands of current systems using innovative methods such as phase-shifting masks, optical proximity correction, and off-axis illumination. In the latter technique, the light incident on the mask is not normal, but arrives at an angle to the perpendicular. This decreases the resolution limit. It is unclear whether the future lithographic process will involve electron beams, EUV, X-rays, ion beams, or a shorter wavelength optical system.

V. ETCHING

In etching the lithographic pattern in the resist is transferred to the underlying layer. This pattern transfer can be accomplished in a solvent liquid or in a reactive gas. These methods are referred to as wet etching and dry etching, respectively. In addition to the etching environment, etching can be classified according to the shape of the etched profiles. In isotropic etching, the etch rate is the same in all directions. The sidewalls are in the shape of a segment of a circle. If the etch rate depends on direction, the etching is said to be anisotropic or directional. The most anisotropic type of etch gives vertical sidewalls; here, the etch rate in the lateral direction is zero. Either type of etch profiles can be desired, depending on the application. In order to insure good coverage of the etched regions in a subsequent film deposition process, an isotropic or close to isotropic profile is often desired. This might be the case, for example, when contact holes are being etched in an insulator. Vertical sidewalls are desired, on the other hand, when the opened regions must be in close proximity to each other. Isotropic and vertical etch profiles are illustrated in Fig. 6.

Figure 6 also shows how these profiles change after a 100% overetch, i.e., after the wafers are left in the reactive environment for twice the time required to etch the film. Isotropic etching clearly leads to a loss of image resolution during overetch. To account for film thickness and etch rate variations in any real process, some overetching is always required. To maximize the use of the wafer surface by devices, vertical etching and deposition processes which can accommodate these profiles, are employed.

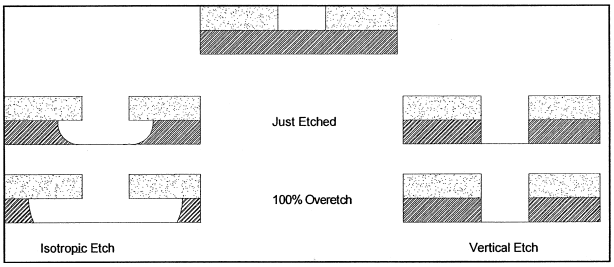


FIGURE 6 An illustration of isotropic and directional etching.

Wet etching is usually isotropic, although anisotropic etchants are available for some crystalline materials. The etch rates depend on the film doping, the temperature and concentration of the etchant, and the agitation of the liquid.

The most common application of wet etching is to open or to clean contact holes in silicon dioxide. This is usually accomplished in a buffered hydrofluoric acid (HF) solution. Isotropic silicon etching is performed in a mixture of nitric and hydrofluoric acids. The (110) planes in silicon are preferentially dissolved in aqueous potassium hydroxide, while hydrazine preferentially dissolves the (111) planes. Hot phosphoric acid is an etchant for silicon nitride and some of the metal silicides. Many organic resists can be removed by typical organic solvents. However, if the resist has undergone a high-temperature bake or been used as an ion implant mask, a much stronger oxidizing solvent is employed. Hot sulfuric-nitric acid mixtures, or mixtures of sulfuric acid and hydrogen peroxide ("Piranha etch") have been used.

Wet etching is primarily used for stripping, or completely removing, films. Pattern delineation is normally achieved by etching in a reactive gas in a process usually called plasma etching. This has several benefits: (1) it is generally cleaner than wet etching, (2) the inlet gases are safer to use, (3) the operation is safer, and (4) the effluent is easier to clean and dispose. The major benefit of plasma etching, however, is the ability to obtain directional profiles with relatively high etch rates.

As the word is used in the electronics industry, a *plasma* is a reactive gas environment formed from a relatively unreactive gas in a high-frequency electric field. The reactive species in the gas include ions and free radicals (atoms or groups of atoms containing unpaired electrons). At any temperature, there are some ions and radicals in a gas as a result of dissociation reactions. As the temperature of the gas increases, the equilibrium concentrations of these species also increase. To obtain a concentration that is large enough to cause measurable etch rates in an otherwise unreactive gas, the temperature would have to be prohibitively high. For example, temperatures at the melting point of silicon (1412°C) are required to cause a significant concentration of oxygen atoms in O_2. In a plasma, however, the temperature which would be measured by a thermocouple is approximately room temperature. For this reason, these plasmas are sometimes referred to as "cold" plasmas or, more properly, glow discharges.

The reactive species in a glow discharge are generated by collisional processes in the gas. In the presence of an electric field, positive ions are accelerated toward the cathode. Electrons and negative ions are accelerated toward the anode. These ions and electrons collide with other gas molecules generating more ions and electrons, dissociating stable molecules into reactive radicals, and exciting

gaseous species into higher energy states. When these excited species relax to the previous state, they may emit a photon whose energy is characteristic of the particular species and excited state. If the photon is in the visible part of the spectrum, the gas "glows." This is how neon lights work.

A glow discharge is not at equilibrium. The translational temperature of a gas is a measure of the kinetic energy or speed of the gas molecules. Since the uncharged species are not accelerated in the electric field, their translational temperature is relatively low. On the other hand, the electrons and ions are accelerated in the field, and so their translational temperatures are higher. Indeed, in high-frequency discharges the electronic temperature is ~30,000 K; the ion temperatures are much lower since the higher mass of the ions prevents them from following the field as closely as the electrons. The discharge produces a "solution" of reactive species (the ions and radicals) in an unreactive diluent (the unreacted gas molecules). The concentration of radicals is quite high, perhaps 30–50%, while that of the ions is low, approximately 0.1%. Some of the radical species can combine with the substrate material to form a volatile compound; others may react with each other to form a deposit.

In principle, a glow discharge can be formed in a DC electric field. In practice, a radio frequency (rf) field is employed. This allows the use of an insulator such as an oxide-covered silicon wafer as one of the electrodes. In addition, high-frequency fields can sustain the discharge at lower gas pressures than DC fields. The frequency most often used is 13.56 MHz, although frequencies from the kilohertz to the gigahertz range are also used. Because the ions are so massive, they cannot respond as well as the electrons to the rapid changes in the field. In a time-averaged sense, the powered electrode acts as the cathode and undergoes bombardment by the positive ions. The grounded electrode acts as the anode and is bombarded by negative ions. Because of the requirement of electrical neutrality and the presence of free electrons, there are more positive ions than negative ions. Therefore, the cathode receives more ions than the anode. In general, as the excitation frequency increases, the flux of ions to the surface increases, and the ion energy decreases.

A wide variety of materials can be etched in a glow discharge. The inlet gas must be so chosen that the radicals formed in the plasma react with the substrate to form a product which is volatile at the ambient temperature and pressure in the plasma. For silicon dioxide the most common etch gas is CF_4 and mixtures of it with oxygen or hydrogen, although a number of other gases have been employed, such as SF_6 and NF_3. For silicon and silicon nitride, CF_4/O_2 mixtures are used. For aluminum, chlorine-containing etch gases such as Cl_2 and CCl_4 are employed because aluminum fluoride is not volatile, but the chloride is. To remove organic material, an oxygen plasma is used. A number of other gases are also employed.

By properly choosing the plasma chemistry and conditions it is possible to obtain a variety of etch profiles, from isotropic to vertical. Ions play a major role in this process. In general, vertical etching is possible only when the ions impinge normal to the surface. The ions appear either to influence the desorption of the products, or to create reaction sites on the surface. The former is sometimes referred to as the "surface inhibitor mechanism of anisotropy," and the latter, "surface damage induced anisotropy." For appropriate chemistries, some of the species present in the gas can form films which deposit on the substrate. The ions remove these films and leave the substrate open to chemical attack by other gas phase radicals. In any event the etch rate in the vertical direction is larger than that in the lateral direction in which there is a negligible ion flux. The degree of isotropy is affected by a number of parameters, in addition to the etch gas. These include the gas pressure, the amount of substrate to be etched (the "loading"), the discharge frequency, and the bias voltage on the wafer.

Any method of increasing the energy or the directionality of the ions increases the extent of anisotropy. Increasing the pressure increases the concentration of reactive species, but the increased number of collisions decreases the ion directionality, and the etch becomes more isotropic. Similarly, increasing the discharge frequency increases the ion concentration but reduces the energy, again making the etch more isotropic. High energy ions affect the electrical properties of the substrate as well as the etch rate. In magnetron etching, the plasma is confined in a magnetic field to increase its ionization efficiency without increasing the ion's energy. In this way, the etch rate is increased without sacrificing directionality or increasing the substrate damage. Electron cyclotron resonance (ECR) is a related technique which also increases the number of ions and reactive species. In some techniques, the plasma is generated in a location remote from the wafers to be etched, and the ion energy is controlled by extracting and accelerating them through a grid.

Several reactor configurations have been used. In a barrel or tunnel configuration, the wafers are loaded into boats similar to those used in oxidation furnaces. The boat is placed in a quartz tube, and the plasma is generated in the gas inside the tube using external electrodes. In many cases, a metal mesh etch tunnel (Faraday cage) is placed inside the quartz tube, with the wafer boat resting on the metal cage. The cage serves to eliminate ion bombardment of the wafers. In the parallel plate reactor system, the wafers are placed on either the grounded or the powered electrode, with the other electrode directly above them.

Some of these systems can accommodate several wafers on the electrode. In others, only one wafer is etched at a time to reduce wafer-to-wafer variation. If the wafers are placed on the powered electrode, they undergo ion bombardment in addition to being exposed to the reactive radicals. This type of etch is usually referred to as reactive ion etch (RIE). With the wafers on the grounded electrode, there is little ion bombardment, and the process is often referred to simply as plasma etching. Finally, it is possible to power both electrodes and thereby adjust the energy with which the ions strike the surface. This configuration is called triode etching. A final reactor geometry is referred to as a hexode. The wafers are placed on a hexagonally shaped cathode, with four or five wafers per side. The inside wall of the bell jar cover forms the grounded anode.

As mentioned above, magnetron etching or electron cyclotron resonance increases the number of reactive species. These techniques are referred to as high-density plasma etching. The ion energy is independently controlled by a separate rf bias applied to the target. This leads to a high ion density at low pressures and provides independent control of the ion flux and energy.

As wafer size increases and the critical dimension decreases, there has been a gradual shift from processing multiple wafers to single wafer processing. This is true not only in etching operations, but in many areas of integrated circuit fabrication. Single wafer processing allows greater control of etching, increasing wafer-to-wafer, as well as within wafer, uniformity. One of the challenges associated with larger diameter wafers is insuring this uniformity over larger areas.

The etch rate is a function of a number of variables. The gas composition is often used to control the etch selectivity (that is, the ratio of the film to substrate etch rate), as well as the etch rate itself. For example, in etching silicon dioxide on top of silicon, it is desirable to have a large etch rate of the oxide as well as a large ratio of oxide to silicon etch rates so that the attack on the underlying silicon is minimized during any overetch that may be required. The addition of H_2 to CF_4 greatly reduces the etch rate of silicon but has a small effect on the etch rate of SiO_2. Similarly, the addition of oxygen to CF_4 increases the etch rate of silicon, but has little effect on that of silicon dioxide. In addition to the inlet gas composition, etch rates depend on the loading, rf power and frequency, gas pressure, and flow rate.

Plasma etching offers the potential of full automation of the etch process by monitoring the emissions from the plasma. As mentioned previously, one of the myriad reactions occurring in the plasma is the excitation of gas species to upper electronic states. These excitations are almost always accompanied by relaxations in which photons are emitted at particular energies. By monitoring the emission at an appropriate wavelength, it is possible to determine when the etching reactions are completed and thereby avoid etching the substrate. For example, when etching aluminum in a chlorine plasma, disappearance of the AlCl emission at about 260 nm indicates that aluminum is no longer being attacked. In addition to monitoring the end point, recent work on low-temperature ($\sim -100^\circ$C) plasma etching has indicated the possibility of removing material one atomic layer at a time. Such a technique, which gives the ultimate in control, is referred to as "digital etching."

Plasmas are used in several ways in the electronics industry in addition to etching. Plasma-enhanced chemical vapor deposition (PECVD) allows deposition of a film at lower temperature than is possible in conventional CVD (see Section VIII). Plasmas are also used in sputter etching and deposition, where the atoms sputtered off the cathode by noble gas ions are deposited on a substrate (see Section X).

VI. DIFFUSION

The transistors in an integrated circuit consist of small regions doped with different amounts of various impurities to produce *p-n* junctions. There are several ways in which a desired local variation in impurity content can be produced: by ion implantation, by growth of a new layer of crystal doped differently from the substrate, and by diffusion of an impurity.

Imagine that we have a wafer of silicon doped uniformly with boron at a concentration of 10^{16} boron atoms per cm^3 of silicon. Boron is an acceptor, making the silicon *p*-type. To produce a *p-n* junction we can diffuse a donor impurity such as phosphorus in from the surface by heating the silicon in a gas containing elemental phosphorus or a phosphorus compound. The *p-n* junction is the surface in the crystal where the boron and phosphorus concentrations are equal. If the boron concentration remains at its original value of $10^{16}/cm^3$ throughout the crystal, then the *p-n* junction is at the point where the phosphorus concentration is also $10^{16}/cm^3$. The spatial distribution of two impurities near the junction determines the electrical characteristics of the device.

Diffusion is carried out in resistance-heated tube furnaces like those used for oxidation. Usually a gaseous diffusion source is passed through the tube. Typical gases are BBr_3, B_2H_6, PH_3, AsH_3, $POCl_3$, $B(CH_3O)_3$, BCl_3, PBr_3, and Sb_3Cl_5 in a carrier gas such as H_2, He, or O_2. Alternatively, wafers may be exposed to ampules with diffusion sources, a liquid diffusant may be applied on the wafers, or a solid source may be used from which the impurity evaporates. Usually O_2 is used as a carrier gas so that oxidation

takes place at the same time as impurity diffusion, simultaneously preparing the wafers for the next lithographic step. The presence of oxygen also prevents halide diffusion sources from roughening the surface of the silicon by etching.

If the diffusion coefficient D for the impurity in the silicon is constant, if the concentration C_s of impurity at the surface of the wafer is constant, if the impurity diffuses in along the entire surface, and if the diffusion time t is too short to permit the impurity to diffuse to the other side of the wafer, then the variation of impurity concentration C with distance x into the substrate is given by:

$$C = C_o + (C_s - C_o)erfc(x/2(Dt)^{1/2}) \qquad (1)$$

where C_o is the initial concentration of that impurity in the crystal and *erfc* is the complementary error function, tabulated in many mathematics handbooks.

High surface concentration C_s is sometimes undesirable in IC fabrication. One way of lowering it is to use a lower partial pressure of the diffusion source in the furnace. However, it is often more convenient to use a two-step process, a predeposition followed by a drive-in. The diffusion source is present during the short predeposition but not during the long drive-in step. If the product of the drive-in time t_d and the diffusion coefficient D_d during drive-in is much longer than that for predeposition, $D_p t_p$, and if no impurity either enters or leaves the silicon during the drive-in, then at the end the final impurity concentration at any depth is given by:

$$C = (2C_{s,p}/\pi)(D_p t_p/D_d t_d)^{1/2} e^{(-x^2/4D_d t_d)} \qquad (2)$$

where $C_{s,p}$ is the surface concentration during predeposition. This is a Gaussian or normal distribution of concentration with depth.

Equations (1) and (2) incorporate many assumptions that are usually not entirely true in practice. Diffusion usually takes place through holes cut in a mask, so that C depends on lateral position as well as on depth. Except for very low C_s with no oxidation, D is not constant. It varies with C, with x, with the presence of other impurities and defects, and is altered when chemical reactions such as oxidation, nitridation, and silicide formation take place at the silicon surface. In addition, some impurities segregate preferentially to the oxide while others prefer the silicon, causing large changes in C_s if oxidation is also taking place during drive-in. Consequently, numerical methods are required to predict diffusion profiles accurately. Even then, uncertainties remain so that some empirical fine-tuning of diffusion conditions is required to achieve desired concentration profiles.

Three basic diffusion mechanisms are thought to be important in silicon, although debate on the details continues. In the vacancy mechanism, a substitutional atom moves to an adjacent empty lattice site (a vacancy). Antimony diffusion in silicon is believed to occur primarily through a vacancy mechanism. In the interstitial mechanism an atom moves interstitially (in between the lattice sites). Many metallic elements such as nickel, copper, and iron are transported through silicon by interstitial diffusion. In the interstitial-kick mechanism, a substitutional atom jumps into an interstitial position, moves interstitially for some distance, and then kicks another substitutional atom off a lattice site and occupies it. Phosphorous diffusion in silicon is an example of the interstitial-kick mechanism. Substitutional impurities appear to be able to form impurity-point defect pairs which diffuse together. Some impurities, such as gold, are thought to diffuse by two mechanisms simultaneously, leading to complex behavior.

When an impurity moves by a single mechanism, the diffusion coefficient depends on temperature:

$$D = D_0 \exp(-E/kT) \qquad (3)$$

where D_0 is a constant, E is an activation energy, k is Boltzmann's constant, and T is the absolute temperature. Thus log D versus $1/T$ plots yield straight lines as shown in Fig. 7. Interstitial impurities that form no bonds with the silicon atoms, such as Li, Cu, and Fe, have the highest diffusion coefficients in Si with the least temperature dependence (least E). Interstitial atoms, such as O, that chemically bond to neighboring Si atoms, have a diffusion coefficient three to four orders of magnitude lower and a

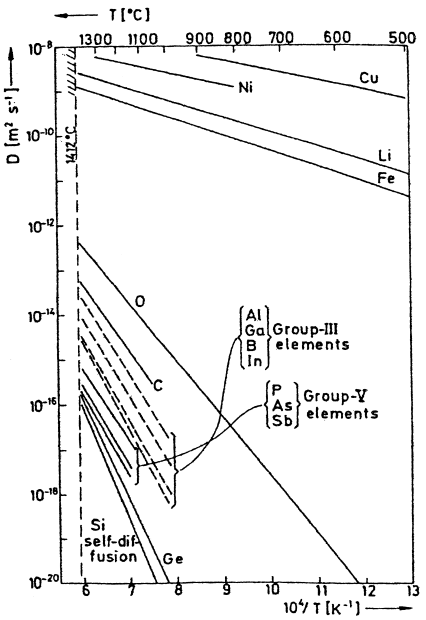


FIGURE 7 Diffusion coefficients for very low impurity concentrations in single-crystal, dislocation-free silicon. [Data from Frank, W., Gosele, U., Mehrer, M., and Murch, G., eds. "Diffusion in Solids II," Academic Press, Orlando, FL.]

much larger E value than the interstitials, because chemical bonds must be broken before they can move. Substitutional impurities move slower still because not only must they break chemical bonds to move, but they must also move via vacancies or must kick another atom from a lattice site.

In addition to the diffusion mechanisms listed above, there are other well-known cases of enhanced diffusion mechanisms. Further discussion on the variation of diffusion rates in the presence of point defects, multiple impurities, and electric field can be found in some of the references listed at the end.

Diffused areas are characterized both electrically and chemically. Electrical characterization is performed by making resistance measurements on diffused areas. The measured resistances are then compared to standards in order to determine the amount of dopant that is electrically active. The chemical concentration of dopant may be determined by secondary ion mass spectrometry (SIMS), where a doped region is bombarded by a primary ion beam that knocks secondary ions from the doped sample. The secondary ions are then mass analyzed in order to determine the sample composition. A comparison of the electrically active and chemical concentrations is useful for investigations of dopant precipitation phenomena and for evaluation of the effect of heat treatments of dopants.

VII. ION IMPLANTATION AND ANNEALING

Ion implantation is an alternative to diffusion to create doped regions. A source of the dopant is vaporized and ionized, and the ions are accelerated to high energies in an electric field. The resulting beam of (in most cases, positive) ions is directed at the wafers. The high-energy ions damage the crystal as they are slowed down by interactions with the lattice. In addition, many of the implanted ions do not reside on a substitutional location. Hence, an annealing step is required to produce electrically activated implanted ions and to remove the damage from the crystal. During implantations, electrical neutrality is achieved by conduction through the wafer, or, in the case of high-dose implants, by simultaneously scanning the wafer surface with an electron beam. In this section, implant systems, implant profiles, and various annealing methods are discussed.

Ion implantation has several advantages over diffusion. It is easier to control both the amount and purity of dopant delivered to the wafer, as well as the depth and profile of the dopant atoms. It is performed at much lower temperatures than diffusion and can accommodate a wide variety of possible dopants. The low temperature makes it possible to use resist as an implant mask, thus potentially eliminating processing steps. The major disadvantage of ion implantation is that a subsequent annealing step is required, although several implants can be annealed at the same time. In addition, it is impossible to grow an oxide at the same time as the implantation proceeds, so no "automatic" passivation is possible (unlike diffusion in an oxidizing atmosphere). However, an implantation can take the place of the predeposition step in diffusion. A high-temperature step then serves to anneal any damage as well as to drive-in the dopant.

A typical ion implanter can be divided into four parts: an ion source, an accelerator, a mass analyzer to remove impurities, and a delivery system. Ionization can be achieved in several ways. In electron impact ionization, a gas containing the dopant is supplied directly or is generated by evaporation of a solid. An electron beam then impinges on the gas to ionize it. Positive ions are extracted from the gas with an electrode. The extracted beam is accelerated in an electric field to energies on the order of 5–500 keV. High energy boron implanters are capable of delivering up to 3 MeV. In these systems, negative boron atoms are extracted from the source and accelerated to about 1 MeV. Three electrons are stripped from each atom in a nitrogen bath, and the resulting B^{2+} ions are accelerated away from the high voltage.

A number of ions of different species and charge states are always present in the beam. The particular ionic species desired is separated from the others by applying a magnetic field at right angles to the beam. Each ion follows a characteristic trajectory in the field depending on its mass and charge. The beam may be focused using other magnets. Typical beam spot sizes for implanters range from 0.01 to 10 cm. To insure a uniform dose, either the beam is scanned over the wafer surface or the wafers are scanned across the beam, or both. The implant may be performed in a batch mode, where many wafers are processed at once, or in a serial mode, where the wafers are fed to the implanter one at a time.

Contamination during implant can arise from several sources. The magnetic field is used to select an ion by its mass to charge ratio. If $^{28}Si^{+}$ is the desired ion, other species, such as $^{56}Fe^{++}$ or $^{12}C^{16}O^{+}$ may be implanted at the same time if iron or hydrocarbons are present in the source or the implanter was previously contaminated by these materials. If resist is used as an implant mask, it can outgas at the low pressure of an implanter. These atoms may be implanted, knocked into the substrate, or deflect desired ions from their path. This problem can be eliminated by a hard bake of the resist before entry into the implanter.

When the ion beam strikes the wafer, its kinetic energy is transformed into heat, which can cause the wafer to reach temperatures of several hundred degrees Celsius. This increased temperature can lead to localized regions

of self-annealing on the wafer. If the heat is not effectively removed, significant variations in as-implanted dopant activation are observed. In addition, if temperature gradients exist across the wafer, the extent of the damage created by the implant can vary across the wafer, also influencing electrical properties.

In a similar fashion, electrostatic charging of the wafer surface can lead to significant variations in the amount of ions different regions receive. This is especially a problem when the wafers are masked by an insulating film such as SiO_2. In this case, the flux must be significantly reduced, or the wafer surface must be continuously neutralized by a flood of electrons.

If the ion beam is directed onto an amorphous substrate, the concentration profile of the implanted dopant is Gaussian. The ion concentration at any location depends on the surface dose which in turn is directly related to the beam area and current, both measurable quantities. Typical beam currents are on the order of 1–10 mA. The mean distance into the wafer is called the projected range, and its standard deviation, the straggle. The Gaussian nature of the implant profile extends to the regions around the mask edge. Close to the edge, the dopant concentration starts to fall off. The standard deviation in this lateral direction is called the transverse straggle, and is slightly less than the straggle. The projected range, straggle, and transverse straggle, all increase with increasing beam energy and with decreasing ion size. These values also depend on the substrate, decreasing with increasing substrate density. For 100-keV implants into silicon, 99.5% of the boron atoms are within 0.5 μm of the surface, whereas for antimony the distance is 0.09 μm. A knowledge of the range and straggle variation with ionic species, energy, and substrate allows the determination of the mask thickness required to stop, say, 99.99% of the incoming ions. Silicon oxide and nitride, as well as resist, are used as implant masks.

As the ions enter the mask material, they can knock some of the mask atoms into the substrate. This is known as the recoil phenomenon. The depth and amount of the recoiled atoms depend on the energy and size of the incident ions, as well as on the thickness of the mask. Recoiled oxygen resulting from implants through a thin oxide layer is a particular problem. These oxygen atoms can initiate dislocations, act as gettering sites, and reduce the after-anneal carrier concentration.

When the ions enter the wafer, they lose energy to the silicon lattice and eventually stop by interacting with the electrons around the lattice atoms and with the atomic nuclei. The former is referred to as electronic stopping, and the latter as nuclear stopping. It is the nuclear stopping events which lead to crystal damage by kicking the substrate atoms out of their lattice positions. If the displacement occurs with sufficient energy, further recoil-type damage can be created deeper in the crystal. Low atomic mass implants create a different type of disordered region than heavy atoms. The light atoms, like boron, are initially slowed by electronic interactions. When sufficient energy has been lost, nuclear stopping predominates. Thus the disordered region is comparatively deep. In an implant with heavier phosphorus or arsenic atoms, nuclear interactions and the resulting crystal damage occur closer to the surface, and the recoiled atoms continue the damage farther into the crystal.

Sometimes the scattered ions may become "channeled" into a (110) direction of the lattice. There are large spaces in this direction, allowing the implanted ion sufficient room to travel long distances without stopping. This phenomenon is more likely to occur when the lattice is relatively undamaged, that is, along the tail of the distribution or at low doses. If tight control of the implant profile is required, the crystal is sometimes damaged with silicon atoms before the dopant implant. This creates an amorphous surface region and avoids channeling. Alternatively, compounds, such as BF_2^+, can be implanted. The increased mass increases the damage near the surface and reduces channeling.

Another method for reducing channeling is to tilt and twist the surface of the wafer relative to the incident ion beam. When mounted in this manner, the wafers no longer present an ordered surface to the incident beam but one that is more random. Variations in the amount of warpage or bow in the wafers, the flex in the mounting mechanism, and in the incident ion beam angle lead to variations in the extent of channeling across the wafers.

An annealing step is always required after an implantation. The wafers are heated to relatively high temperatures not only to repair the damaged crystal, but also to activate the dopant atoms, since they often do not stop on substitutional locations. The higher the temperature of the annealing step, the greater the recovery of the crystal. Two types of annealing behavior are distinguished, depending on the extent of crystal damage. If the implant created an amorphous layer, the anneal causes regrowth of the crystal through the mechanism of solid phase epitaxy. If there is a region of partial disorder, caused, for example, by a low-dose implant, the point defects and dislocations are more difficult to eliminate.

At $\sim$1100°C, the carrier lifetimes are completely restored, but there is still some residual crystal damage. In many cases, heating to high temperature is prohibited by other factors. Above 500°C there is significant dissolution of silicon into any aluminum which may be present; the aluminum melts at 660°C. As mentioned in Section VI, the diffusion coefficient of all impurities increases with increased temperature. All annealing is accompanied by diffusion. Throughout the manufacturing process, it is

important to stay within the "thermal budget" of the device: the maximum time at elevated temperature. If this budget is exceeded, the electrical properties of the device are ruined.

In addition to furnace annealing, lasers and incoherent light sources have been used. In laser annealing, an infrared or ultraviolet laser is scanned across the surface of the wafer. The energy is sufficient to activate the implanted atoms, but the heating and cooling is sufficiently rapid to minimize diffusion. However, large thermal stresses are set up in the wafer when localized regions are heated to high temperature. This results in the formation of dislocations. Indeed, backside damaging with a laser has been used as a means of gettering impurities. The most popular annealing method in use today is known as rapid thermal annealing. High-power arc lamps are used to heat the silicon wafer to 1100–1200°C in approximately 10 sec. These have the advantage of treating the entire wafer at once, increasing throughput and reducing the thermal stress. In addition, rapid thermal processing using this same type of equipment is used for oxidation and nitridation.

Ion implantation has been used to create silicon-on-insulator (SOI) structures. Buried layers of silicon dioxide or silicon nitride have been made by deep implants of oxygen or nitrogen. If the dose is sufficiently high, a region of SiO_2 or Si_3N_4 is formed several micrometers below the surface after heat treatment. Wafers with buried oxide layers made in this way are commercially available.

As the device size shrinks, the electrically active region in the silicon becomes thinner and closer to the surface. Implants as shallow as 30 nm or less are required. This requires the use of focused, high-current, low-energy (<5 keV) beams, the use of multi-atom species (BF_2^+), or pre-amorphization with a silicon implant. New technologies are also being pursued. Gas immersion laser doping (GILD) is a technique in which the wafer is placed in an atmosphere containing the desired dopant. A laser pulse is used to melt the wafer locally, and the dopant dissolves in the melt. Another method is plasma immersion ion implantation (PIII), where a high-density plasma is used to dope, rather than etch, the substrate.

VIII. CHEMICAL VAPOR DEPOSITION

Chemical vapor deposition (CVD) refers to deposition of a film on a solid substrate by chemical reaction of a gas. In CVD one may obtain a single crystal layer, a polycrystalline material, or an amorphous material.

The form produced depends primarily on the substrate, the deposition rate and the temperature. At high temperatures and low deposition rates on a clean single crystal substrate, the atoms produced by the chemical reaction have sufficient time and surface mobility to move around the surface of the substrate and find low energy sites. A single crystal film results unless the mismatch between substrate and film crystal lattices is too large. As the temperature is decreased or the deposition rate is increased, the adsorbed atoms have less opportunity to find a low energy site before being covered by more atoms. Nucleation of new grains occurs to produce a polycrystalline material. With continued decreases in temperature or increases in growth rate, the grain size becomes smaller and smaller until adsorbed atoms are not able to move at all before they are buried. An amorphous layer results.

Chemical vapor deposition is used to grow single crystal silicon films, polycrystalline silicon (poly) gate electrodes on oxide, phospho-silicate glass (PSG), or boro-phospho-silicate glass (BPSG) layers to getter impurities, silicon nitride and silicon oxide as masks and as dielectrics, various conductors such as metals and metal silicides as gate electrodes and interconnect lines, and so on.

Often a film may be deposited by more than one method. Silicon can, for example, be deposited by any one of the following reactions:

$$SiH_4 \rightarrow Si + 2H_2$$

$$SiH_2Cl_2 \rightarrow Si + 2HCl$$

$$SiHCl_3 + H_2 \rightarrow Si + 3HCl$$

These and other reactions can be thermally driven and/or enhanced by a plasma discharge process. Plasma-assisted deposition is commonly known as PECVD. Many PECVD films can be deposited at low temperatures, avoiding thermal budget problems.

Silicon dioxide may be deposited using many different reaction gases and reactor conditions. Relatively high temperatures are used when high-quality oxides with good insulating properties are required. Low-temperature oxides may be used in applications where compatibility with low melting point materials is required or in masking applications where insulator quality is not the prime concern.

The impurities previously listed as diffusion sources (Section VI) can be added to the gas stream to produce doped polycrystalline or oxide layers. Silicon compounds can be reacted with ammonia to produce silicon nitride and with nitrous oxide, to form silicon oxide. Metals can be deposited using metal halide or organometallic sources. It should be pointed out that these and many of the other gases used in IC manufacturing are very hazardous. Silane (SiH_4) is toxic and pyrophoric (combusts on exposure to air). Arsine (AsH_3) is very toxic, with concentrations as low a 6 ppm classified as "immediately dangerous to life or health."

Some materials may be selectively deposited only on areas of a substrate that chemically interact with the source

gas. For example, in "selective tungsten" technology, the substrate Si reduces the WF_6 source gas, whereas SiO_2 does not. This difference allows tungsten to be deposited only on exposed silicon. After the silicon is buried under the tungsten film, the tungsten metal itself promotes tungsten deposition only over the already deposited tungsten.

Chemical vapor deposition may be carried out either in a hot- or cold-wall reactor. The distinction between these two types depends on whether or not the wall plays a role in the dissociation reactions. A hot-wall reactor is substantially the same as those used for diffusion and for oxidation. The wafers are stacked in a slotted quartz holder, but usually farther apart than in oxidation and in diffusion to allow transport of reactants in between the wafers and transport of products out. Deposition occurs on the tube walls, as well as the wafers, necessitating occasional cleaning to avoid flaking off of these deposits. The yield is also reduced by this wall deposition.

In a cold-wall CVD reactor the wafers are usually placed on a graphite susceptor, which is heated inductively. A water-cooled coil wrapped around the outside of the glass wall carries high-frequency current that induces current to flow in the graphite, and the electrical resistance of the graphite causes it to heat up. Typically the graphite is in the shape of a vertical barrel with flat, slightly inclined walls inside a glass bell jar. The wafers are laid in circular depressions machined in the walls of the graphite susceptor. Alternatively, radiant energy can be used to heat the wafers. Because the glass container walls are cool, little or no deposition takes place on them. The primary limitation in cold-wall CVD is the small number of wafers that are processed in each run.

Low-pressure CVD (LPCVD) is a popular form of hot-wall processing. At atmospheric pressure uniform deposition rates across each wafer requires that the distance between the wafers be relatively large, allowing reactants to diffuse into these spaces. By lowering the pressure to a few torr or less, the mean free path of the gas molecules increases to several centimeters, allowing up to several hundred wafers to be processed in each run.

In general the CVD rate increases as the reactant partial pressures in the gas increase, the product partial pressures decrease, the temperature increases, and the flow rate through the reactor increases. Deposition is a multistep process, involving chemical reactions in the gas, chemical reactions at the deposition surface, diffusive and convective mass transfer in the gas, and diffusion on the surface. The chemical reactions are much more temperature-dependent than the diffusion steps. Consequently at low temperatures the deposition rate is reaction-controlled, with the logarithm of the deposition rate usually varying linearly with 1/T as shown in Fig. 8. At high temperatures the rate is mass transport controlled, with log rate vs. l/T

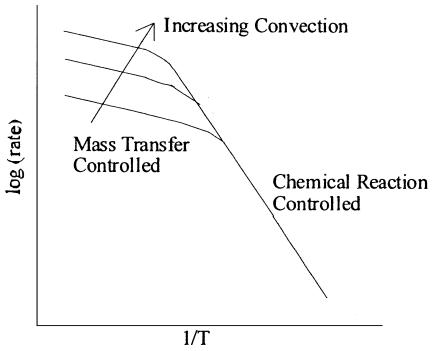


FIGURE 8 Chemical vapor deposition rate as a function of temperature.

having a much smaller slope (smaller activation energy) than for the reaction-controlled portion, and with the rate increasing with increasing convection in the gas.

As indicated above, a plasma discharge may be used to accelerate chemical vapor deposition. The primary difference between etching and PECVD is the composition of the gas and the products it forms in the discharge. In etching, the gas species form volatile products with the wafer; in PECVD, the gas species deposit on the wafer. One of PECVD's characteristics is that it allows deposition to be carried out at lower temperatures. For example, it is possible to deposit an amorphous film from silane (SiH_4) on any substrate down even to room temperature. Unlike the film deposited at 800°C, the PECVD film deposited at room temperature contains roughly as many atoms of hydrogen as of silicon. That is, it is a silicon hydride, from which solar cells may be fabricated but not yet integrated circuits.

A high-intensity light may also be used to enhance CVD. This is called photon-assisted CVD or PACVD. The incident photons may act both to heat the substrate and to cause photochemical reactions. If a laser is used as a light source, the deposition may be localized to the neighborhood of the beam. The beam may be moved directly to write lines on a substrate. However, additional development is required to fabricate submicrometer structures; therefore, the primary application may be in printed circuit boards or hybrids, rather than for integrated circuits.

IX. EPITAXY

Epitaxy is a crystallographic term meaning that the crystal lattices of two substances in contact with one another have a preferred orientation to one another. Epitaxial growth may be achieved either by chemical vapor deposition (as described in Section VIII) or by techniques using molecular beams or growth layers from solution. Examples of epitaxial growth include the formation of a single crystal

silicon film with one doping level on a single crystal silicon substrate with other doping, and single crystal silicon deposition on a sapphire substrate (SOS).

Two techniques have grown popular for III-V film deposition, and may yet prove useful in silicon devices. These are molecular beam epitaxial growth (MBE) and liquid phase epitaxial growth (LPE). In MBE the wafer is placed in a very high vacuum, such that the mean free path of gaseous molecules is much longer than the size of the chamber. A source of one type of desired atom is placed in a chamber with a small hole aimed at the wafer, producing an atomic beam. The use of several such sources can be used to produce doped layers, compounds, and very rapid changes in composition of the grown layers (e.g., superlattices).

Liquid phase epitaxy is really growth of layers from a solution. As practiced with the III-V compounds, a wafer is placed on a graphite slider which is moved to place the wafer in contact with a solution contained in a graphite cavity. Lowering the temperature below the equilibrium value (the solubility) causes growth to occur. Layers of different composition can be grown on top of one another by moving the wafer from one solution composition to another.

Silicon on insulator is being examined as a means of avoiding parasitic capacitance and latch-up. Some SOI structures may also permit three-dimensional integrated circuits to be fabricated. One form of SOI in current use is SOS. An epitaxial silicon film is grown by CVD on a single crystal sapphire (Al_2O_3) wafer. Each device is formed in a silicon island produced by etching the silicon around it down to the sapphire substrate. Sapphire is an electrical insulator but a good thermal conductor.

Current SOI effort is concentrated on growth of single or large-grained silicon on top of a silicon wafer. Many techniques for accomplishing this have been investigated, and some are in production. One method is to cut periodic holes in the insulator and do CVD under conditions such that nucleation of silicon does not occur on the bare insulator. Silicon grows epitaxially from the substrate up through the holes and then spreads laterally over the surface of the insulator.

Another method for SOI is to use CVD to deposit polycrystalline silicon on the insulator, melt the poly layer, and cause it to refreeze. Surface tension tends to cause the molten silicon to ball up, and so the poly layer is usually covered by a thin layer of oxide or nitride, completing a four-layer sandwich as shown in Fig. 9. One technique used is first to cut holes in the bottom oxide so the polysilicon is in contact with the substrate. Then when the poly layer is melted and refrozen, growth can take place from the substrate up through the holes to produce a single crystal layer. Alternatively, a laser can be used to melt a narrow zone in the poly and move it slowly along the surface.

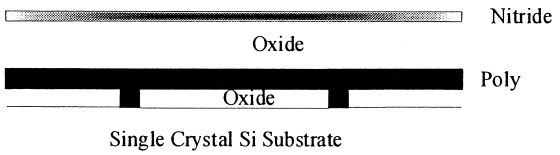


FIGURE 9 One method of producing silicon on insulator (SOI). The poly layer is heated by melting from above, then refrozen.

The two most popular SOI techniques are ion implanting and wafer bonding. In the former, known as SIMOX (separation by implantation of oxygen), is discussed in Section VII. Wafer bonding, or anodic bonding, involves the direct bonding of silicon to silicon dioxide in the presence of heat and an electric field. Two wafers, one with an oxide layer, are bonded together. The assembly is then thinned by mechanical lapping.

X. METALLIZATION AND CONTACTS

Conductors must be deposited and patterned on an integrated circuit to act as gate electrodes in FETs, connect devices together on a chip, and allow the chip to be connected to the outside world. In modern ICs, several layers of metals are used to allow for increased density of components on the chip, as well as to decrease the signal path, thus increasing the device speed. A variety of pure metals (Al, Cu, Au, Ti, W, Pb, Sn), alloys (Al-Cu, TiW), and compounds (metal silicides and nitrides) are used as conductors in integrated circuits. Ideally, they should satisfy a number of requirements, many of which are conflicting. They should be highly conductive, easily deposited, corrosion resistant, readily etched, stable, capable of withstanding high temperature, and adherent to the substrate.

The most common uses for metals in ICs are for contact materials and interconnects. Commonly used contact materials include silicides, Ti, Cr, and Ta. A contact material should be capable of reducing native oxides on the substrate in order to insure reproducible contact. Contact layers are usually the first layer in a multilayer metallization and therefore must exhibit excellent adhesion to the substrate. The most common interconnect metal currently in use for ICs is aluminum or mixtures of aluminum with silicon and/or copper. Copper is often the interconnect of choice for many packaging applications and is becoming increasingly common for interconnect levels as well. Polycrystalline silicon alone or with a metal silicide is often used as a gate electrode because, unlike aluminum, it can be heated to high temperature with no adverse effects

on the silicon substrate. Tungsten metallizations are also used in applications where refractory properties are required or in situations where step coverage or hole filling is needed. These materials are not without their problems, however, and in many applications are being replaced by refractory metals and compounds.

Two types of metal-semiconductor contacts are possible, rectifying and ohmic. The rectifying contact, also called the Schottky diode, can be incorporated in the device design as an active circuit element. The metal acts as an infinite source of holes and electrons. Consequently, when the metal and semiconductor are brought together, a *p-n* junction forms at the interface. As with any diode, current flows in one direction only. Rectifying behavior can be eliminated by heavily doping the semiconductor in the contact region before metallization. The result is called an ohmic contact, which allows current to flow easily in both directions.

Metals are deposited by several methods, the primary ones being CVD and physical vapor deposition (PVD) (i.e., evaporation and sputtering). Film growth by CVD is discussed in Section VIII. Tungsten, deposited using a mixture of WF_6 and H_2, is one example. Physical vapor deposition is perhaps the most popular group of deposition methods. In evaporation, a source of the metal to be deposited is heated at low pressure, 10^{-6} torr or less. Evaporation sources may be heated by resistance, rf induction, or electron beam methods. A low pressure is required to reduce surface contamination as well as to allow the metal to evaporate. Even at a millionth of a torr, ambient gas species impinge on the substrate at a rate of about 1 monolayer/sec. This can be a significant source of contamination in some applications. As in CVD, the morphology and properties of the deposited film depend on the substrate temperature and the rate at which the evaporated atoms strike the surface.

Sputter deposition is performed in a plasma (glow discharge), usually at a higher pressure than in plasma etching, about 0.001–0.1 torr. In most cases, the gas is one of the noble gases, argon being the most popular. The material to be deposited, called the target, is placed on the cathode. The plasma-generated positive ions strike the surface with sufficient energy to remove, or sputter, target atoms. These atoms deposit all over the inside of the chamber, including the substrate which usually forms the anode. Cathode bias, target temperature, gas species, and pressure control the sputter rate. Substrate bias and temperature control the film morphology and deposition rate. Sputtering may also be performed in a reactive environment such as a mixture of argon and nitrogen. For example, sputtering titanium with nitrogen gas leads to a titanium nitride film.

For interconnect applications where metal fill of a high aspect ratio contact hole is necessary, a technique called collimation can be employed. By using a honeycombed, solid angle restricting structure between the target and the substrate, the flux of metal atoms impinging on the wafer is restricted to those primarily in the normal direction. These directional species have better fill capability than their uncollimated counterparts. A similar effect can be obtained by increasing the throw distance between the target and the substrate without the collimator. Both collimation and increased throw distance lead to a decrease in the deposition rate on the wafer and an increase in the amount deposited on the collimator or the chamber walls. In both cases, frequent cleaning of the tool is required. Further gains in hole-filling ability can be obtained by a form of PVD called ionized PVD. Ionizing the metal atoms and applying an rf bias to the substrate results in a more directional fill.

In recent years, copper metallization has emerged as a possible successor to aluminum. Due to its higher conductivity, copper offers the potential of increased device speed and smaller wires compared with aluminum. Copper may be deposited by electroplating, the same technique used in packaging applications.

Various problems have been observed with the multilevel metallization used in ICs, including step coverage, hillocks, silicon solubility, and atomic migration due to mechanical or electrical stress. Most metallization problems have been addressed by materials or structural changes (or both). Poor step coverage is often observed with PVD films. Step coverage problems have been addressed by minimizing topography and using processes such as CVD that provide improved step coverage. Hillocks are protrusions from the metallization that are caused by differences in thermal expansion coefficient between the metal film and the Si substrate. The addition of alloying elements such as Cu and multilayer metallizations have lessened the susceptibility to hillock formation. The origin of the silicon solubility problem is that when heated, the silicon in the contact regions dissolves in the aluminum and diffuses along the metal line. Upon cooling, the silicon precipitates. Etch pits are formed in the underlying silicon and become filled with aluminum, causing a metal to substrate short circuit. One way to minimize this problem is to use a silicon-aluminum mixture for metallization. However, since the solubility increases with increasing temperature, there is always some dissolution and precipitation of the silicon when the temperature varies. There is a practical upper limit of about 500°C on wafers when aluminum is present. A second and common method of reducing the solubility is to introduce one or more "barrier layers" between the silicon and the aluminum. Both

silicides and titanium-tungsten mixtures have been used for this purpose.

Electromigration defects are observed during reliability stressing. As current flows through the thin aluminum lines, it causes atom migration along the grain boundaries. Small variations in the line thickness are amplified, resulting in voids or holes in the metal lines. The atoms which leave the void areas pile up to form hillocks of metal. Eventually the consequence is an open or a short circuit and device failure. One method of minimizing electromigration is to incorporate small amounts of copper into the aluminum. The copper apparently "plugs up" the grain boundaries, reducing the migration. Stress migration is similar to electromigration, with the exception that the driving force is mechanical stress, not current. Alloying with copper also reduces the susceptibility to stress migration.

When multiple layers of metals are employed, the different levels are connected by holes (vias) formed through interlevel dielectric materials. Tungsten is often used for these interconnects. These insulation regions can be inorganic materials like silicon dioxide or silicon nitride, an organic material, or combinations of these. The most popular class of organic insulators is polyimide. Several different kinds of polyimides are available, but all are characterized by being insoluble, stable to relatively high temperatures (about 400°C) and impervious to water vapor. In most cases, the polymer is dispensed as a solution of polyamic acid, spin-coated on the wafer, and cured to allow the imidization reaction to proceed. In addition to its desirable physical properties, the spin-coated polyimide is more planar than typical inorganic materials.

XI. PLANARIZATION

Throughout the fabrication process, films are continually deposited and etched. Consequently, the surface of the silicon wafer develops a significant amount of topography, with comparatively large hills and deep valleys. Two problems result. First, it may be difficult to achieve a uniform focus with the lithographic tool over its entire field of view, since the typical depth of focus for modern instruments can be 0.5 μm or less. Second, it is impossible to achieve more than two or three layers of metal interconnects. This leads to longer signal paths and an underutilization of the silicon surface.

For these reasons, there is considerable interest in being able to planarize wafers, and several techniques have been developed. Early work focused on the use of a planarizing layer, such as resist, polyimide or "spin-on glass," a silicon-based polymer which is deposited as a liquid and then cured at high temperature to form a solid. While the addition of a liquid coating does make the surface more planar, it is usually far from flat. As illustrated in Fig. 10, there can be significant differences in the "degree" of planarity. On one extreme is a conformal film, where the film surface reflects the details of the substrate. Liquid coatings usually result in partial or local planarization depending on the size and separation of the underlying features. Reflow glasses demonstrate this type of behavior. For a liquid on a substrate with a regular topography, the time required for planarization is proportional to the fluid viscosity and the fourth power of the distance between surface features, and is inversely proportional to the surface tension and the fourth power of the film thickness. Any liquid film will planarize eventually, but the time required may be prohibitively long, and liquid films are of no value. If the film is hardened by solvent evaporation or chemical reaction, the film will shrink, and the substrate topography will reappear.

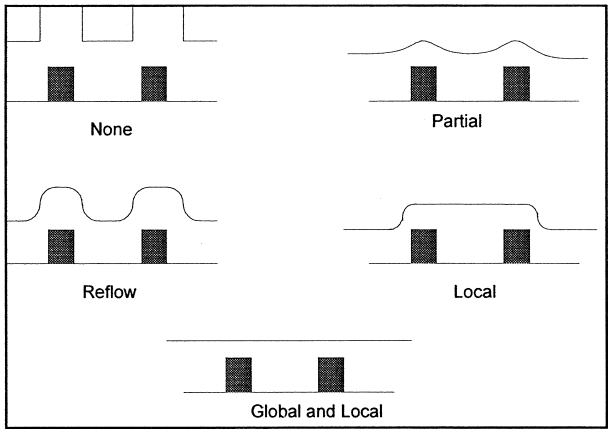


FIGURE 10 Differences in planarization.

The planarization method which has received the most interest in recent years is chemical mechanical polishing or chemical mechanical planarization (CMP). This is similar to the lapping mentioned in Section II to achieve flat, polished single crystal wafers. It is based on the much older technology of glass polishing. The wafers are held by a carrier, in most cases above a polish pad. The pad is usually polyurethane or a polyurethane-impregnated felt or foam. The wafer is moved along the pad in the presence of an aqueous slurry. The liquid reacts with the wafer surface, and the particles in the slurry remove the reaction products by mechanical action. The relative importance of reaction and mechanical abrasion differs for different materials. In all cases, the mechanistic details are very complex and not well understood.

A wide variety of process variables are associated with the polisher, the wafer, the polish pad, and the slurry. In general, the removal rate is proportional to the pressure

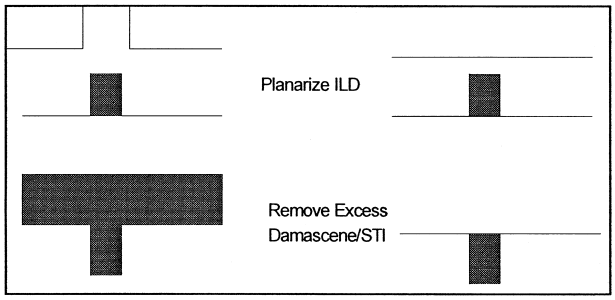


FIGURE 11 Applications of CMP.

which forces the wafer against the pad, and the relative velocity between the wafer and the pad. This simple relationship is called Preston's equation. The proportionality constant depends on (among other variables) the film type and how it was deposited, the pattern density, the pad material and its stiffness, the composition of the slurry, and the nature, size, and concentration of the solids.

Chemical mechanical planarization enables a number of new technologies. As illustrated in Fig. 11, an interlevel dielectric (ILD) can be deposited and then planarized to improve the efficiency of the next metal layer. It is not uncommon to have six or more levels of metal on a chip. In another application, a trench can be cut into the substrate and then a layer of silicon dioxide deposited. The excess oxide is removed from the surface leaving an insulating region between two active areas. This shallow trench isolation (STI) process is used to prevent "latch-up" of transistors through the substrate. Chemical mechanical planarization is also used in Damascene metallization. Instead of depositing a metal and then patterning it, the metal is applied to a patterned substrate and the excess is removed. Both the via and the metal line may be deposited at the same time. This is particularly useful when employing metals which are difficult to etch, such as copper.

There are a number of problems associated with polishing. The polish rate varies across the wafer surface. This within wafer nonuniformity (WIWNU) means that some material must always be "overpolished." Consequently, there must be a large difference in the polish rate between the film and the underlying "polish stop." In essentially all polishers, there is a region close to the wafer edge from which material is removed at a significantly lower rate than in the central portion. This edge exclusion reduces the useful wafer area. Finally, the polish rate may change during processing as the pad undergoes a break-in and later an aging process. For some pads, the polish rate increases during the break-in period. Later, as the pad ages, the rate decreases and the nonuniformity increases. Most polishers provide for pad conditioning which continually refreshes the pad surface.

A number of defects in the polished wafer often occur. Especially when polishing soft materials such as copper, the particles in the slurry can scratch the film. Soft polish pads are good to achieve global planarity since they can conform to the surface, but soft pads can also lead to "dishing" or removing material between features. Residual slurry and other types of debris are left on the surface after polishing. As a result, wafers must be scrupulously cleaned after polishing and before they are sent to the next fabrication step. As should be expected, CMP is inherently a very dirty process. To reduce contamination, the CMP area is often kept separate from the other process areas, and personnel, as well as wafer boats and boxes, are not shared between the two.

XII. PACKAGING

After chip fabrication is complete, the chip must be packaged. Packaging serves to protect the chip from the outside world and allows electrical connection to it. The packaging sequence includes testing and dicing (wafer is cut into chips); electrically "good" chips are then electrically connected to the package and encapsulated.

First, the backside of the chip is electrically connected to the packaging substrate either with a conductive polymeric material or by a metal bond. The connections to the device side of the chip are usually made in one of three ways. Wire bonding is the most commonly used technique. A gold or aluminum wire is thermocompression or thermosonically bonded to the last level of metal on the chip. The other end is attached to the electrical circuits on the package substrate. Wire bonding is a low cost technique, but one drawback is that the density of chip pads that can be bonded is limited.

A second interconnection method is known as tape-automated bonding. A gold or copper bump is formed on the chip and then plated copper beams are thermocompression bonded to all the bumps on the chip simultaneously. The copper beams are supported on tape that is fed from a reel. After the bond to the bumped chip is made, the tape plus chip combination is collected on a second reel, where it will wait until ready for bonding to the package substrate and separation from the tape frame. Tape-automated bonding offers a higher pad density than wire bonding but is somewhat more costly.

The third bonding option is the "flip-chip" technology. Solder balls are formed on the chip bonding sites. The chip is turned face down so that the solder balls are in direct contact with the metal pads on the package substrate. The solder is then heated in order to bond with the substrate pads. The "flip-chip" technology offers a high pad density, and pad designs are not limited to the chip periphery, as in wire bonding and tape-automated bonding.

After the interconnections are completed, the packages are encapsulated to protect them from the environment. The encapsulation technology depends upon the application. Chips may be hermitically sealed for high-reliability applications, whereas a plastic encapsulant may be used for low-cost applications. Many different package shapes are also available. Issues such as cost, performance, reliability, size, and ability to dissipate heat help in determining package type.

XIII. RELIABILITY

In order to insure that reliable devices have been produced by the IC manufacturing process, a sampling of chips is tested. Because the failure rate due to wear under normal use conditions is expected to be quite low, accelerated testing must be done. Failure characteristics may be studied in relatively short times by increasing temperature, voltage, current, and humidity above normal use conditions. Thermal cycling from a low temperature to a high temperature is also used in order to test the mechanical integrity of the chip and the package.

Reliability tests often combine several accelerating factors, with each test directed toward understanding a particular failure mode. Voltage-temperature stresses are used to investigate the effects of mobile ionic species in oxides. High-applied electric fields combined with temperature are used to study dielectric breakdown and charge injection. Voltage, temperature and humidity are used to accelerate and detect corrosion of chip and package metallizations. Corrosion in ICs is often caused by processing residues or contaminants such as chloride ions. High current densities combined with high temperatures are used to evaluate electromigration. High-temperature and high-intrinsic stresses are used as the accelerating factor for stress migration (Section X). Repeated thermal cycling simulates device turn on and also storage conditions. Wiring and dielectrics are stressed by differences in thermal expansion coefficients during thermal cycling. The data generated by such stresses are used to improve device processing and to ensure product reliability.

SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • CHEMICAL VAPOR DEPOSITION • CRYSTAL GROWTH • EPITAXIAL TECHNOLOGY FOR INTEGRATED CIRCUIT MANUFACTURING • FIELD-EFFECT TRANSISTORS • INTEGRATED CIRCUIT PACKAGING • MOLECULAR BEAM EPITAXY, SEMICONDUCTORS • SILICON, HYDROGENATED AMORPHOUS • SUPERLATTICES • ULTRA LARGE SCALE INTEGRATION (ULSI) DESIGN

BIBLIOGRAPHY

El-Kareh, B. (1995). "Fundamentals of Semiconductor Processing Technology," Kluwer Academic Publishers, Boston.

Ghandhi, S. K. (1983). "VLSI Fabrication Principles," Wiley-Interscience, New York.

Glendinning, W. B., and Helbert, J. N. (1991). "Handbook of VLSI Microlithography," Noyes Publications, New Jersey.

Kern, W. (1983). "Handbook of Semiconductor Cleaning Technology," Noyes Publications, New Jersey.

Lau, S. S. (1990). "Electronic Materials Science," Macmillan, New York.

Lee, H. H. (1990). "Fundamentals of Microelectronics Processing," McGraw-Hill, New York.

O'Mara, W. C., *et al.* (1990). "Handbook of Semiconductor Silicon Technology," Noyes Publications, New Jersey.

Ravi, K. V. (1981). "Imperfections and Impurities in Semiconductor Silicon," Wiley-Interscience, New York.

Rosnagel, S., *et al.* (1990). "Handbook of Plasma Processing Technology," Noyes Publications, New Jersey.

Sapoval, B., and Hermann, C. (1995). "Physics of Semiconductors," Springer-Verlag, New York.

Sheats, J. M., and Smith, B.W., eds. (1998). "Microlithography: Science and Technology," Marcel Dekker, New York.

Sze, S. M., ed. (1988). "VLSI Technology," McGraw-Hill, New York.

Thompson, L. F., *et al.* (1994). "Introduction to Microlithography," American Chemical Society, Washington, D.C.



Integrated Circuit Packaging

Frank E. Andros
Bahgat G. Sammakia
Integrated Electronics Engineering Center,
State University of New York at Binghamton

I. Definition and Classification of Electronic Packages
II. Bonding Technology
III. Design Considerations
IV. Emerging Technologies that Rely on Electronics Packaging

GLOSSARY

Ball grid array (BGA) A surface mount component which contains balls as the interconnect medium between the substrate and the printed circuit card or board.

Chip scale package (CSP) Surface mount BGA component whose package is less than 1.2 times larger than the chip and has a BGA pitch of 1.0 mm or less.

Column grid array (CGA) Surface mount component that utilizes an array of columns (typically solder) to interconnect the component to the printed circuit card or board. Most common on ceramic substrates where there is a large difference in the coefficient of thermal expansion between the substrate and card. Columns being more flexible than balls results in better reliability.

Component Electrically functional subassembly, which can be passive or active. Passive refers to resistors, capacitors, inductors, etc. Active refers to those containing semiconductor die.

Core Element of a printed circuit card or board consisting of a dielectric with conductors on one or both sides. Multiple cores are laminated together to form a multilayer card or board.

Die A semiconductor element (typically square or rectangular) with electrically functional integrated circuits on one side. Also referred to as a chip.

Micro-electro-mechanical-system (MEMS) Integrated micro systems that contain mechanical and electrical components fabricated on the surface of a semiconductor device.

Multichip module (MCM) A component containing two or more active devices bonded to the same substrate.

Pin-in-hole (PIH) Typically used in reference to components or connectors with pins that are to be soldered to printed circuit cards. The pins are inserted into metallized vias (through holes) in the card, and solder is applied to fill the space between the pin and via side walls forming a robust electrical and mechanical interconnection.

Printed circuit board (PCB) An organic board containing conductors (on its surface and/or internally)

that allow electrical communication between elements mounted on it.

Substrate A space transformer between the die and the printed circuit board to which the die is electrically and mechanically attached.

Surface mount technology (SMT) The technology used to attach electronic components or elements to metallized pads on the surface of printed circuit cards or boards. The interconnection may be solder (most common), adhesive, or a bonded interface.

Via Hole in a printed circuit card, board, or substrate that is metallized and serves to electrically connect a conductor on one layer to a conductor on a different layer. Three classifications: (1) thru, interconnect from one side to the other; (2) buried, from one plane to an adjacent plane; (3) blind, on the outside sufrace of the card or board to an internal plane.

I. DEFINITION AND CLASSIFICATION OF ELECTRONIC PACKAGES

Electronics packaging is commonly defined as "all parts of an electronic system with the exception of the integrated devices." The electronic system itself may be a personal computer, laptop, server, or a workstation. Systems also include many consumer electronics, appliances, automobiles, weapon systems, airplanes, medical diagnostic and surgical instruments, and biomedical devices. Typical hardware that would be considered a part of the electronic package includes printed circuit boards, chip carriers, connectors, cables, power supplies, fans and blowers, and the system box or chassis among many others. The above definition of packaging includes not only the hardware used in the electronic system itself, but also includes the manufacturing and assembly processes for the components and the system. This very broad definition of packaging therefore encompasses a wide range of technical and scientific fields and is intrinsically multidisciplinary. Just about every engineering discipline and many scientific, financial, and marketing disciplines are involved. All of these areas need to be considered during the development qualification and manufacturing of any electronic product in order to ensure acceptable performance, reliability, and financial/marketing success. This is shown schematically in Fig. 1. It is common to classify electronic packages according to different levels. These include zeroth level, first level, second level, third level, and system level as shown schematically in Fig. 2.

A. Zeroth Level

The "die" or "chip" is often referred to as the "zeroth" level of packaging and in most cases is not considered part of packaging. Die are semiconductor devices, most commonly silicon (Si), gallium arsenide (GaAs), silicon germainium (SiGe), and a host of other lesser known materials, that contain electrically functional integrated circuits (ICs) capable of performing specific tasks such as storing information, decision making, etc. They are typically square or rectangular in shape and are formed by sawing wafers (thin slices sawed from an ingot of the parent semiconductor material), Fig. 3.

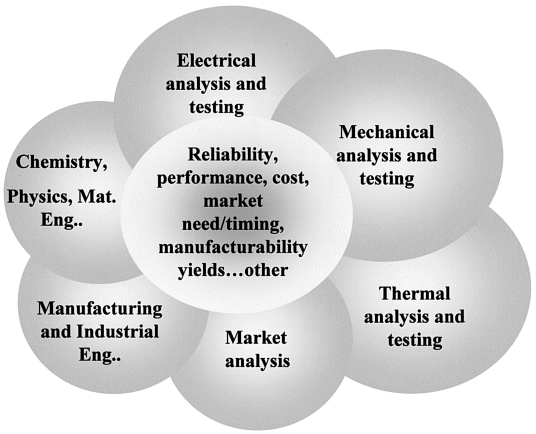


FIGURE 1 Key areas of research and development involved in electronics packaging.

Wafers, and therefore die, vary in thickness depending on the wafer diameter. Larger wafers are thicker to minimize fragility during handling and processing. Thinner wafers may be produced by backside grinding after the parent wafer has been circuitized (doping and metallization). Thin die are desirable for those technology applications that require lightweight, thin profiles. The majority of the die thickness serves as an electrically nonfunctional substrate to facilitate handling and processing the ICs. The electrically functional portion of the die is contained in a very thin layer on the die surface and may contain up to four layers of circuitization. Some typical die properties important in packaging are given in Table I.

B. First Level

The first level package is also commonly referred to as the module. The first level package or module includes the chip, the chip carrier, the encapsulation system that covers the chip, and the module cap, cover, or heat sink as the case may be. The primary function of the first level package is to electrically and mechanically connect the die to a substrate and to provide protection for the die and its interconnects.

The primary and most common functions of a first level package are

- Mechanical support and protection for the chips
- Thermal management or power dissipation

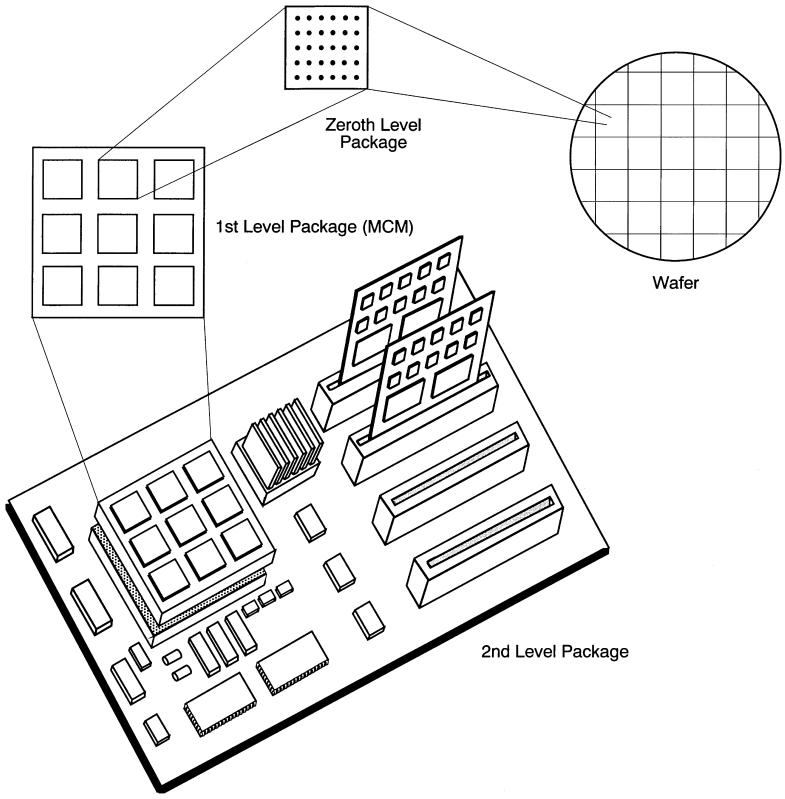


FIGURE 2 Schematic diagram depicting the nomenclature used to define various levels of packaging.

- Signal distribution from the chips to the printed circuit board
- Power distribution to the chips
- Chemical and environmental protection of the chips

There are many different types of first level packages. They are classified in different ways. Some common classifications are

- Chip to chip carrier interconnect technology (flip chip, wire bond, TAB, etc.)
- Chip carrier to PCB interconnect technology [pin-in-hole, ball grid array (BGA), J-lead, etc.]
- Chip carrier material (ceramic, organic)
- Molding compound (plastic)
- Relative size of the chip to the package [TSOP, chip scale package (CSP)]

It should also be noted that sometimes the first level packages are classified by combining several of the above classifications, for example, *organic flip chip*, or *wire bond ceramic BGA*. Figure 4 shows examples of several first level packages: pin-in-hole packages and surface mount packages. Figure 5 shows schematic cross sections of several

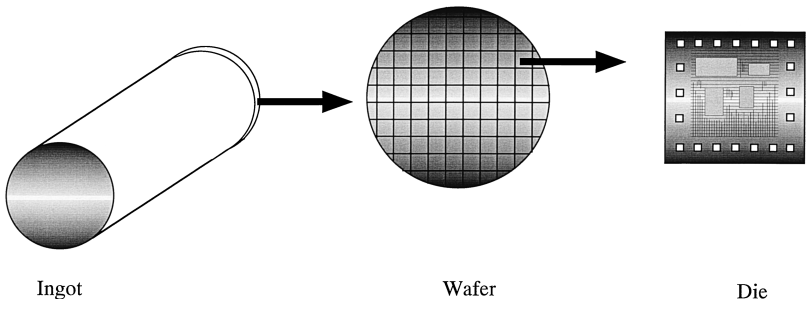


FIGURE 3 Schematic diagram showing the major die fabrication steps.

TABLE I Physical and Electrical Properties of Some Common Packaging Materials[a]

	Physical characteristics				Electrical characteristics		
Material	CTE ppm/°C	Thermal conductivity, W/m · K	Ultimate tensile strength, MPa	Young's modulus, GPa	Resistivity Ω cm	Dielectric constantat 1 MHz, $\times 10^3$	Dissipation factor at 1 MHz %
Dielectrics							
AIN	2.89–3.79	319–195	—	345	10^{12}	8.9–10	0.08
Alumina (96%)	5.4–7.2	46–31	179	303	$>10^{14}$	9.34	0.2
BeO	5.9–7.2	275–184	138	312	$>10^{14}$	7.35	—
Cordierite	7.9	4	245	—	$>10^{13}$	5	0.5
Mullite	5	6.7	—	—	$>10^{14}$	5.4	—
Polyimide	20	0.155	103	1.37	10^{16}	3.4	0.2
PTFE	75	0.26	24.5	0.4	—	2.1	—
SiO_2 (CVD)	0.57	0.1	98	44	$>10^{15}$	4.42	—
Air		0.03				1.0	0
Conductors							
Ag	19.7	418	145	75.8	1.59×10^{-6}		
Al	23.1	218	146.5	70.2–66.6	2.65×10^{-6}		
Au	14.2	319	340	57	2.35×10^{-6}		
Cu	16.7	378	310	116	1.67×10^{-6}		
Mo	5.0	139	690	325	5.2×10^{-6}		
Ni	13.1	92	820–709	225–217	6.84×10^{-6}		
Pd	9.0	82.5	152	145	10.8×10^{-6}		
Sn/Pb (63/37)	24.1	51	35–42	14.9			
Sn/Pb (10/90)		36					
W	4.5	178–159	621	388	5.64×10^{-6}		
Substrates							
Alloy 42	5.3	15.6	588	145	57×10^{-6}	—	—
FR4	13.6–15.4	0.29	276	71	—	5.2	—
Si(110)	2.6–3.25	148–98.9	185	169	4×10^{-12}	12.1	—
GaAs	5.9						
SiC	3.4	120	—	439	4×10^{-13}	9.72	5
SiN	1.1–2.17	40.2	106	150	—	4.2	—
Cu/Invar/Cu	17.7	394	586	160	—	—	—
Kovar	6.11–4.98	16.3–17.9	520	138	49×10^{-6}	—	—
Epoxies, encapsulants							
Ag filled Epoxy	$<x_1 - 46$	1.3–50			1×10^{-4}		
Molding Compound	$<x_2 - 240$ $<x_1 - 12$–18 $<x_2 - 41$–65	0.66	20.0	$e_1 - 11.7$ $e_2 - 0.1$	5×10^{12}	5.0	
BT Epoxy	14	0.19					
FR-4	16	0.11	310–379	17.2	1×10^{10}	4.5–5.2	0.01–0.025
Epoxy Encapsulant	20–65	0.17–0.21			10^{12}–10^{13}	3.5–4.8	0.01–0.03
Silicone	60–90	0.15–0.30				2.6	0.1–0.2

[a] (Adapted from Moore, T. M., and McKenna, R. G.)

different packages including cavity up, cavity down, flip chip, and wire bond as well as hermetic and nonhermetic configurations.

C. Second Level

A second level package is typically a printed circuit board with several components on it as shown in Fig. 6. Printed circuit boards are composed of epoxy resin and fiberglass layers metallized on both sides. This structure is referred to as a core. Using photo processes circuits are formed from the metal layers. The circuit board is fabricated by laminating several cores with epoxy resin layers in between the cores. Holes are then drilled and metalized to act as interconnections between the different planes. Other types of interconnections between planes include

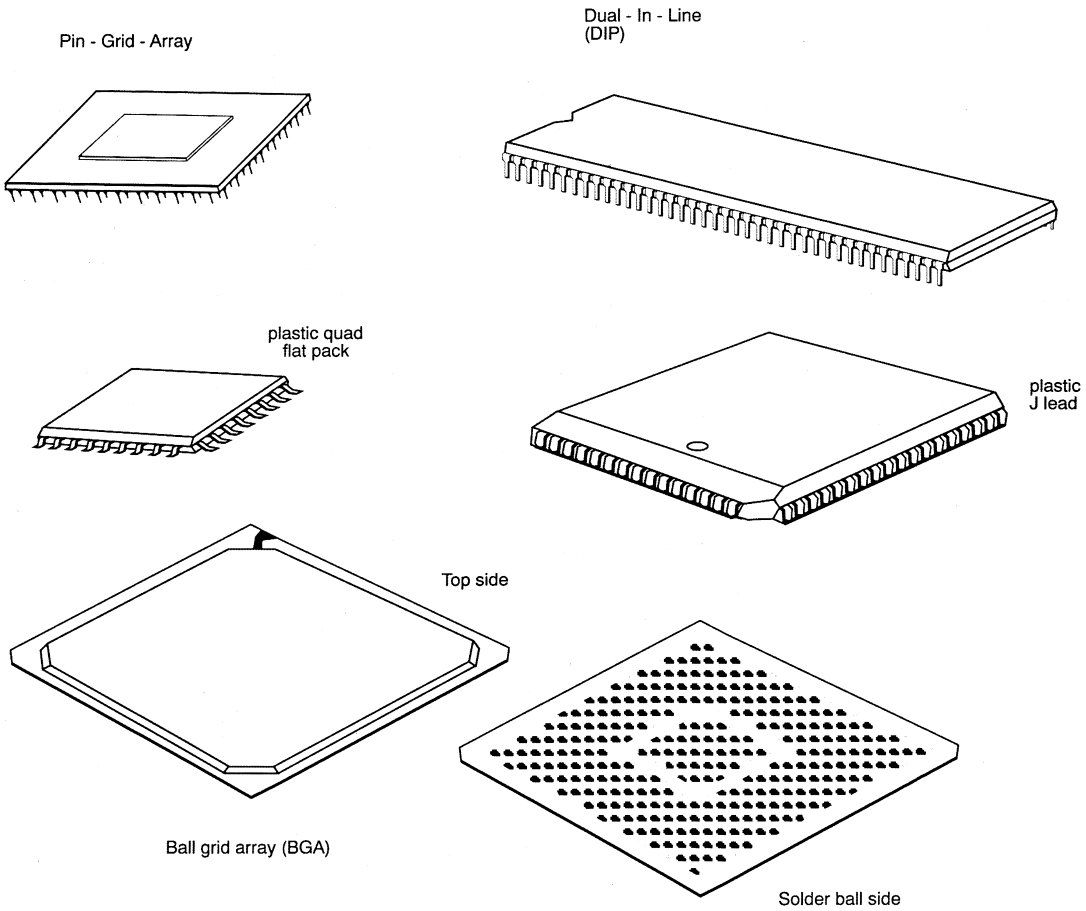


FIGURE 4 Examples of various first level packages: pin-in-hole components (PGA and DIP) and surface mount components (PQFP, J Lead and BGA). (after Harper)

blind and buried plated through holes (PTHs). The reason for the classification of packages into first, second, and third levels is that the considerations and concerns that arise in each level of packaging, and often the analytical and experimental techniques used in addressing them are unique. Three key areas where the classification is important are thermal management, mechanical design, and reliability. These topics are discussed in later sections after the reader is introduced to more details about interconnect technologies and systems level packaging.

II. BONDING TECHNOLOGY

There are three basic bonding technologies used today to electrically interconnect the die and first level package as illustrated in Fig. 7: (1) wirebond; (2) flip chip; and (3) tape-automated bonding (TAB).

A. Wire Bonding

Wire bond is the oldest and by far the most commonly used technology accounting for more than 90% of all die interconnects. This technique has been in production since the early 1960s. Wire bonding is used with ceramic, laminate, lead frame, and tape (flex) substrates; it has well-established reliability and a large supporting infrastructure (bonding and handling equipment, materials suppliers, established processes and controls, specifications, etc.).

After the die has been mechanically affixed to a substrate or lead frame, wires are bonded, one at a time, between the aluminum I/O pads on the die and substrate pads which may be plated with nickel, gold, palladium, or other inert metals. Bonding to oxidized surfaces (copper, commonly used for substrate metallization and bond pads, readily oxidizes) can result in unreliable interconnections, hence the need for a nonoxidizing metal plating on the pads. Lead frame materials (used in many plastic molded components) of choice are plated Alloy 42 and copper.

The most common wire is gold or aluminum, although copper and silver have also been used. Wire bonds to die pads can be ball or wedge bonds and wedge bonds are used exclusively to join the wire to the substrate pads. Ball bond diameters can vary from 2.5 to 4 wire diameters

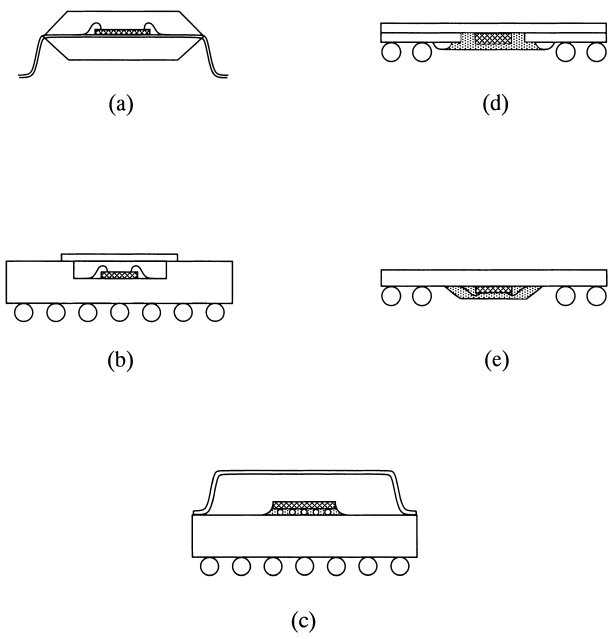


FIGURE 5 Examples of cavity up (a, b, c) and cavity down (d, e) packages: (a) molded plastic quad flat pac (b) hermetic wirebond multi-layer ceramic BGA (c) multi-layer ceramic BGA with flip chip and underfill (d) wirebond laminate BGA, epoxy dispensed (e) glob top wirebond laminate BGA.

depending on material, bond pressure, and ultrasonic energy. Wedge bonds are less, typically 1.5–2 wire diameters. Wedge-wedge bonds can provide a lower profile (less loop height) than ball-wedge bonds and therefore may result in a thinner, lighter finished component. These attributes are important for hand-held, portable electronics applications. Figure 8 shows the process sequence for ball-wedge bonding. Multiple methods may be used to form a reliable metallurgical joint between the wire and pad: (1) thermocompression (uses pressure, temperature, and time); (2) ultrasonic (uses ultrasonic energy, pressure, and time); and (3) thermosonic (uses ultrasonic energy, pressure, temperature, and time). The latter is the most pop-

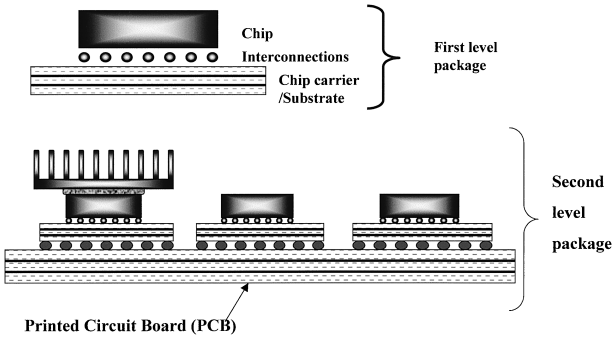


FIGURE 6 Schematic diagram of first and second level packages.

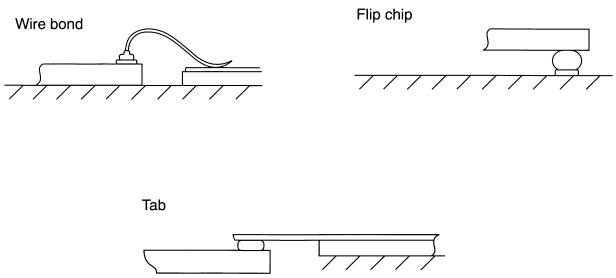


FIGURE 7 Die interconnect options for IC packages: (a) wire-bond, (b) flip chip, (c) tape automated bond.

ular bonding methodology. A wire-bonded die is shown schematically in Fig. 9.

Wire bonding is a peripheral interconnect technology, i.e., all bonds are relegated to the die edge. Die pads are on the order of 100 μm square when 25-μm wire is used and pad pitch is typically 150 μm. In some instances a double-staggered peripheral pattern is used to increase the I/O density. The substrates are designed such that one wire does not cross another to eliminate the potential for electrical shorts. Wire lengths and distance from die pad to substrate pad are limited to minimize wire "droop" or wire "wash" during subsequent component processing. This is particularly critical for injection molded components to minimize wire-to-wire electrical shorting. A common rule of thumb is that chip pad to substrate pad spacing should be less than 100 wire diameters.

B. Tape-Automated Bonding (TAB)

Tape-automated bonding is a technology specific to flexible circuit substrates. Die bonds to the flex are called inner lead bonds (ILB) and flex bonds to the circuit card are called outer lead bonds (OLB). Figure 10 shows a schematic representation of a TAB chip carrier. The flex circuit (tape) is supplied on reels, similar to movie film with sprocket holes on both edges for automated feeding during the bonding operation. The flex has a window or opening in the substrate at the die bond site where copper conductors plated with nickel and gold protrude in cantilever fashion. Typical conductor thickness is 25 μm with minimum width and pitch at 25 and 75 μm, respectively.

Inner lead bonding can be completed with bumped die or with bumped leads. Die bumping is carried out by electroplating gold, approximately 25-μm thick, on the die pads. Leads can be bumped by several methods: transferring premanufactured bumps to the leads with a bonding operation, melting the lead tips with a laser, and by including an additional photo/plating step during tape manufacture. Thermocompression bonding using pressure, temperature, and time is the most common bonding technology. A die is picked and placed on a heated

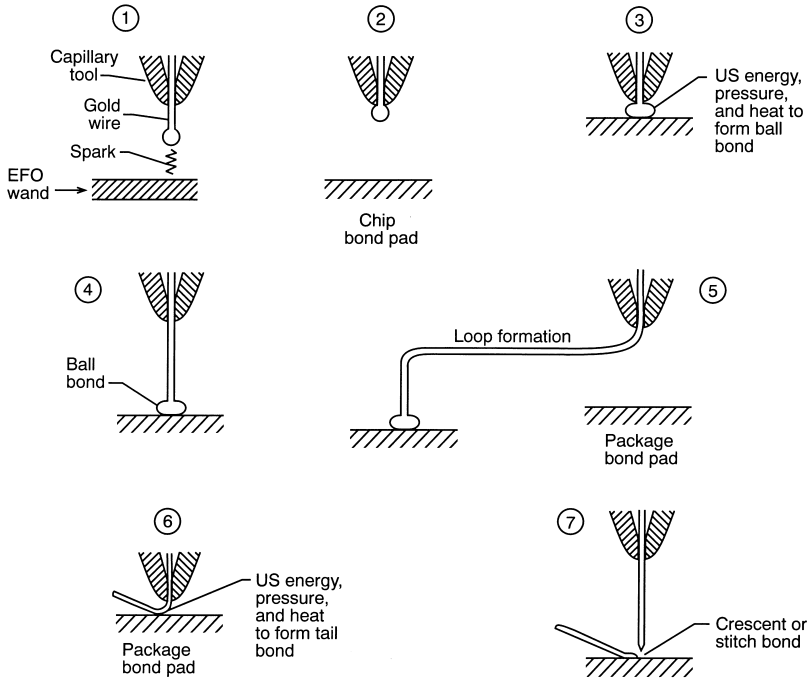


FIGURE 8 Process sequence for wirebonding a die to a chip carrier, showing ball and wedge bonding (after Harper, 1997).

lower thermode, the flex is fed to the bond station from a continuous reel and optically aligned with the die bumps. An upper thermode (550°C typically) applies pressure to the flex leads forming gold-gold diffusion bonds. Gang bonding (making all interconnections at once) minimizes process time. Single point bonding is used for prototype hardware due to the lengthy delivery time for thermodes.

C. Flip Chip Bonding

When a die is joined to a substrate with its functional surface facing the substrate, this is referred to as flip chip bonding. With flip chip the die pads are bumped with a conductive material that will serve as the interconnect providing both the electrical and mechanical connection between die and substrate (Fig. 7). This technology offers several advantages over other bonding technologies. It can be an area array interconnect as opposed to peripheral die interconnect for wire bond and TAB. This allows higher I/O capability on the die. It also allows the capability of bringing power to the die at any point on the surface as opposed to busing it from the periphery to the center of the die, and it offers the shortest connection from die to substrate. These factors enhance the electrical performance of the component. In addition the bonding process exerts little or no pressure on the die, therefore, bonding over active circuitry can be utilized with flip chip. Figure 11 shows a cross section of a solder flip chip on a laminate substrate.

The interconnection materials in use today are solder or conductive adhesives. Typical joint lengths are 50–125 μm with pad-to-pad pitches of 200–350 μm. The solder interconnect is sometimes referred to as C4 (controlled collapse chip connect) technology, a name coined by IBM, who was a key developer and user of flip chip for several years.

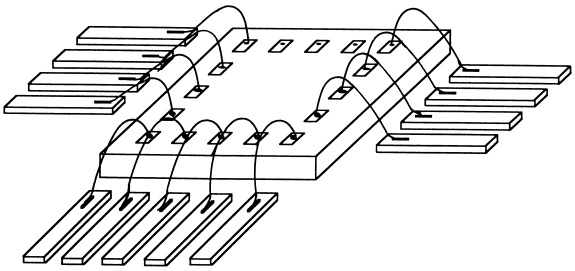


FIGURE 9 Schematic diagram showing a wirebonded die.

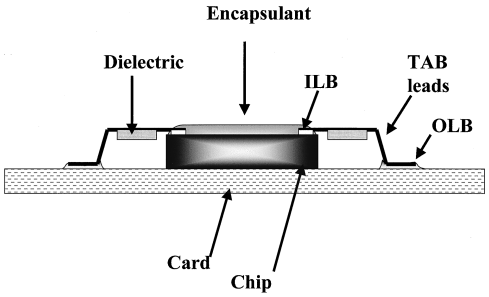


FIGURE 10 TAB chip carrier soldered to a circuit card.

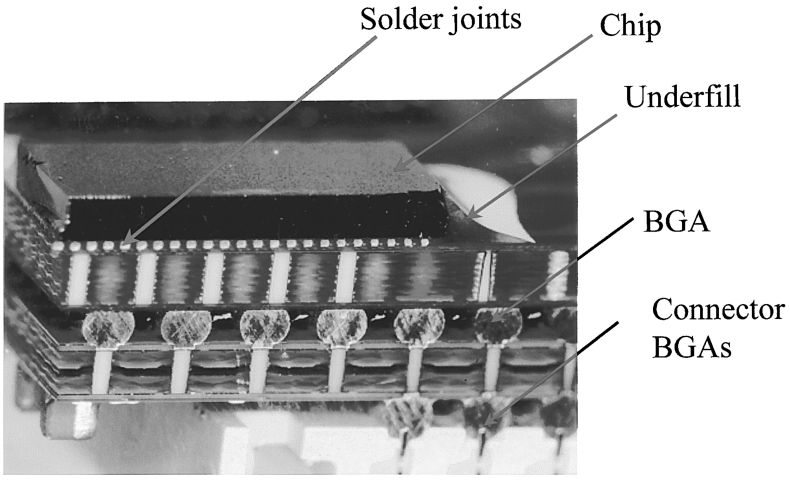


FIGURE 11 A cross-section of a flip chip package.

High-melt solders (3Sn/97Pb and 5Sn/95Pb) have been used on ceramic substrates, while low-melt or eutectic solder (63Sn/37Pb) has been used on laminate chip carriers. High-melt solders were the materials of choice for ceramic chip carriers before organic chip carriers came into being. With ceramic substrates a hierarchy of solders was used so that the chip interconnects (high-melt) did not reflow during subsequent card assembly solder operations which use eutectic solder. Application of flip chip solder interconnect to laminate chip carriers initiated the use of low-melt because the laminate substrates could not withstand the high temperature required for high-melt solders.

Die are bumped prior to wafer dicing. Aluminum pads are metallized to provide a solderable surface, a typical structure is shown in Fig. 12. Solder is either evaporated through a mask or electroplated on the copper die pads and then reflowed to form solder balls on the pads. Attachment of die to substrates is an automated process which includes fluxing, alignment of die solder balls to substrate pads, die placement, and reflow. Placement accuracy is not critical for solder reflow flip chip, 50% ball-pad misalignment can be tolerated because the surface tension in the molten solder will exert sufficient force on the die to align the die and substrate pads.

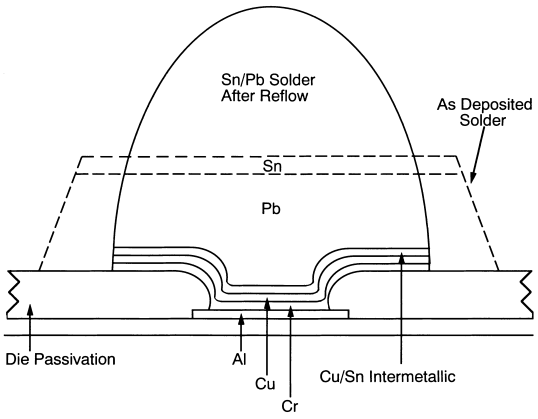


FIGURE 12 A schematic diagram showing one of several possible flip chip solder joint metallurgies (1999 Prismark Annual Report).

D. Conductive Adhesives

Conductive adhesives are the other interconnect medium in use today. Isotropic adhesives (electrically conductive in all directions), e.g., silver-filled epoxy, have been used in conjunction with gold bumped die to interconnect die to substrates. Die pads are bumped by completing a ball bond on a wire-bond tool and breaking the free end of the wire off at the ball (Fig. 13 and 14). Balls are planarized, conductive epoxy applied to the balls or substrate, die are placed on the substrate, and the epoxy is oven cured (Fig. 15). Chip bumping can also be carried out by plating gold bumps on the die.

Anisotopic adhesives (materials that conduct in one direction) are commonly partially cured or B-stage epoxies in sheet format containing conductive particles. The film thickness is typically 50 μm and the particles are solid nickel or nickel/gold-plated polyimide spheres, 5–10 μm in diameter. Die and substrate pads are aligned with the film adhesive between, and sufficient pressure and temperature are applied through a thermode to compress the film to the point where the conductive particles are crushed between the die and substrate pads creating a conductive path. The elevated temperature completes the cure of the adhesive (Fig. 16). The density of conductive particles in the film is so low that there is no potential for

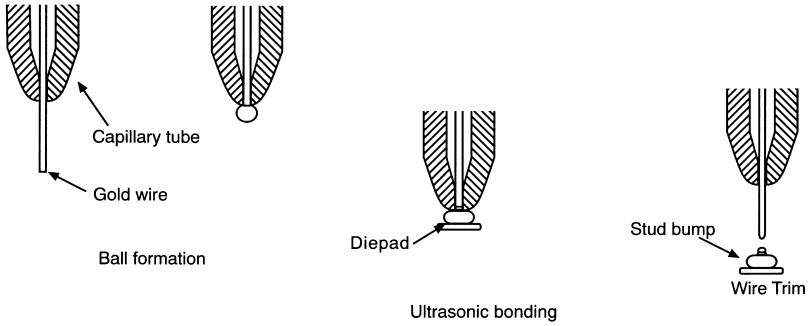


FIGURE 13 Process sequence for bump formation to be used for flip chip stud-bump interconnection.

shorting of adjacent interconnects. This bonding technology is commonly used today to bond driver display die and flex circuits to the indium tin oxide (ITO) metallization on glass for displays. A cross section of an anisotropic adhesive interconnect is shown in Fig. 17. In this example the spheres are 10-μm diameter polyimide, plated with nickel and gold. Work is currently under way to develop materials with adequate electrical resistance and reliability for chip carrier applications.

Industry use of flip chip joining technology is currently limited but it is increasing due to the advantages it offers over peripheral interconnect technologies. However, there are some disadvantages. Infrastructure does not exist to the same degree as with peripheral techniques. High-density interconnect at the die drives higher density circuitization into the substrate forcing the use of "build-up" technology which increases substrate cost. Reliability of the joints are an issue due to the mismatch in the coefficient of thermal expansion between the die (3 ppm/°C) and the laminate substrate (17 ppm/°C). This can result in fatigue of the joints during on-off cycling. Underfill materials, which fill the gap between the die and substrate, have been developed to resolve this issue, however, the additional processing time and equipment required to dispense and cure the underfill result in increased cost.

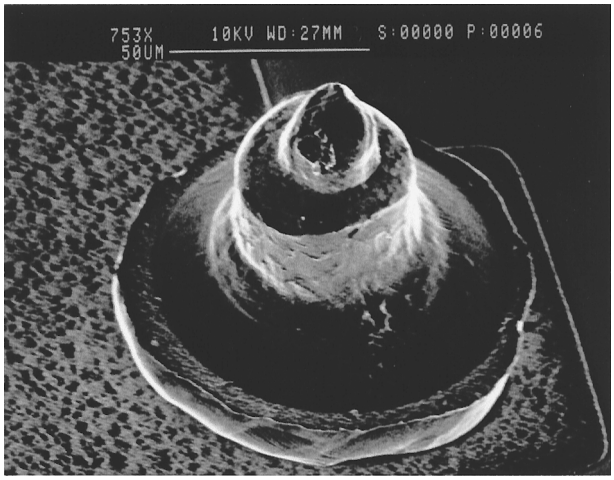


FIGURE 14 Scanning electron microscope photo of a stud bump.

E. Die Bonding

Chip carriers can be classified as cavity-up or cavity-down packages. Cavity-up implies that the functional surface of the die faces away from the component to card interconnects and cavity-down, vice versa. Figure 5 shows examples of cavity-up and -down packages. Cavity-down substrates provide fewer package I/Os due to depopulation of the die site. However, they can provide substantial thermal enhancement since the die is back-bonded directly to a heat sink. This is particularly important for high power dissipation applications, such as, microprocessors and ASICs.

Die can be bonded to substrates or lead frames using epoxies, adhesives, solders, and by diffusion bonding. The most common material used is silver-filled epoxy which is both electrically and thermally conductive. These materials can be accurately dispensed from syringes on to the substrate. After the die is placed, the epoxy is cured in a belt oven. The entire process is automated and offers a low-cost die attach methodology. Adhesives are used for some applications; anisotropic conductive adhesives, mentioned earlier in conjunction with flip chip bonding, form the conductive joint and bonds the die to the substrate. Very low-cost commercial applications use adhesive films to tack the die in place prior to wire bonding. Solder has been used to bond die with back side metallization directly

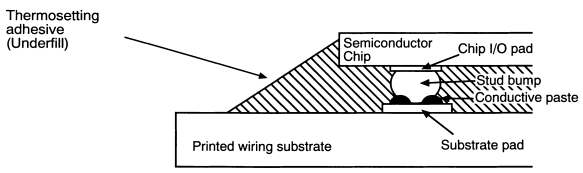


FIGURE 15 A completed flip chip stud-bump assembly using conductive adhesive.

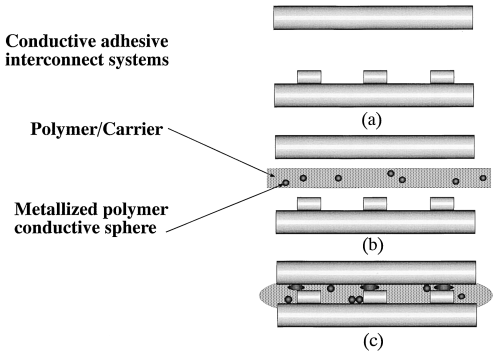


FIGURE 16 A schematic diagram showing the assembly process steps for an anisotropic conductive adhesive interconnect system.

to PCBs to enhance thermal dissipation. One such application involves bonding power transistors on power supply cards. Diffusion bonding, used primarily with ceramic substrates, involves placing a thin gold foil between the die and substrate die pad. A scrubbing motion is applied to the die while the substrate is maintained at a very high temperature. This results in the formation of a gold-silicon eutectic at the interface. These bonds have very high die shear strength and are used for high reliability applications.

F. Die/Bond Encapsulation

Packaging objectives include protection of the die and reliable interconnections. This is accomplished, for most packages, by encapsulating the functional side of the die and fragile interconnects. Figure 5 shows several examples of encapsulated die. The exception is ceramic substrates that have lids or covers. These are typically hermetic packages for high reliability applications.

Molding compounds are one of the most widely used encapsulants. Injection molding technology has advanced to the point where fragile wire-bond wires can be molded without shorting. Molding is used on wire-bonded J-lead and gull wing lead frame packages and with flex and laminate substrates for BGAs and CSPs.

Glob-top encapsulants are epoxy-based materials that are dispensed on the functional side of the die covering the die and interconnects. Both TAB and wire-bond interconnected die utilize this methodology.

Underfill is an epoxy-based encapsulant specifically for flip chip applications. It fills the space between the die and substrate while surrounding the interconnects and forming fillets at the die edges. The underfill material is dispensed adjacent to one or two sides of the die and it flows by capillary action across the die gap. The function of the underfill is to distribute the stress, resulting from the die-substrate coefficient of thermal expansion (CTE) mismatch, over the entire die surface. Without the underfill the stress would be applied only to the flip chip joints.

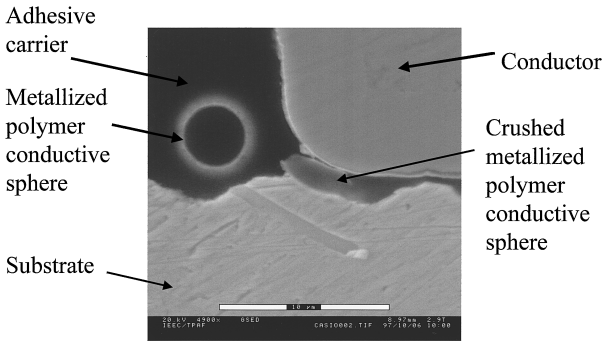


FIGURE 17 A scanning electron micrograph of an anisotropic conductive adhesive interconnection. Metallized polymer spheres are 10 μm in diameter.

G. Finishing Operations

After the die and interconnects are encapsulated the product part number, date code, and manufacturers name or trade mark are marked on the component by laser scribing or epoxy screening. In the case of epoxy, there is an oven cure required. Some components are assembled in strip format, i.e., four or more components tied together by the substrate. Most high volume manufactured components are assembled in this manner with the exception of ceramic, substrate-based carriers. Individual components must be excised (cut from the strips), leads formed (J-lead, gull wing, etc.), and leads coated with a material that will not oxidize. Typically coatings for copper and alloy 42 leads and pins are solder or tin. Components that will be plugged into sockets, rather than soldered to cards, have leads or pads that have been plated with gold or palladium and, therefore, require no tinning. Ball grid array, CGA, and CSP packages must have solder balls or columns attached. A final electrical test must be carried out to insure that the component functions properly. Some high reliability components are "burned-in," where they are exposed to high temperatures to cause failure in defective die. A final inspection for defects is completed on a sample population. Lastly the components are loaded into trays, in bubble pack, or on reels for shipping to the second level assembly facility where they will be attached to a PCB.

H. Card Assembly

Card assembly refers to the process of attaching components to PCBs. Eutectic tin lead solder with a melting point of 183°C is the interconnect material of choice. However, other solders may be used for specific applications. Higher strength joints can be obtained with 62Sn/36Pb/2Ag. If

solder temperature hierarchy is important, then lower meting point solders, such as indium-based materials are applicable. There are two basic soldering techniques that are used today—wave soldering and surface mount soldering. Wave soldering is used only with PIH components where as surface mount can be used for attaching PIH and SMT components. Card assemblies can be entirely PIH or SMT, or a combination of both. Also components can be placed on one or both sides of the card.

I. Wave Soldering

The wave soldering operation involves moving a PCB with PIH components across a molten wave of solder. In this process only the bottom side of the card and pins protruding from the circuit card PTHs come into contact with the solder. Components are placed, leads are clinched to hold the components, the assembly is preheated, fluxed, passed over the wave, cleaned, and dried. A typical solder joint for a PIH component is shown in Fig. 18.

One significant drawback of PIH technology is that the PCB must have through holes or vias to accommodate the leads. For high-density area array packages, card wiring was impeded by the presence of the holes forcing additional layers in the PCB in order to escape the component footprint. Demand for increasing component I/O density in conjunction with decreasing physical size and the limitation of PIH technology was the primary reason for development of surface mount technology.

J. SMT Soldering

Surface mount technology entails attaching the component leads to surface pads on the circuit card with solder. The copper pads are coated to inhibit oxidation. Coatings may be tin, solder, nickel/gold, palladium, or an organic solderability preservative (OSP). Solder paste consisting of a carrier, flux, and solder balls is screened through a metal stencil on to the card pads. Eutectic solder paste is most commonly used. Stencil thickness can vary from 100 to 500 μm depending on the pad pitch. Pad pitch in the range of 250–300 μm typically requires a 100-μm thick stencil. The majority of surface mount applications can be successfully assembled with 150–200 μm stencils. Some manufacturers also screen or dispense adhesives on the card to tack and keep components in place through the reflow process. Components are picked and placed by automated tools that use sophisticated alignment systems to insure that component leads and PCB pads are properly registered. Laser and optical or vision tools are most common, the latter allows component leads to be inspected as well. Assemblies are reflowed in a multizone belt oven at temperatures that range from 210 to 240°C when using eutectic solder paste. Cleaning may be required depending on the flux type contained in the solder paste and reliability requirements for the system. Electrical test and inspection for defects complete the process.

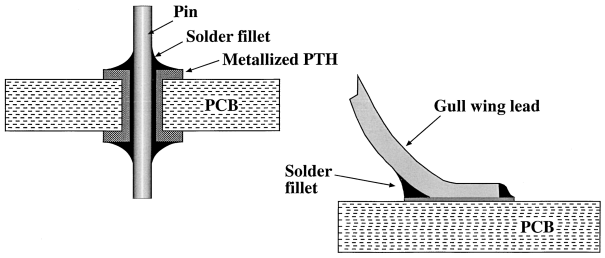


FIGURE 18 Schematics of a Pin in Hole (PIH) and gull wing surface mount solder interconnect.

Defective parts, whether a visual or a test fail, may be reworked or repaired. Repair can mean removing a solder bridge between two leads or soldering a lead that was open. Rework is the process of removing a component, cleaning the removal site, preparing the site for a new component, and attaching the new component. This process requires localized heating on the PCB, so that solder joints on neighboring components are not reflowed during removal and attach steps. Rework is a semiautomated process that is costly. The decision to rework assemblies will depend on their value.

III. DESIGN CONSIDERATIONS

A. Electrical Design Considerations

As discussed earlier one of the primary functions of an electronics package is to serve as an interconnect system for the chips (processors and memory), the storage media, and the I/O devices, among others. Thus the primary electrical design objectives for the packages in a system are to provide paths that minimize noise generation and maintain the best possible signal integrity. It is therefore desirable to minimize parasitic effects such as delta-I noise in the power distribution system and cross talk between interconnects. Delta-I noise (also referred to as switching noise) occurs due to parasitic inductance of a power distribution system since it does not change the supply current at the same rate as the switching circuits. Cross talk, on the other hand, is caused by the mutual capacitances and inductances between the interconnects. This arises due to the need to keep the electrical signal lines in close proximity to each other in order to minimize the signal flight time. Some of the key package design attributes that impact the overall system electrical performance are materials selection, signal and power line dimensions, distance between signal

lines, distance between signal lines and power lines, and overall signal flight path. Other design attributes that effect the system's electrical performance include the speed, voltage swing, terminal impedance, and the noise tolerance of the receiver and the driver circuits. Additional complexities in the electrical analysis arise due to the location of line discontinuities which are caused by the presence of connectors and via holes. Additional reading material on the electrical design of electronics packages may be found in Tummala *et al.* (1997) and Brown (1999).

B. Thermal, Mechanicals, and Reliability Issues at Different Levels of Packaging

One of the key reliability concerns in packaging is due to mechanical issues that arise from thermal gradients. Packages are intrinsically composed of many different materials, with substantially different properties, by design. Furthermore packages have many interfaces which are typically attached together by either a lamination process, an adhesive material, or solder joints. All of these interfaces are subjected to mechanical stresses during assembly, rework, and field operation, and each of them is a potential failure site. As an example consider Fig. 11 which is a cross-sectional view of a typical flip chip organic chip carrier with BGA interconnections. The package consists of the Si die, attached to an organic substrate by solder bumps (C4) of about 127 μm diameter. The space between the chip and the chip carrier is filled with an underfill material which typically consists of an epoxy that is highly filled with a solid that has a high Young's modulus and a fairly low CTE. The entire first level package (chip and substrate) is attached to the PCB by means of the solder BGA, which provides electrical connectivity as well as mechanical support. The solder BGA is about 900 μm. In this typical package the variation in properties is substantial.

The Si die has a CTE of about 3, the substrate and the PCB typically have an effective CTE (they are actually composed of layers of resin and woven glass fiber, with layers of copper laminated within the structure) in the range of 15–20 ppm/°C. The underfill material has an effective CTE of about 25–30 ppm/°C. Some of these CTE values are significantly higher if the temperature is above the glass transition point of the material. This wide range of material properties and the fact that the package is typically subjected to a temperature distribution mean that all of the interfaces between the different layers, and the solder bumps, experience thermomechanical stresses. These stresses arise during the assembly process, in the form of residual stresses. As the package is then operated in the field, the resulting temperature distribution causes a cyclical stress that recurs whenever the die is powered up. This can occur during an on-off cycle of the system, or if the die temperature varies significantly during normal operations. In general thermomechanical issues arise during the development, manufacture, assembly, and field operation of electronic packages. These issues can be classified, according to stage of the life of the product, into the following general areas:

- Design choices based on the mechanical and thermal performance of the package—these choices are made very early in the development stage and are typically based on a combination of predictive modeling and experimental measurements on prototype hardware.
- Residual stresses that arise during the assembly and manufacturing processes of the packages—these residual stresses may impact the reliability of the product as well as manufacturing yields.
- Issues that arise during repair and rework of the packages. These typically arise when a specific component needs to be reworked or removed from an assembly. Issues include potential damage to the printed circuit board and to adjacent components.
- Issues that arise during shipping and storage of the packages.
- Issues that arise in stress testing during the qualification stage of the development of a new package. Stress testing is intended to be an accelerated field simulation. Typical stress tests include accelerated thermal cycling and ship shock testing among many others.
- Issues that arise during field operation of the system.
- Each of the above areas is governed by a number of different variables. For example, design choices, stress testing, and field issues are all dependent upon the actual application conditions. The required life of the product, in terms of power on hours (POH), or years of operation, is one of the key parameters. The life of the product is a function of the operating environment and the actual product characteristics. On the other hand, assembly and rework issues are a function of material properties, overall package geometry, and manufacturing or assembly process parameters such as temperature/time excursions.

As an example to demonstrate the thermal, mechanical, reliability, and electrical considerations in systems design consider the IBM 3081 processor. That module has a multichip ceramic substrate with 100–118 chips, maximum of 4 W per chip, 300 W per module, and a maximum chip junction temperature of 85°C. The cooling medium in this case is water. The power dissipation in this unit is at 20 W/cm^2 at the chip level or 4 W/cm^2 at the module level. The thermal conduction module (TCM) is designed to remove heat from these high-powered modules, see Chu *et al.* (1982). In the TCM, shown in Fig. 19, heat from the

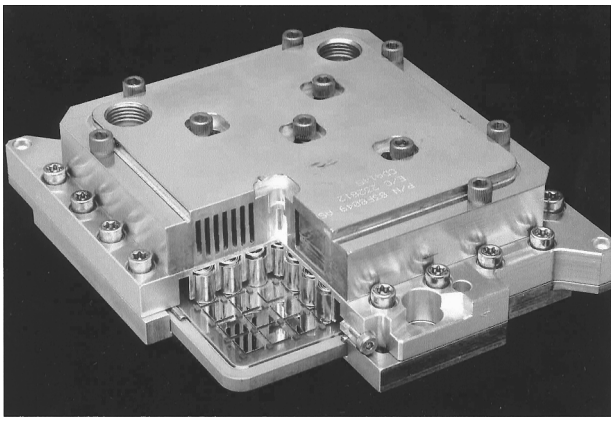


FIGURE 19 Photograph of an IBM Thermal Conduction Module (TCM).

chip is conducted through a series of solids to the water. A spring-loaded metal piston is in contact with the chip. The spring assembly is filled with helium gas to reduce interfacial contact resistance. Thermal conductivity of helium is five to six times higher than the thermal conductivity of air. Thermal conduction module requirements are outlined by Oktay and Kammerer (1982).

It is convenient to think of and evaluate a thermal resistance from the die junction up to the edge of the package (the conduction zone). This thermal resistance is conventionally referred to as the "internal thermal resistance," R_{int}. The internal resistance is typically a sum of conduction in bulk materials, spreading resistances, area constrictions, and contact resistances. The definition of the thermal resistance is $R_{th} = (\text{Delta T})/Q$, where Delta T is the temperature rise relative to the surroundings and Q is the power dissipation in the chip. The nominal internal resistance of the TCM, from the chip junction to the TCM hat, is about 8°C/W, and $R_{ext.} = 0.02$°C/W, this results in a chip junction temperature of 55°C ± 8°C, which is well below the required 85°C. The 8°C uncertainty in the above estimates is due to variability in surface roughness, geometry, chip tilt, etc., which are inherent in any product.

The development of the TCM predated high modulus underfill materials. Here also C4 fatigue due to thermo-

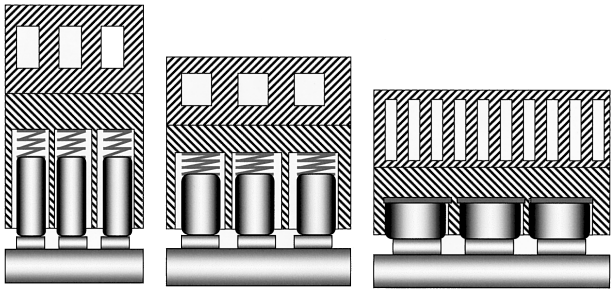


FIGURE 20 Schematic diagram showing the evolution of the TCM design over time.

TABLE II Showing the Relative Incremental Improvements made to the TCM [From Simons, R. E. (1995)]

IBM system	3081	ES/3090	ES9000
Year	1980	1985	1990
Max chip power (W)	4	7	27
Max module power (W)	300	600	2000
R_{int} (°C/W)	11.6	7.7	1.8
Chip heat flux (W/cm^2)	19	33	64
Module heat flux (W/cm^2)	3.7	5.3	11.8

mechanical cycling was a primary design concern. The interfaces between the pistons and the chip and between the pistons and the housing are designed to maintain a known and allowable load on the chips, and hence the C4s. However, the thermal resistance at these interfaces is about 3°C/W each, resulting in a total internal resistance of about 9°C/W. This relatively high thermal resistance would be substantially reduced in future designs as discussed below.

There have been several improvements made to the TCM as new systems were developed, as illustrated in Fig. 20. Increases in cooling capability were achieved by refining the tolerances in each subsequent design, Simons (1995). Emphasis was placed on reducing the length of the piston, refining the tolerance between the piston and housing, and replacing the helium with oil. The aluminum pistons were also replaced with copper ones. Table II shows the relative incremental improvements in the TCM.

IV. EMERGING TECHNOLOGIES THAT RELY ON ELECTRONICS PACKAGING; MEMS, PHOTONICS, AND BIOMEDICAL APPLICATIONS

There are several new emerging technologies and applications that are intrinsically based on electronics packaging, such as MEMS, Photonics, and Biomedical devices. For many of these new applications most of the emphasis has been on demonstrating the feasibility of the devices, rather than on packaging, qualification, reliability, and mass production. This is particularly true in the case of MEMS wherein very little work has been reported on the results of failure modes, failure rates, and acceleration factors in stress testing of packages. One of the key issues needed in order to accomplish that is physics-based modeling of these devices and technologies. These areas present excellent opportunities for research and development activities by both academic and industrial scientists. Some of the key research areas include the development of analytical and experimental algorithms and methodologies to address these new fields. In the area of modeling, for

example, there are key areas such as thermal and mechanical analyses, which will require new analytical tools. Although the physical effects and attributes have not changed and are well known, effects that were previously negligible in large-scale problems (such as electrostatic effects) may dominate in applications like MEMS. In the area of experimental research scale is very important. Many of the experimental tools needed to conduct research at the microscale may not be available or may require substantial development. These issues make emerging technologies in electronics packaging challenging and exciting, with the potential for substantial new discoveries and findings.

SEE ALSO THE FOLLOWING ARTICLES

BIPOLAR TRANSISTORS • EPITAXIAL TECHNOLOGY FOR INTEGRATED CIRCUIT MANUFACTURING • INTEGRATED CIRCUIT MANUFACTURE

BIBLIOGRAPHY

Brown, W. D. (1999). "Advanced Electronic Packaging with Emphasis on Multichip Modules," IEEE Press.

Chu, R. C., Hwang, U. P., and Simons, R. E. (1982). Conduction cooling for an LSI package—A one-dimensional approach. *IBM J. Res. Dev*. **26**(1), 45–54.

Harper, C. A. (1997). "Electronic Packaging and Interconnection Handbook," McGraw-Hill.

Oktay, S., and Kammerer, H. C. (1982). A conduction cooled module for high performance LSI devices. *IBM J. Res. Dev.* **26**(1), 55–66.

Prismark Partners, "The Electronics Industry Report," Prismark Partners Annual Report, 1998–99.

Simons, Robert, E. (1995). The Evolution of IBM High Performance Cooling Technology, *IEEE Transactions on Components, Packaging and Manufacturing Technology-Part A*, Vol. 18, No. 4.

Tummala, R. R., Rymaszewski, E. J. and Klopfenstein, A. G. (1997). "Microelectronics Packaging Handbook," Second edition, Chapman & Hall.

Moore, T. M., and McKenna, R. G. (1993). "Characterization of Integrated Circuit Packaging Materials," Manning Publications Company.



Molecular Electronics

Mark A. Ratner
Northwestern University

Mark A. Reed
Yale University

I. The Molecular Scale and Device Function
II. Molecular Wire Junctions: Interconnects at the Molecular Scale
III. Molecular Wires: Modeling Circuit Conductance
IV. Molecular Electronics: Beyond Single Wires

GLOSSARY

Conductometry The measurement of the conductivity of a particular circuit, used to measure the amount of an analyte material that is present within that circuit.

Mesoscopic Refers to a length scale between nanometers (1 billionth of a meter) and micrometers (1 millionth of a meter). More generally, it can be used to mean length scales between the macroscopically observable and the molecular scale.

Molecular recognition The ability of two or more molecules to bind to one another, because of effective attractive interactions arising from their molecular structures.

Nanoscale Refers to length scales of the order of one nanometer, or one billionth of a meter. Most molecular structures exist at the nanoscale.

Oligonucleotide A short segment of double helical DNA; the term refers to a number of base pairs somewhere between one and one hundred.

Rectification The process of resolving an alternating current into direct currents; in molecular electronics, a rectifier is a molecular circuit that passes charge asymmetrically from right to left versus left to right.

Self-assembly Molecular self-assembly refers to the ability of molecules to use their mutual interactions to arrange themselves into particular structures.

THE TERM MOLECULAR electronics is generally used to include those electronic properties of molecular-based devices that are determined by the behavior of the individual molecules. Since molecules have characteristic lengths between 0.2 and 20 nm, molecular electronic devices will generally have at least one nanoscale dimension. Molecular electronic devices that have been demonstrated include such species as molecular conductors, rectifiers, transistors, photovoltaics, light modulators, and light-emitting diodes. Because molecules have characteristic sizes in the nanometer range, their electronic energy levels are quantized, and exhibit spacings corresponding to energies from the near ultraviolet to the infrared. These energy spacings can be controlled or tailored by synthetic chemistry, to make a given molecule with desired electronic energy level spacings and electronic charge distributions. The characteristic molecular electronics applications follow directly from this discreteness of energy at the molecular scale.

I. THE MOLECULAR SCALE AND DEVICE FUNCTION

A. Size and Structure

The information technology advances that have transformed both the economy and the communications of mankind are largely based on semiconductor devices. Integrated circuits and chip technologies are based on lithographic fabrication—that is, on the carving and assembly of silicon wafers to provide integrated devices, which together yield both architecture and function. The characteristic size scale of individual transistors in semiconductor chips has been dropping by roughly a factor of two every 18 months. This behavior is generally referred to as Moore's Law, and is the chief underlying reason that semiconductor technology becomes both cheaper and better with time.

In the natural world of biology and biochemistry, organization is done in precisely the opposite way: instead of carving a small structure from a large one, larger structures are prepared by assembly of smaller components. The intellectual father of molecular electronics is probably Richard Feynman, who in 1960 suggested (in a talk entitled "There's Plenty of Room at the Bottom") that devices could indeed be made both much smaller and much more functional if they were constructed by atomic or molecular assembly. Such biological structures as nucleic acids, helix-bundle proteins, cells, and organisms like people exemplify the richness of structure and function that becomes available simply by permitting molecules to assemble into structures defined on the nanoscale.

The rules of assembly that define traditional silicon electronics are straightforward: one requires some good conformal interaction between semiconductor and insulator layers, such as is offered by silicon and silicon dioxide. The rules of assembly for molecular structures are far more complex—the differing intermolecular forces, ranging from covalent bonds to hydrogen bonds to hydrophobic interactions, act as geometric constraints on the final structure. Choosing appropriate molecules that will assemble into the desired device structure remains a complex challenge, so that molecular electronics is, at the beginning of the century, just beginning as a well-defined area of science and technology.

Because molecules do interact with one another by various forces, a group of molecules will generally assemble into a particular structure: familiar examples include the interface between olive oil and water, liquid crystal displays, and nearly any biological structure from a leaf to an eyelash. The most important organizing motif for molecular electronic structures, then, will be the weak intermolecular forces that act to determine, form, and stabilize such assemblies. Designing molecules such that they will assemble into particular patterns is a new challenge for the chemistry community, which has, in its synthetic efforts, previously focused on the synthesis of individual molecules, without regard to their interactions. Nevertheless the field of molecular self-assembly, including directed assembly, nanoscale stamping, molecular lithographies, and self-assembly chemical linkages, is developing rapidly.

Because molecular electronics depends upon the characteristics of individual molecules held together in an assembly, development of molecular electronic devices and structures requires understanding both of the individual molecule and of the larger assembly in which the molecules are held. The simplest possible molecular electronic structures, then, would be single molecules. Accordingly, molecular electronics is sometimes limited in its meaning to the properties of individual molecules serving as electrical junctions, rectifiers, or other circuit elements. A far more interesting panoply of phenomenon and function, however, can clearly be found in molecular assemblies; it is the electronic properties of molecular assembles that will constitute the challenge and reward of molecular electronics.

B. Device Function

By device, we will mean a structure that gives an understandable (and in the best case controllable and unique) response to a given stimulus. Molecular electronics has usually been taken to imply controlled conductivity or electronic response, because this has been the dominant consideration of research in the area, and will also be the dominant consideration here. Nevertheless extensions to more obvious, and perhaps more significant, sensing, optical, and hybrid applications may take fuller advantage of the recognition properties of molecules. These are discussed in Section IV.

The simplest possible molecular device is just a wire, that is, an object that will carry current between two macroscopically defined end points, or electrodes. An index of the immaturity of molecular electronics is the fact that the true conductance of any given single molecule wire circuit has been definitively measured only for wires made of nanotubes. In particular, carbon nanotubes have been shown both to conduct current in a coherent fashion over long distances, and to provide transistor type responses when a third (gate) terminal is used. Figure 1 shows a molecular wire junction consisting of a nanotube connecting two lithographically prepared platinum electrodes; it also shows the rather complex current/voltage characteristic found in this single molecule device. Figure 2 shows the nanotube structure utilized as a transistor, where voltage

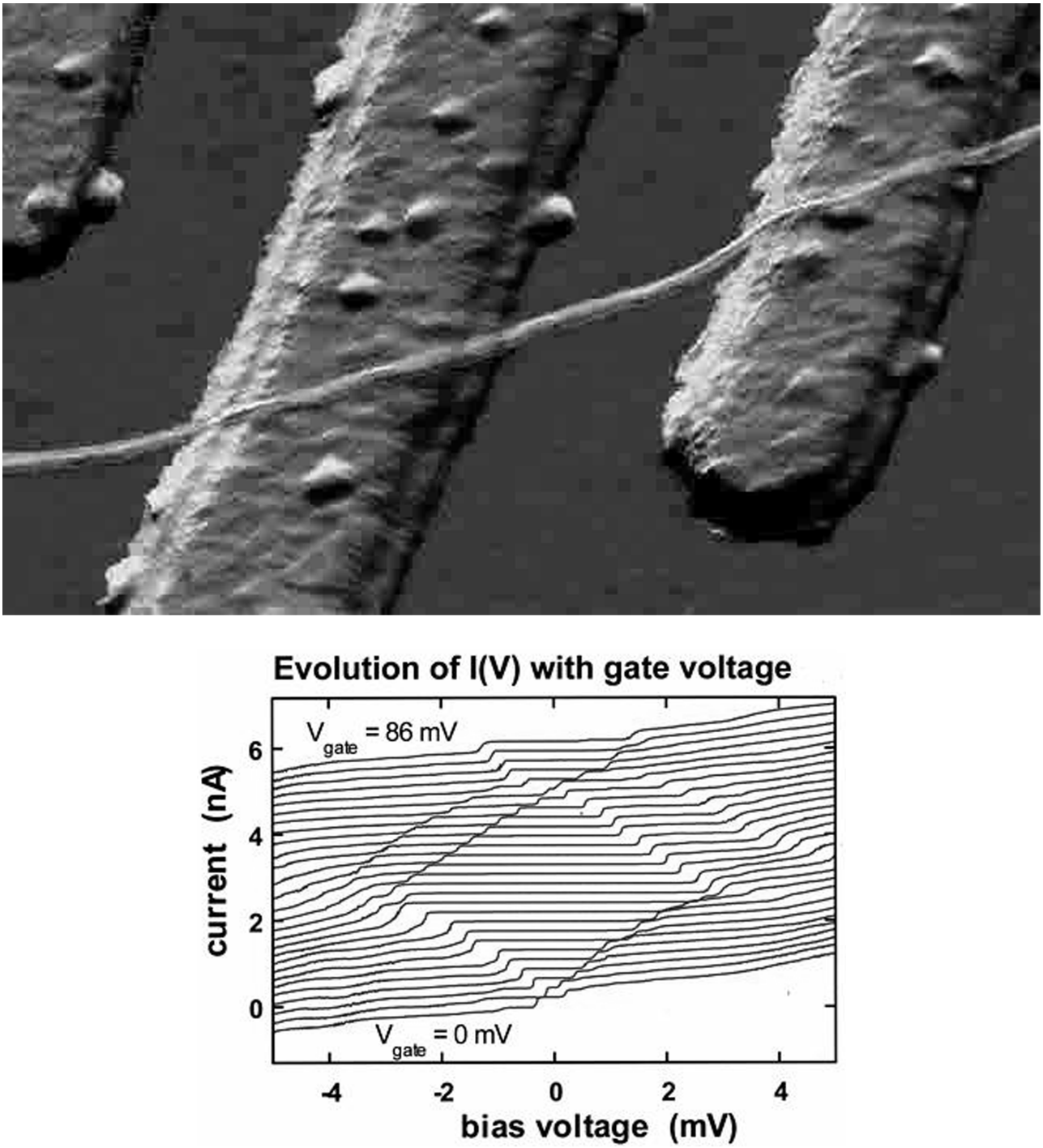


FIGURE 1 Molecular wire junction consisting of a nanotube (smaller structure) connecting to lithographically prepared nanowire electrodes (larger structures). Also shown is the current/voltage characteristic of the nanotube junction, showing response that does not obey Ohm's law. [From Dekker, C. (1999). *Phys. Today* **52,** 22.]

applied to the gate electrode causes charge density within the molecule to distort, with conductance is substantially smaller than that in the unperturbed nanotube. This nanotube transistor, then, constitutes a single two-state device, defined on the molecular length scale.

The nanotube example demonstrates the complexity, the promise, and the challenge of molecular electronics. The nanotubes themselves can vary in length up to several microns, although their cross sections are of the order 1 nm. They constitute single molecules

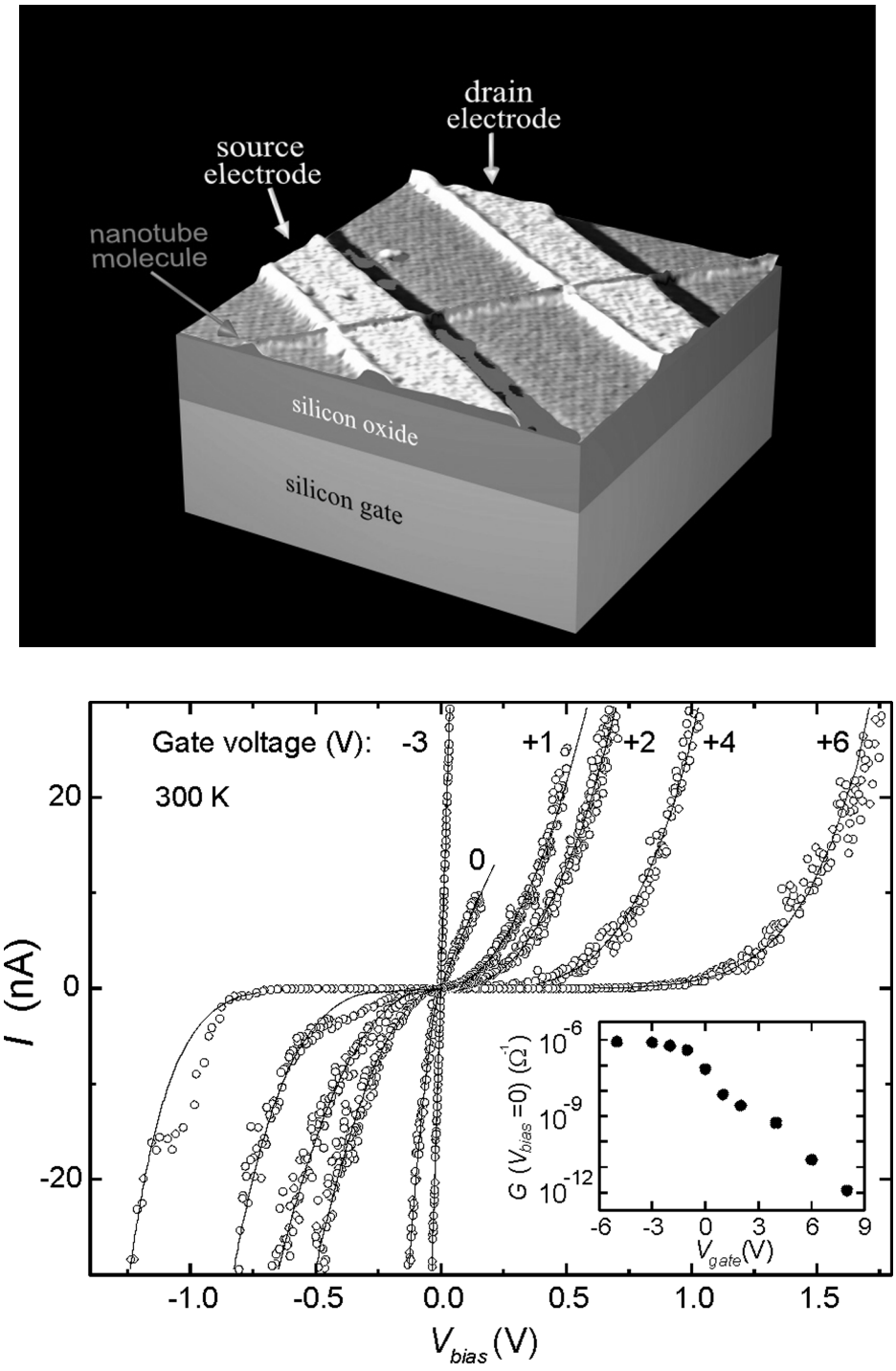


FIGURE 2 A single carbon structure utilized as a field effect transistor. [From Tans, S. J., Verschueren, A. R. M., and Dekker, C. (1998). *Nature* **393,** 49.]

by one definition, although knowing the precise number of carbon atoms in a carbon nanotube or silicon atoms is a silicon nanotube is not generally useful. In this regard, nanotubes differ even from quite large organic molecules such as hemoglobin or chlorophyll, whose precise molecular structures are known and the precise number and disposition of atoms are well defined.

Because molecules have discrete energy levels, they also have well-defined optical transitions. The natural world is marked by lovely colors, many of which arise simply from electronic absorption from one molecular energy level to another. Accordingly, there is extensive interest in molecular optoelectronics, that is, in the interaction of electrical and optical energetics on the molecular scale. Indeed, the first large-scale application of molecular electronics was certainly electrophotography or xerography, in which organic molecules provide a transport channel for excitons, thereby permitting an electrical mapping of light and dark patterns on a document to be copied.

II. MOLECULAR WIRE JUNCTIONS: INTERCONNECTS AT THE MOLECULAR SCALE

A. Mesoscale Properties

Macroscopic structures usually exhibit homogeneous properties as a function of size: a ruby cut in half is still ruby red, a length of copper wire cut in half will show the same conductivity as the original copper wire, and a cup of water will boil at the same temperature as a quart of water. One of the reasons for the current fascination with nanotechnology is that when matter is defined at the nanoscale, nearly all properties become size dependent. The usual examples of this involve the many colors in stained glass windows, which are derived from nanoscale-sized clusters of gold and differing semiconductors. Faraday showed quite conclusively 150 years ago that gold sols prepared with different numbers of gold atoms exhibited a wide range of color. Similarly, most materials in the 1–100 nm range have physical and chemical properties (including color, hardness, melting point, structure, heat capacity, reactivity) that will change with particular size.

One therefore expects that molecular electronics can be tuned by tuning the size and structures of the molecules, and it is precisely this tunability that constitutes part of the promise of the field. This also makes the field far more complex, because the delocalization and coherence properties of the electronic wave functions defined at the nanoscale can be changed by interactions among subunits. This means, for instance, that the conductance of a molecular wire circuit comprising a single molecule stretched between two macroscopic electrodes will depend not only on the molecule itself, but also on the electrodes and the molecule/electrode interface. We should not speak of the conductance of the molecular wire, but rather of the molecular wire junction that is defined by the electrodes and the molecule acting as a unit.

B. Molecular Wire Circuit Measurements

Landauer showed in the 1950s that the conductance of a nanoscale linkage between macroscopic electrodes depends on the transmission probability of an electron wave starting at one macroscopic electrode, passing through the nanoscopic junction, and emerging into the other electrode lead. He showed that in the simplest case of elastic scattering, the current was quantized, that is, the current that could pass through the molecular junction was fixed by the scattering probabilities for all the given states that were available as the electron passes through the wire. This quantized conductance of Landauer's has been seen directly in a number of experiments, including those shown in Figure 3, corresponding to electrical conduction through a wire consisting of a series of gold atoms. The atomic unit of conductance is given by $g_e = 2e^2/h$ where e and h are, respectively, the charge of the electron and Planck's constant. This quantization is shown quite clearly in the conductance of gold wire circuits.

The simplest molecular wire circuits consist of a single molecule linking two macroscopic electrodes, much as the nanotube does in Fig. 1. The assembly process between molecule and electrode, however, is not optimized in Fig. 1. The nanotubes touch the platinum electrodes in many places, and the actual physical structure of the interface, and therefore its molecular electronic transport properties, are not reproducible from junction to junction. Accordingly, much of the study of molecular electronics has focused on different self-assembled structures, whose geometries can be fixed more reproducibly. Particular examples include the interaction between a molecular thio group (that is, a sulfur singly bonded to a carbon) and a less reactive metal such as gold, copper, or silver. Here it is relatively clear from a number of molecular-scale measurements that the sulfur atom lies above the center of a triangle of surface metal atoms. These thio/metal structures are relatively easy to prepare, quite robust, can be annealed and improved, and will support many different molecular components. Therefore they are widespread in the general area of molecular electronics; Section III deals with them in discussing modeling of transport.

Historically, the conductance of molecular wire junctions developed from the preparation of self-assembled monolayer films, that can be only one molecule thick, but extend to macroscopic dimensions along the surface. Although microwave conduction measurements can obtain the conductivity of isolated molecules at finite frequency (where conductance corresponds to moving charge back and forth along the molecule), measurements of actual circuit conductance began with inelastic tunneling spectroscopy of thin film structures. Here it was demonstrated that characteristic molecular vibrations can actually be

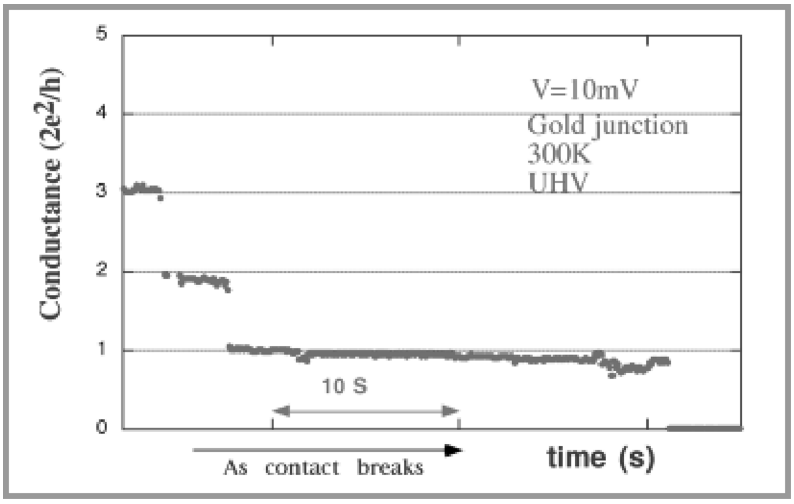


FIGURE 3 Quantized conductance of a wire consisting of a series of gold atoms. Atomic unit of conductance is given by $g_e = 2e^2/h$, where e and h are, respectively, the charge of the electron and Planck's constant. [From Zhou, C., Muller, C. J., Reed, M. A., Burgin, T. P., and Tour, J. M. (1998). *In* "Molecular Electronics" (J. Jortner, and M. A. Ratner, eds.), Blackwell, London.]

seen in the transport measurements, and that they modulate the current passing between the supporting electrode on which the molecular film is created and the probe electrode.

In the 1970s, conductivity in polymeric solids and in stacked inorganic compounds was investigated extensively. In such systems, however, the overall transport is more complex, since it involves not one molecule but many. Actual measurement of molecular conductance really depends on two important experimental advances: the first is self-assembly to produce a molecular wire circuit consisting of one or a very few molecules, and the second is an appropriate measurement technique. Scanning probe microscopy has been used extensively in the analysis of molecular wire circuits.

The first reported number for the resistance of a molecular wire circuit was for an alkane thiol assembled between a gold surface and a gold nanodot (Fig. 4). The deduced value for the molecular resistance was near 12 MΩ which is roughly 1000 times the atomic unit of resistance (which is $1/g_e = 12.9\,k\Omega$). This measurement was made not on a single molecule, but rather on a series of parallel molecules bridging between the nanodot and the gold surface. The estimate for the resistance of a single molecule was based on assuming independent transport through all the molecules linking a particular dot with the surface, and assuming that precisely six molecules would fit on the particular nanodot face that was used. Other measurements of few-molecule conductance have been made utilizing mechanically controllable break junctions and nanopore structures, both of which are shown in Fig. 5. There have also been measurements utilizing mixed adlayer films, in which the molecular wire circuit is found in the midst of other molecules, whose charge transport is assumed negligible. By suitable kinetic preparation schemes, one can vary the number of active wires down as small as three. All of these measurements are complex and difficult, because they require preparation of an electrode set at some mesoscopic length scale. A simpler technique involves use of mercury electrodes, which will also lead to self-assembled adlayers when thiol-containing molecules are placed on them.

In all of the examples just cited, conductance measurements have been made on traditional organic molecules, some examples which are given in Table I. The usual qualitative understandings of physical organic chemistry suggest that the gap between empty and full molecular orbitals will be smallest in systems with extensively delocalized π-electrons, such as are found in planar unsaturated hydrocarbons. These considerations also suggest that the presence of saturated linkages between unsaturated subunits of the wire can lead to local resistances or throttles, and that transport through such throttles will be inefficient.

C. Observed Molecular Wire Transport Behavior

Table I presents schematically some of the molecular wire transport structures that have been studied and some of the striking results that have been obtained. For short strands of gold atoms quantized conductance is observed. For conductive nanotubes, one again observes quantized conductance following Landauer. For short organic molecules, particularly those that form junctions using the thio group, the conductances are substantially lower than the single conductance quantum. Additionally, such

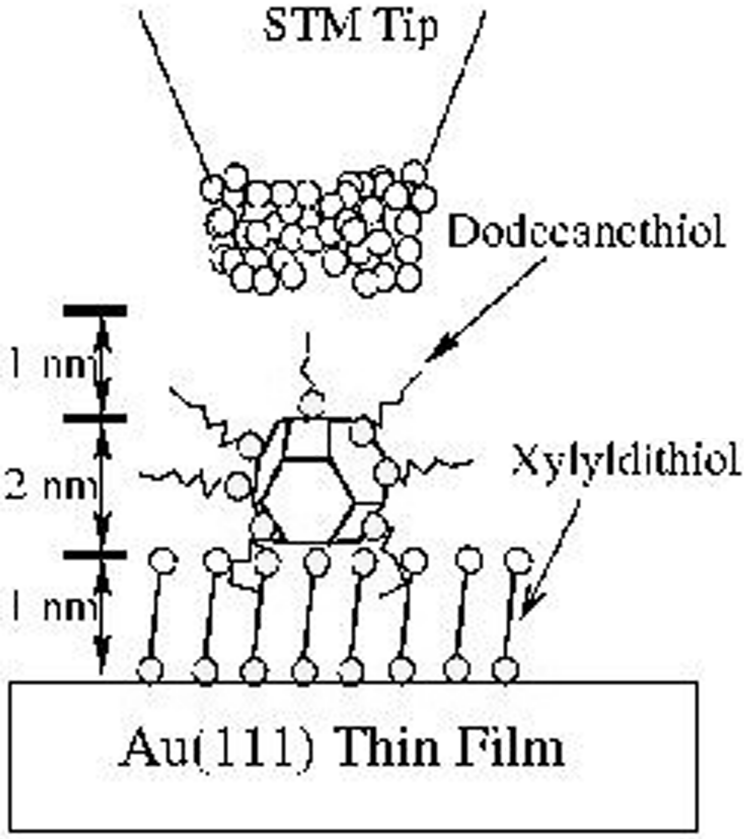


FIGURE 4 The first reported resistance for molecular wire circuits was for an organic dithiol assembled between a gold surface and a gold nanodot (see Table I). The value deduced for the molecular resistance was 12 MΩ which is roughly 1000 times the atomic unit of resistance (which is $1/g_e = 12.9\,k\Omega$). [From Dorogi, M. *et al.* (1995). *Phys. Rev.* **B52,** 9071.]

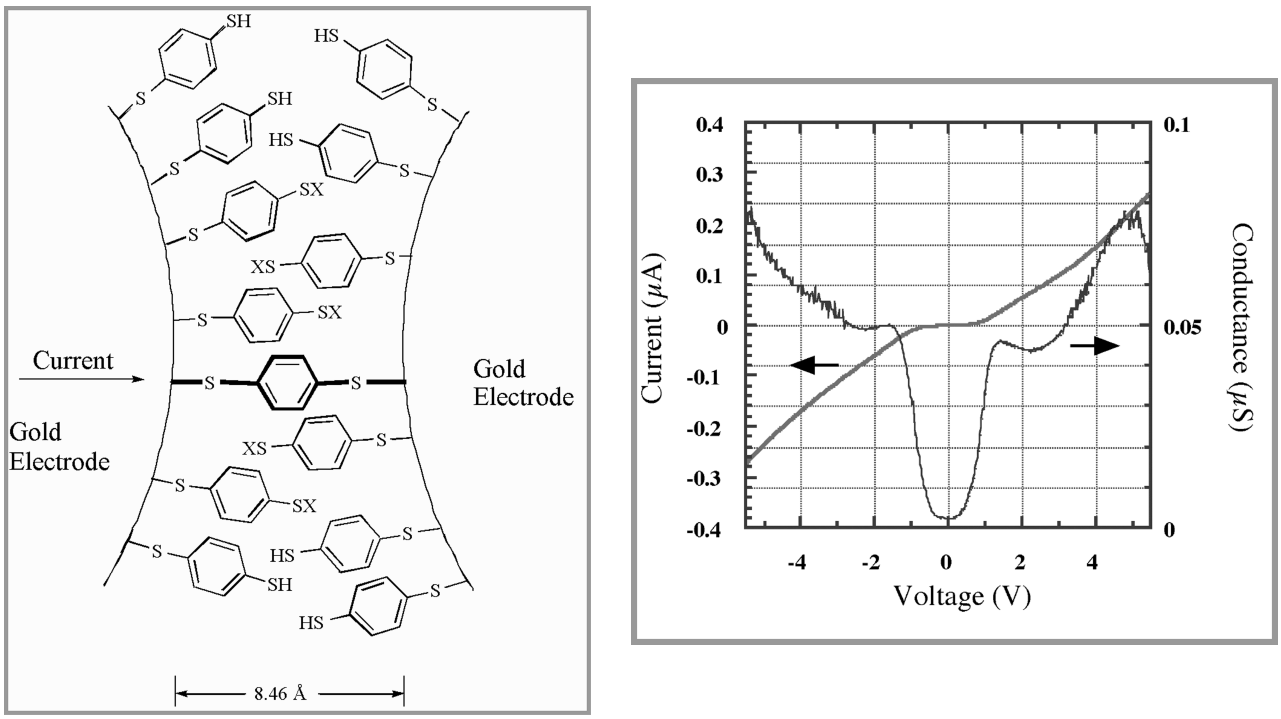
circuits generally demonstrate exponential decay of conductance with molecular length, as is expected based on long-range electron transfer rates in molecules with throttle-type linkages between termini. Finally, long extended chain molecules show conductance that can have striking thermal signatures, including activation behavior or nonmetallic resistance. Actual measurements of individual molecular wires, as pointed out in Section I, have been reported only for nanotubes. Nevertheless, the partial evidence obtained so far, as well as the results from closely related measurements of intramolecular electron transfer, suggest that molecular wire transport can occur by several different mechanisms.

III. MOLECULAR WIRES: MODELING CIRCUIT CONDUCTANCE

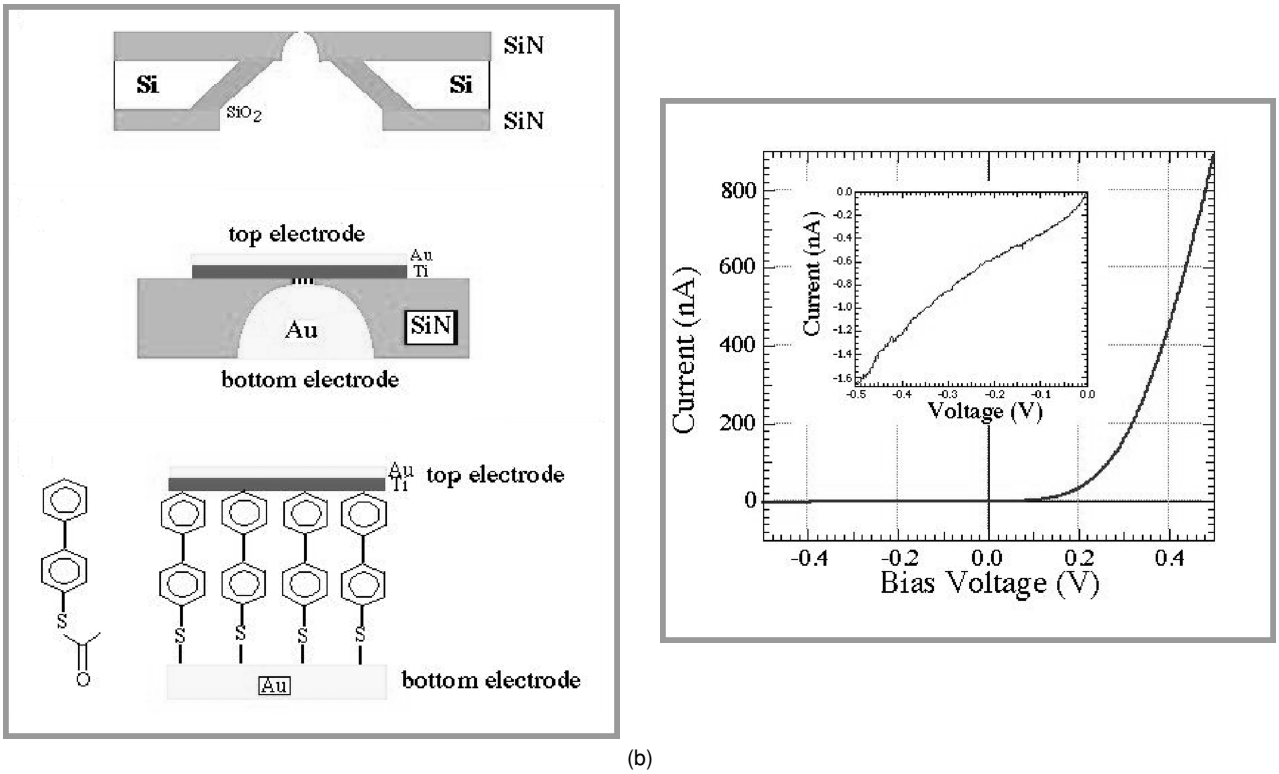
A. Coherent Transport

The molecular wire junction or molecular interconnect is the simplest molecular device. Understanding the behavior of such junctions illustrates both the complexity and the control that characterize molecular electronics. Conceptually, the difficulty with all of molecular electronics is that one deals with an interaction between the discrete electronic levels of a molecule and the continuous electronic levels of the macroscopic structure (electrodes) with which the molecule interacts. The energy levels are schematically shown in Fig. 6. Note that in the metal the spacing between electronic energy levels essentially vanishes, whereas the molecular levels themselves are discrete and separated by characteristic frequencies. The presence of the "bath" electrode states means that the individual molecular states are no longer stationary structures, but that either an electron or a hole placed in one of the molecular levels will eventually decay into the electrode. Therefore the molecular energies will acquire both a width (corresponding to a lifetime) and a shift (corresponding to a change in energy) due to the interaction arising from the contact with the electrodes.

Although the coherent conductance $g_e = \frac{2e^2}{h}$ was originally derived for a constricted metallic junction, one can



(a)



(b)

FIGURE 5 Measures of the molecular conductance utilizing (a) a mechanically controllable break junction and (b) a nanopore. [From Reed, M. A. (1999). *Proc. IEEE* **87,** 652; Zhou, C. *et al.* (1997). *Appl. Phys. Lett.* **71,** 611.]

TABLE I Some Measured Resistances of Molecular Wire Junctions[a]

Electrodes	Molecular wire	Resistance[a]	Ref.
Gold	Au atom chains	1.0, .5, .33, .25. . .[b]	[c]
Gold, Pt/Ir	S—CH_2—C_6H_4—CH_2—S	700	[d]
Gold	(CN—C_6H_4—C≡C)$_2$—C_6H_3 C_2H_5	2500	[e]
Gold/Pt	XX	330,000	[f]
Gold	S—C_6H_4—S	1700	[c]
Gold	H_3C $(CH_2)_n$S	800,000 (n = 10), 3,000 (n = 6)	[g]
Gold/Ti	YY	600,000,000	[i]
Pt	Carbon nanotube	0.5	[j]

Note: XX is a carotenoid, $C_6H_6(CH_3)_3$ CHCH CCH_3 CH CHCHC CH_3 CHCHCHCHCCH$_3$ CH CHCHCCH$_3$ CHCH C_6H_4 CH_2S. YY is the molecule of Fig. 11; this resistance is at the peak.

[a] The resistances are in units of $(g_e)^{-1} \cong 12.9$ kΩ. They refer to electrode/molecule/electrode junctions.

[b] These measurements are for chains of 1, 2, 3. . . independent channels, see Fig. 2.

[c] Zhou, C., Muller, C. J., Reed, M. A., Burgin, T. P., and Tour, J. M. (1998). "Molecular Electronics" (J. Jortner, and M. A. Ratner, eds.) Blackwell, London.

[d] Dorogi, M., *et al*. (1995). *Phys. Rev.* **B52,** 9071.

[e] Andres, R. P., *et al.* (1996). *Science* **273,** 1690.

[f] Leatherman, G., *et al.* (1999). *J. Phys. Chem.* **B103,** 4006.

[g] Wold, D. J., and Frisbie, D. (2000). *J. Am. Chem. Soc.* **122,** 2970.

[i] Chen, J., Reed, M. A., Rawlett, A. M. Tour, J. M. (1999). *Science* **296,** 1550.

[j] Dekker, C. (1999). *Phys. Today* **52,** 22.

straightforwardly generalize the Landauer approach to describe the coherent conductance in molecular junctions. That is, when the molecular bridge is relatively short so that the time for the electron to tunnel from one end to the other is small compared to the periods of molecular vibrations, one can adopt the coherent formalism. Under these conditions, the conductance for very small voltages is given as

$$g = g_e \Delta_L G_{LR}^2 \Delta_R \tag{1}$$

The notation here is schematic: Δ and G are, respectively, the self-energy of the electron sitting on the end of the molecule, and the propagation amplitude for the electron to move from one end of the molecule to the other. The self-energy term Δ depends on the nature of the metal and the bonding of the molecule to the metal. The Green's function is most simply written

$$G_{LR}(E) = \sum_n \frac{C_{Ln} C_{Rn}}{E - E_n - \sum_{nn}} \tag{2}$$

Here the coefficients C_{Ln} and C_{Rn} are, respectively, the mixing of the atomic orbitals of the left and the right atomic ends into a particular molecular orbital. The energy of that molecular orbital is E_n, and its self-energy, due to interaction with the metal, is $\sum_{nn}$. The real and imaginary parts of $\sum_{nn}$ describe energy shifts and lifetimes that vanish in the free molecule. Equation (2) shows that, for transport through a given molecular wire to be efficient, there should be a relevant molecular orbital of energy close to the injection energy E, and that that molecular orbital should mix strongly with metal both on the left and the right end. When the injection is perfectly resonant (when E equals E_n for any given n) the results of Eq. (1) become g=g_e. In general, however, injection does not occur on resonance, and the molecular orbitals do not mix perfectly with both initial and final ends of the molecular wire. Therefore, one normally expects molecular conductance in structures such as those in Fig. 4 or 5 to be smaller than the atomic unit of conductance. Accordingly, the quoted molecular conductances for several molecular bridges, as shown in Table I, are substantially smaller than g_e.

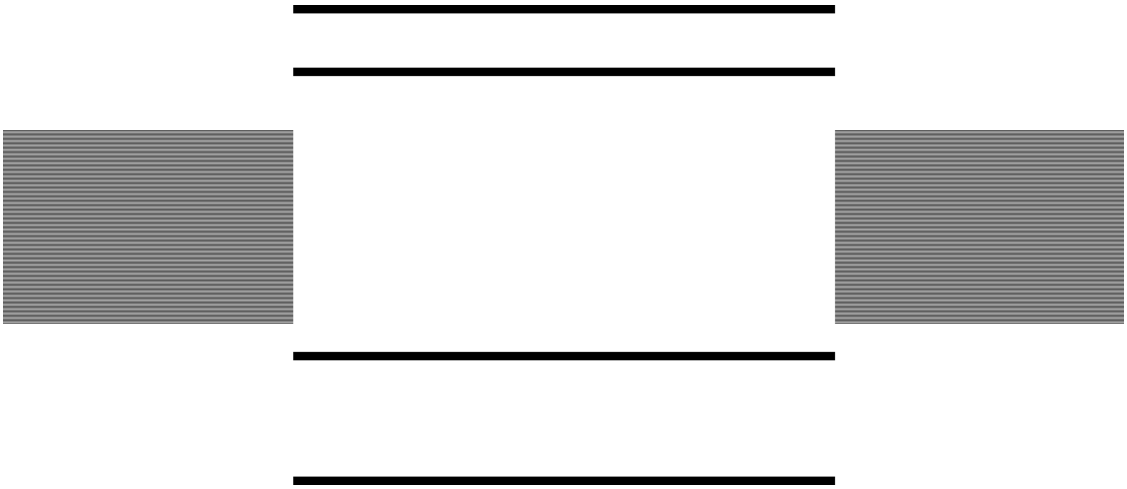


FIGURE 6 A schematic of the energy levels of metal contacts (right and left, continuous energy levels) and a molecule (center, discrete energy levels).

B. Complexities in Molecular Circuits: Voltages, Changes, Injection

The simple form in Eq. (1) ignores any interaction between the injected electron and the vibrational modes of the molecule. Indeed, if the electron is localized on the molecular bridge, it will change the geometry of the molecule, and can form a localized polaron state that will move in a diffusive fashion, and will not obey Eq. (1) at all. Therefore, while Eq. (1) is an important generalization of the Landauer approach to wire circuits, more general considerations are necessary before an accurate general understanding of molecular wire conductance can be established.

The first complications arise from the nature of the potential drop. When a field is applied across the molecular wire junction, the actual profile of the voltage is given by a solution to the Poisson equation, and that in turn depends on the charge of the molecule. One then expects a series of effects, including noninteger populations on the molecule. The complexity of doing electronic structure calculation for noninteger charge is substantial, and therefore a useful approach is the so-called extended molecule structure of Fig. 7. This schematic shows the molecular bridge linked to a small subset of metallic atoms. Those metallic atoms in turn are fully screened from the remainder of the charge distribution within the metal. Therefore the extended molecule can be approximated as neutral, and its Green's function and self-energies can be substituted into Eq. (1). This is important because it permits consideration of excited electron transfer states, charge interchange between metal and molecule, and actual calculation of the nonlinear current/voltage characteristics.

Figure 8 shows the computed charge distribution for an extended molecule structure consisting of *p*-benzenedithiol plus three gold atoms on each end. Notice the substantial change in the population on the molecule, due essentially to the greater electronegativity of sulfur compared to gold. These charge shifts occur even in the absence of an applied field and they must be taken into account for an accurate calculation of the conductance. Indeed, quite elaborate and accurate calculations of coherent conductance using the full self-consistent density functional theory have been given.

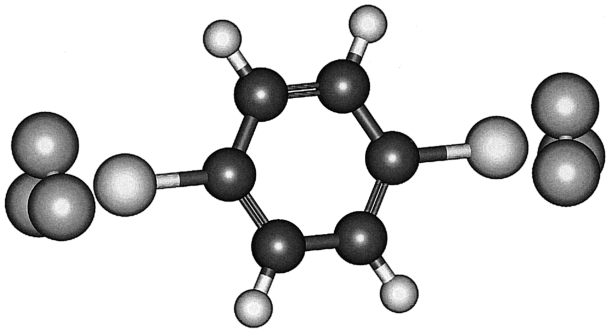


FIGURE 7 The "extended molecule" used in the calculation of tunneling conductance in molecular wire junctions. The molecule is parabenzenedithiol with the two terminal hydrogens missing, and the sulfurs coordinated by a triangle of gold atoms.

One very useful concept for understanding mechanism in molecular wire transport is that of the Landauer/Buttiker tunneling time. This time, denoted τ_{LB}, is the actual time of contact of the transferring electron with the molecular bridge. Generalization of the original Landauer/Buttiker derivation suggests that the tunneling time is given by

$$\tau_{LB} = \frac{Nh}{2\pi\,\Delta E} \tag{3}$$

Here ΔE is the energy gap between the Fermi level of the metal and the closest available energy level of the molecule, and N is the length of the molecule, measured in repeat units of the bridge. This time should be compared with the characteristic vibrational period of the molecule. If these times become comparable, then vibronic interaction on the bridge must be considered, and the simple elastic scattering formula of Eq. (1) will not hold. Conversely, for short molecules or large gaps (ΔE) the tunneling time will be substantially less than any period of molecular motion, and the purely elastic behavior of Eq. (1) should dominate.

The characteristic length dependence of the two dominant sorts of motion within the molecular circuit (elastic tunneling and the inelastic injection onto the bridge) should be quite different. The behaviors in the two regimes are given by

$$\text{coherent: } g = g(R_o)\exp\{-\beta(R - R_o)\} \tag{4}$$

$$\text{injection: } g = \frac{1}{A + BR} \tag{5}$$

Here A + B are system-dependent constants, and R and R_o are the overall junction length and a reference length. It follows from this relationship that at low temperatures or for short wires the elastic, Landauer-type tunneling [Eq. (4)] should dominate. As the wires become longer, however, the exponential decay expected for the elastic scattering mechanism simply makes the current flow too inefficient, and we expect to see, instead, the very slow distance decay and activated temperature behavior, characteristic of the injection mechanism.

C. Observed Behavior in Wire Junctions

Some of these effects have indeed been seen in molecular junctions. Experiments on metal-molecular layer-metal structures reveal an activated temperature, as shown in Fig. 9. The figure shows an activated temperature dependence that is voltage dependent, indicating thermionic emission at the metal-molecule barrier. Analysis of the slope of this activation as a function of bias can

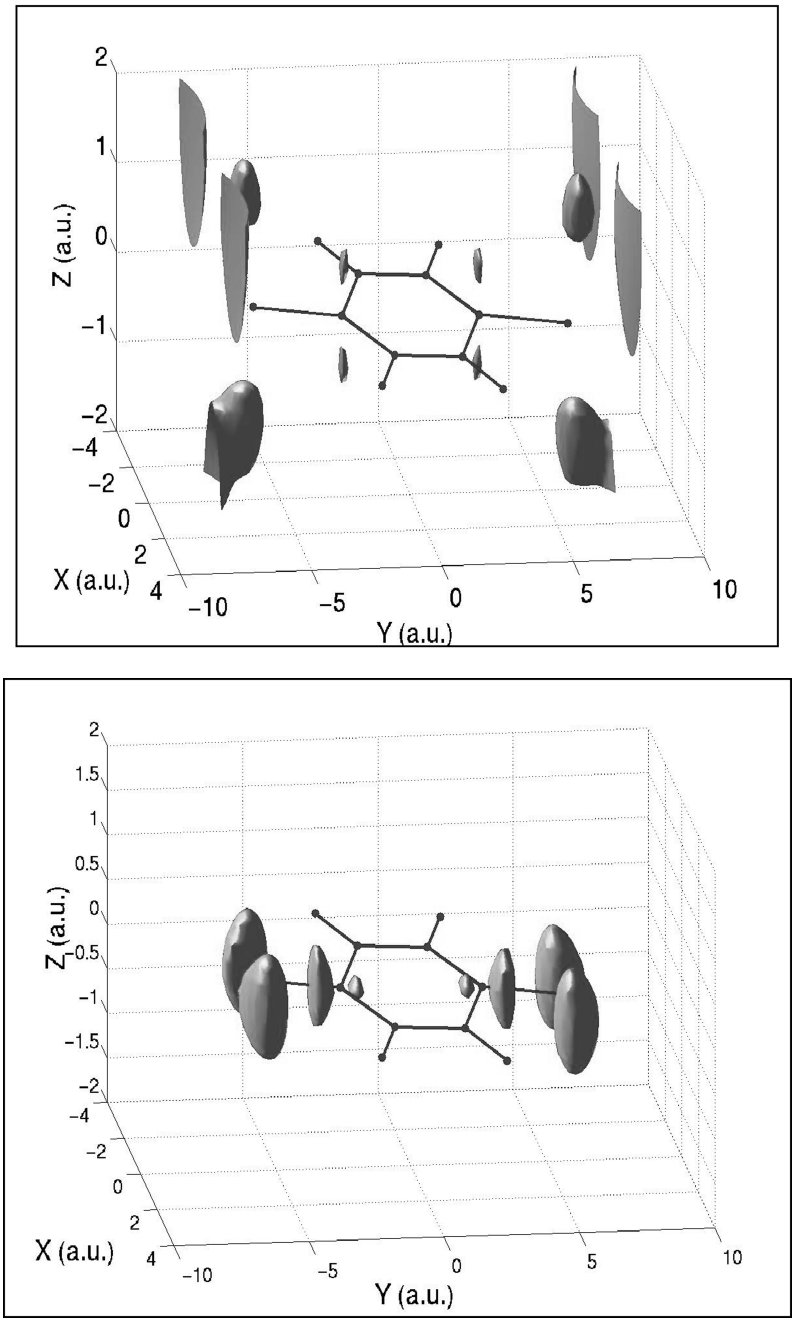


FIGURE 8 The computed charge distribution for the extended molecular structure of Fig. 8, consisting of parabenzenedithiol with three gold atoms at each end. The calculation also uses a Green's function method to describe the bulk gold. Upper figure shows electron density loss; lower figure shows gain (especially at sulfur). [From Xue, A., Datta, S., and Ratner, M. A., *J. Chem. Phys.* (submitted).]

quantitatively determine this barrier, as shown on the right of the Au-isonitrile junction.

D. Other Phenomena in Molecular Wire Circuits

Our discussion in this section has focused on the simple linear current/voltage relationship and even there we found complexities depending on mechanistic conditions and the geometric and energetic nature of the wires. More complex phenomena include those below.

Nonlinear current/voltage relationships—As expected from Eq. (2), the current/voltage characteristic will generally exhibit maxima when there is a resonance between the Fermi level under an applied field and a molecular energy level under that same field.

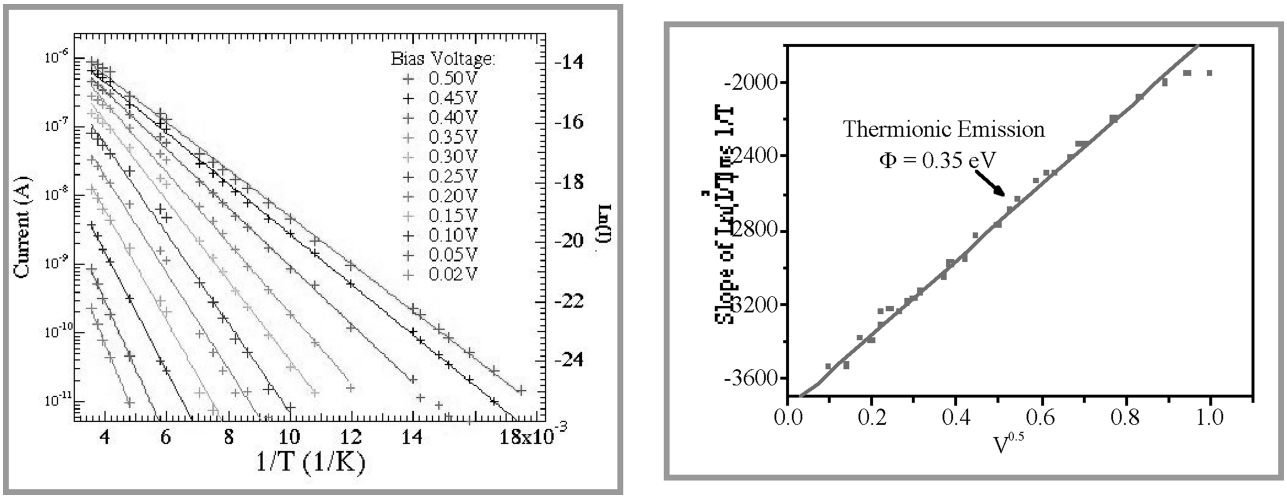


FIGURE 9 Activated temperature dependence typical of a molecule/metal contact. The analysis of the activation as a function of bias can quantitatively determine the barrier, as shown on the right for a gold/isonitrile junction. [From Reed, M. A. (1999). *Proc. IEEE* **87,** 652; Zhou, C. *et al.* (1997). *Appl. Phys. Lett.* **71,** 611.]

Figure 1 shows the characteristic in the case of nanotube wires.

Coulomb blockade—This term simply refers to the coulomb capacitance corresponding to a mesoscopic-scale structure. It was originally coined to describe quantum dot structures, but molecules are simply small perfect quantum dots. Coulomb blockade refers to the fact that the current as a function of voltage will have steps, corresponding to a charging energy of a given molecular structure. Characteristic coulomb blockade structure is shown in Fig. 10—this actually comes not from a molecular wire, but simply from isolated molecular structures probed with the scanning probe current measurement.

Rectification—Equation (1) suggests the current is proportional to the spectral density. If the molecule has asymmetric spectral density (different on the right than on the left), the current is expected to be asymmetric. This rectification behavior was suggested long ago, and has been observed in a number of thin film measurements—it corresponds to one of the first identifiable molecular-scale phenomena of molecular electronics.

Negative differential resistance—The structure of Eq. (2) suggests that as the voltage is increased beyond a resonance, there should be a sharp drop in the current. Formally, this response is a negative differential resistance. Just such behavior has been seen in a molecular wire consisting of π-type systems either separated by methylene groups (effective throttles) from the electrodes or built from moieties that form localized states in the molecule. Figure 11 shows large negative resistance observed in a molecular junction of the second type, where the nitroamine ring is responsible for the peaked current-voltage characteristic.

The transport mechanisms suggested here begin to describe phenomena available within molecular wire junctions. One could expect a series of mechanisms, including (in order of mechanistic complexity): elastic nonresonant tunneling, elastic resonant tunneling as in nanotubes or gold wires of Fig. 1, injection/diffusive transport as in Fig. 9, and gated charge motion (corresponding to the molecule changing its geometry as the electron moves along the wire).

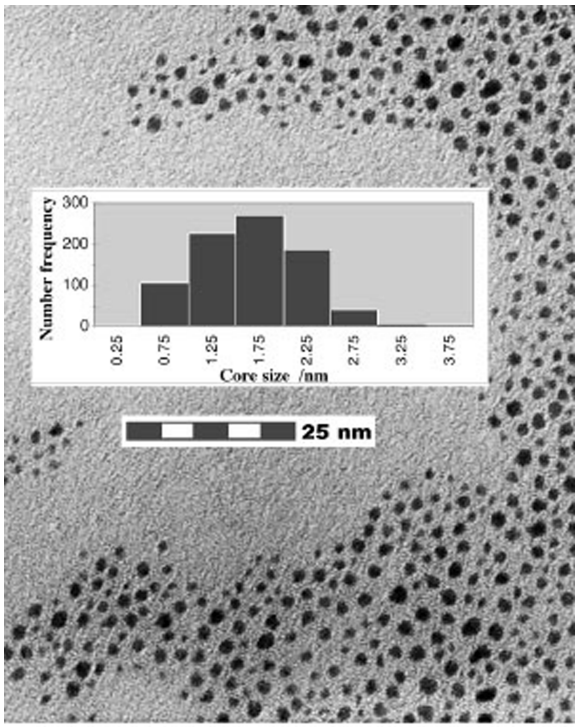
Many of these mechanisms have been observed in isolated systems as an electron transfer rate constant, but one expects all of them to be visible in certain limits of the molecular wire structure. For example, the molecular transistor behavior of the nanotube shown in Fig. 2 is a manifestation of the interaction between molecular structure and charge transport, and is therefore a sort of structural gating of molecular conductance.

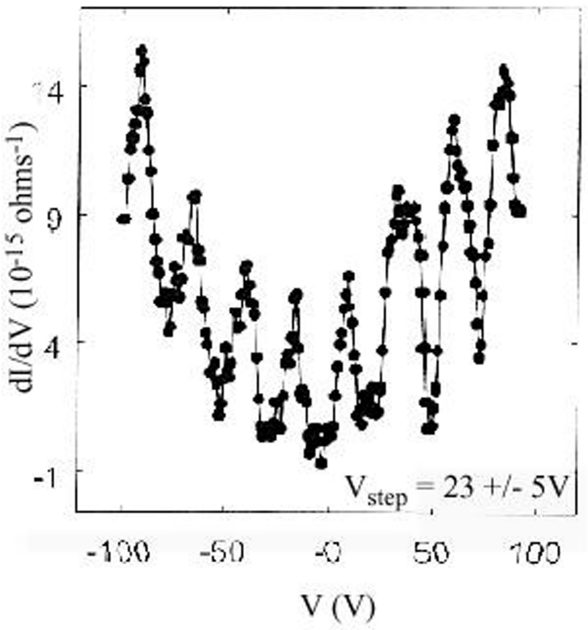
IV. MOLECULAR ELECTRONICS: BEYOND SINGLE WIRES

Molecular wire interconnects, even such functional interconnects as molecular rectifiers or negative differential resistance structures or geometrically gated molecular transistors, are only the most obvious application of molecular electronics. Some other structures that have been extensively studied involve the interaction of molecules with optical and electronic probes simultaneously, and the interaction of molecules with one another.

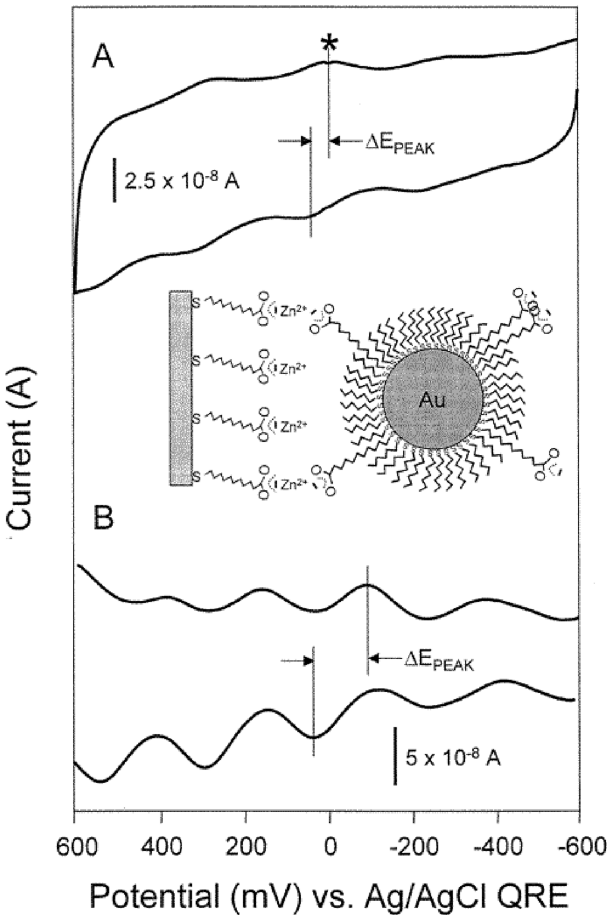
A. Molecular Optoelectronics

Electrophotography, as discussed in Section I, was the first application of molecular scale electronics. The common





(a)



(b)

FIGURE 10 Coulomb blockade structure observed in the current/voltage characteristic of metallic nanospheres in an electrochemical cell. The steps in the current reflect the capacitative charging of the gold nanodots.

motif of molecular optoelectronics is the combination of electrical and optical signal transduction in molecules.

Phenomena of interest in molecular optoelectronics include electro-optic modulation processes, in which molecules excited by resonant radiation change their conductance, or molecules in a static electric field change optical behavior, thus providing modulation of optical and electronic signals. Perhaps the most striking area of progress in molecular optoelectronics has been molecular or polymeric light-emitting diode (LED) structures.

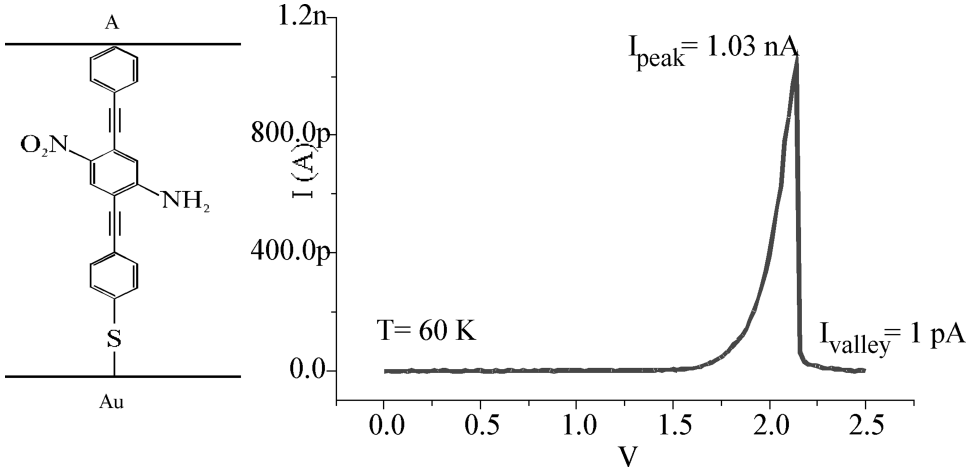


FIGURE 11 Measured negative differential resistance in a molecular junction. [From Chen, J., Wang, W., Reed, M. A., Rawlett, A. W., Price, D. W., and Tour, J. M. (2000). *Appl. Phys. Lett.* **77,** 1224.]

Figure 12 shows a characteristic LED: here electronic and hole current are simultaneously passed from electrodes into a thin film molecular structure; the hole and electron passed through the electrode interface are injected into transport layers and recombined on a molecular luminescence center. Here they combine to form an exciton, a molecular electronically excited state. This state can decay radiatively or nonradiatively. The overall structure provides an optical signal induced by an electronic current, that is, an LED.

Molecular LEDs have made spectacular progress in the last decade, and are probably the most active area of current molecular electronics. Their stability, quantum yields, voltage response, and other properties have been improved by a combination of molecular structure optimization, creative synthetic methodology, and appropriate modeling studies. Current molecular LEDs can perform throughout the visible spectrum and the near infrared, have external quantum efficiencies near 4%, and are very promising for wide area displays, including applications from goggle displays to full room lighting.

The molecular aspects of these LEDs arise both from the current injection problem (actually quite a similar problem to that occurring into the molecular wire circuit) and

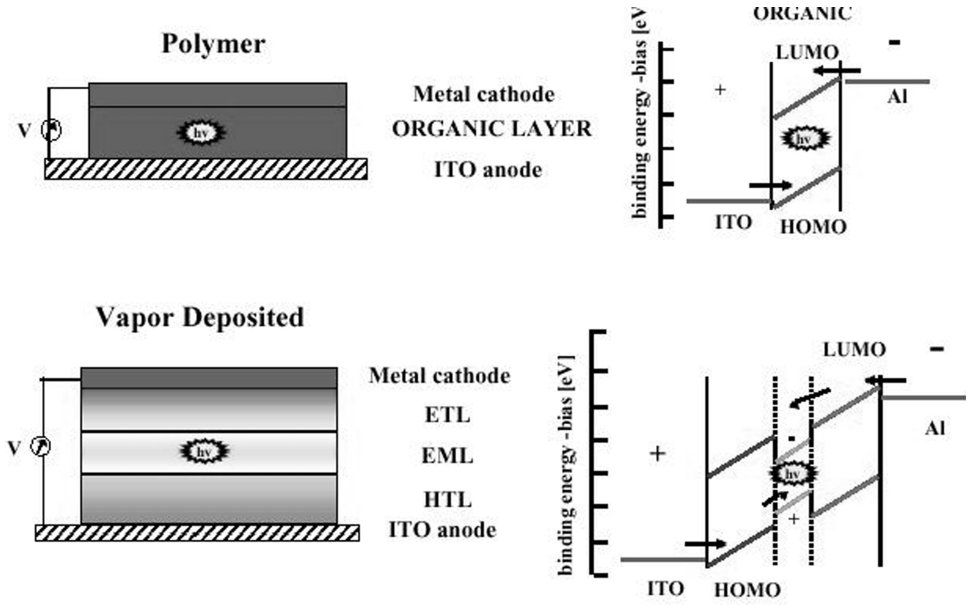


FIGURE 12 A schematic of a molecular LED structure. The electrons are introduced through the cathode and holes are introduced at the anode. Electrons and holes diffuse together and recombine in the molecular emission center, which is normally a large organic or organometallic molecule. [From T. J. Marks, personal communication.]

from optimization of the color, quantum yield, and device stability. By changing the chromophore molecule, one can change the efficiency and frequency of the LED.

When the emitted radiation in an LED is reflected from an interface, one can observe amplified emission and in favorable cases, actual lasing. The electro-optic amplified emission has now been observed both with mirrors and with scattering centers to enforce the optical feedback. Understanding this phenomenon sufficiently to predict lasing onset is a complex problem involving the emission process itself and efficiency of reflections. Quasi-states for the photon seem an appropriate way to discuss the random laser, which offers several advantages in construction, possible quantum yield, and novelty.

B. Molecular Sensors: Recognition Motifs

Molecular scale recognition phenomena can be transduced to become observable using several electromagnetic signatures. Because the passage of electronic current through a molecule depends on the electronic structure of that molecule, conductometric sensors based on molecular recognition at the single molecule level are a rapidly developing area. For example, in a film of gold nanodots functionalized with oligonucleotides, the conductivity of the array will be changed by several orders of magnitude upon binding the complementary oligonucleotide, essentially due to reduction of both capacitive and tunneling barriers for conduction. This application of molecular electronics depends intimately on molecular recognition at the single molecule level, but the observation made is in fact macroscopic.

Extensive work on so-called laboratory-on-a-chip structures is based on the same idea. For example, if a single strand oligonucleotide is present in a self-assembled monolayer on an electrode, binding the complementary oligonucleotide will produce a double helix that can facilitate transport through the adlayer. Again, a conductometric response curve can be constructed and can lead to sensing of individual DNA molecules.

The selectivities of conductive probes for particular nucleotides are quite remarkable. It has been shown that a mismatch of a single base pair between two analytes can give a reproducible, reliable change in conduction. This suggests that such molecular sensors can be used both in wide-scan medical monitoring and in ultrasensitive toxic chemical or biological detection.

Sensing methods based on molecular recognition thus promise effective and reliable applications of molecular electronics that lie well outside the sphere of conventional semiconductor electronics.

An older application of molecular electronic sensing is the so-called CHEMFET transistor. Figure 13 shows the operating principle: when the target molecule is adsorbed on the channel between the source and drain electrodes, it will modify the electrostatic potential acting through the channel. This will change the conductance between source and drain. Although microscopic CHEMFETS have long been known, recent work at the Weizmann Institute has indicated that nanoscale CHEMFETS, which they call MOCSERS, can analyze attomolar amounts of such analytes as water, ammonia, or ethanol.

In biosensing, recognition of particular DNA fragments can provide remarkably efficient monitoring of toxins

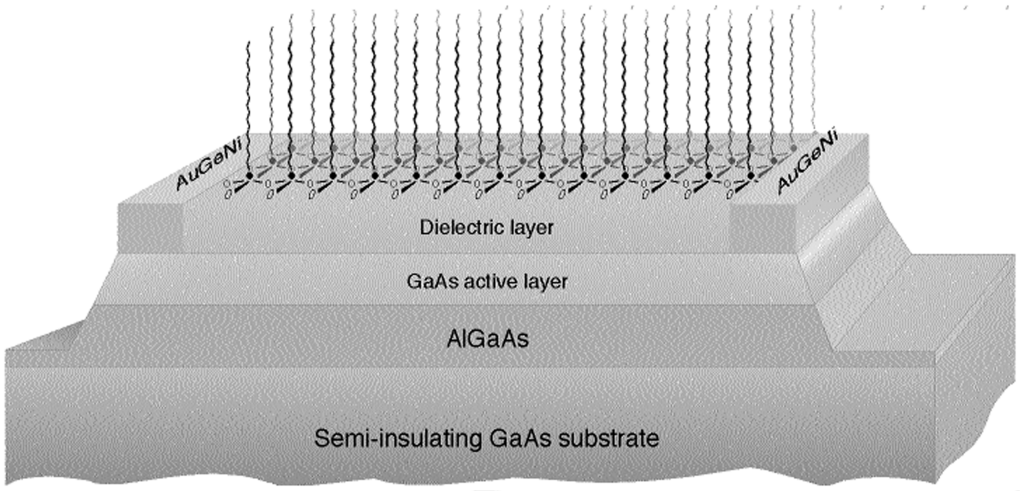


FIGURE 13 A schematic of a nanoscale CHEMFET (chemical field effect transistor). The undoped GaAs region between the cross-hatched AuGeNi electrodes contains a molecular adlayer; those molecules interacting with other molecules change the potential acting through the AlGaAs channel, and therefore modify the current. [From Wu, D. G., Ashkenazy, G., Schvarts, D., Ussyshkin, R. V., Naaman, R., Shanzer, A., and Cahen, D. (2000). *Angew. Chem. Int. Ed.* **39,** 4496].

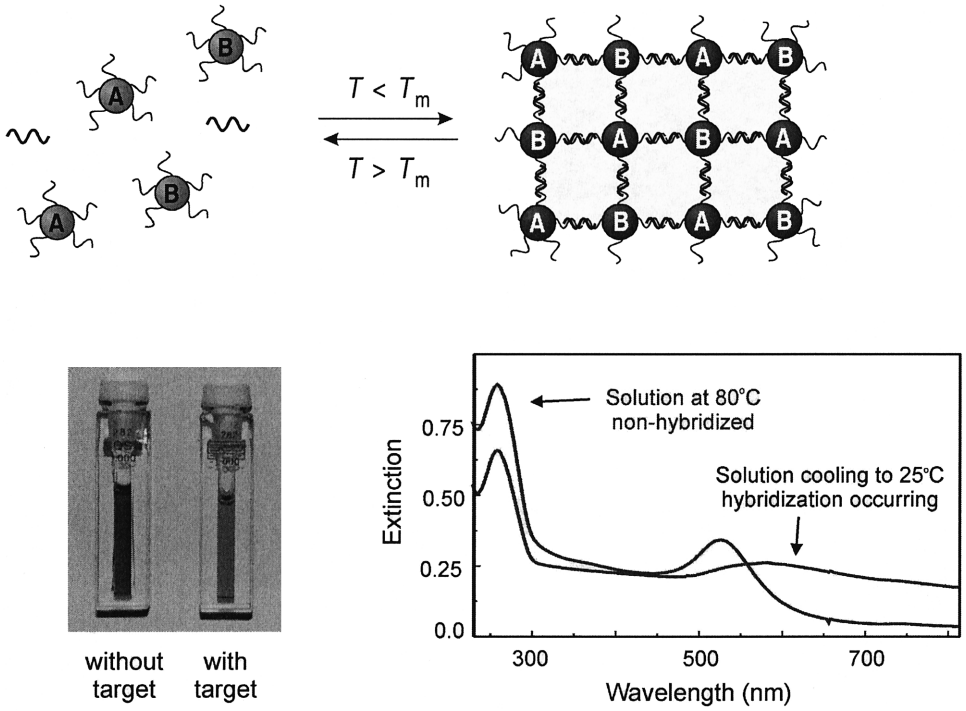


FIGURE 14 Schematic illustration of sensing modality for a particular target. The gold nanospheres schematically indicated by A and B bind a doubly complementary oligonucleotide target (the wiggly shape on the left of the arrows). The material, below the melting point, comes together to form the array on the right, where the distances between the gold spheres are reduced. This results in the color change illustrated in the actual samples at the lower left. [From C. A. Mirkin, personal communication.]

and organisms. The latter recognition process can be monitored optically. At Northwestern, Mirkin and Letsinger and their collaborators have demonstrated remarkable sensitivity (10^{-15} molar, single base pair misfit) for sensing using single strand oligonucleotide targets functionalized on gold nanospheres. Upon binding the target DNA, the nanospheres come closer together, and the color changes sharply (Fig. 14).

The structures just discussed are part of a striking development of molecular scale electronics prototypes. Understanding how electronics, molecular structure, and optics interact at the appropriate nanoscale is a substantial challenge for current condensed matter science, one that has enormous promise in nanoelectronics, bioelectronics, and bioanalytical chemistry. The fundamental ability to make molecular scale structures using self-assembly methods and combinatorics has provided a set of structures whose properties can be varied essentially at will. Development of scanning probe techniques permits one to characterize the electronic and optical electronic response of molecules at size scales corresponding to single molecules. Finally, modeling capabilities permit calculations, in favorable cases, of the structure/functional relations for molecular electronics response.

Extensive applications to molecular engineering will depend upon progress in making molecular assemblies relatively inexpensive and very replicable. The fundamental problem of molecular electronics is beginning to be addressed and we are beginning to understand how to design electromagnetic response on the molecular scale.

SEE ALSO THE FOLLOWING ARTICLES

CIRCUIT THEORY • LIGHT EMITTING DIODES (LEDS) • MICRO-MECHANICAL DEVICES • NANOSTRUCTURED MATERIALS

BIBLIOGRAPHY

Ratner, M. A., and Jortner, J., eds. (1997). "Molecular Electronics," Blackwell, London.

Ariram, A., and Ratner, M. A., eds. (1998). "Molecular Electronics: Science and Technology," *Ann. N.Y. Acad. Sci.* **852,** New York Academy of Science, New York.

Joachim, C., and Roth, S. (1997). "Atomic and Molecular Wires," Kluwer Academic, Dordrecht.

Ellenbogen, J. C., and Love, J. C. (2000). *Proc. IEEE* **88,** 386–426.

Joachim, C., Gimzewski, J. K., and Aviram, A. (2000). *Nature* **408,** 541–548,
Lent, C. S. (2000). *Science* **288,** 1597.
Mirkin, C. A., and Ratner, M. A. (1992). *Annu. Rev. Phys. Chem.* **43,** 719–754,
Petty, M. C., *et al.* (1995). "Introduction to Molecular Electronics," Oxford.
Reed, M. A. (2000). *MRS Bull.* **26,** 113.
Tour, J. M. (2000). *Accounts Chem. Res.* **33,** 791–804.
Tsui, R., Siragusa, L., and Goronkin, H. (2000). *C.R. Acad. Sci. IV-Phys.* **I,** 875–883.
Wada, Y., Tsukada, M., and Fujihira, M., *et al.* (2000). *Jpn J. Appl. Phys.* **I,** 3835–3849.
Ward, M. D. (2001). *J. Chem. Educ.* **78,** 321–328.
Wolkow, R. A. (1999). *Annu. Rev. Phys. Chem.* **50,** 413–441.



Nanoelectronics

D. J. Paul
University of Cambridge

I. Introduction
II. Silicon Nanoelectronics and Ultimate CMOS
III. Single Electron Devices
IV. Quantum Mechanical Tunnel Devices
V. Spin Nanoelectronics or Spintronics
VI. Fabrication Technology
VII. Self-Assembly
VIII. Molecular Nanoelectronics
IX. Alternative Architectures
X. Summary and Future Prospects

GLOSSARY

Nanoelectronics The field of devices controlled using the electronic properties of materials with dimensions below 100 nm in size.

Quantum mechanical tunneling The ability to use the wave properties of an electron to allow transmission (electron tunneling) through a thin potential barrier.

Single electron devices Devices that have switching properties controlled by the addition or subtraction of one electron or through which only one electron may be transported at any one time.

Spin nanoelectronics The use of the electron spin for computation or memory storage.

Transistor A three-terminal device used as a switch or amplifier where one terminal controls the electron current flowing between the other two terminals.

NANOELECTRONICS is the field of computation and control in the nanometer scale regime (below one thousandth the diameter of a human hair) using the electronic properties of materials. Useful logical computation or storage of information may be achieved by a number of different concepts and devices. Logical circuits may be used for computation, communication, control systems, or storage of information. This article reviews the major nanoelectronic devices, fabrication techniques, and architectures. The devices include silicon transistors, single electron transistors, resonant tunneling diodes, magnetic spin devices, and molecular devices.

I. INTRODUCTION

Many new technologies appeared during the 20th century. If one had to decide on which new technology

had the largest impact on mankind, the microelectronics industry would certainly be one of the main contenders. Microelectronic components in the form of microprocessors and memory are used in computers, audiovisual components from hi-fis and videos to televisions, cars (the smallest Daimler-Benz car has over 60 microprocessors), communications systems including telephones and mobile phones, banking, credit cards, cookers, heating controllers, toasters, food processors—the list is almost endless. It is almost impossible to live in the western world today and not use a microelectronic component in everyday life or eat food that has not had a microchip involved in the management, harvesting, processing, distribution, or sale of the product.

The microelectronics industry has only been around for the last 50 years or so. The first bipolar transistor was demonstrated at Bell Laboratories in 1948 and the first field effect transistor (FET) appeared in 1960. Since that date, the microelectronics industry has been growing at an exponential rate, which was first analyzed by Gordon Moore, founder of Intel, in 1965. Moore showed that the size of a transistor gate is halved every 18 months, a feature that is now referred to as Moore's Law. This halving in size has been driven by economics. The smaller the gate, the faster the transistor can switch, the less power it consumes, and the larger the number of transistors that can be integrated onto the one silicon chip. The increase in number of transistors and the associated higher manufacturing yields reduces the cost per transistor, and therefore it is economical to scale the transistor to smaller and smaller sizes.

At the start of 2001, the gate length on the silicon metal oxide semiconductor field effect transistors (MOSFETs) on the latest Intel Pentium IV microprocessor chip will be 100 nm with a gate oxide thickness of 2 nm and a total of 28 million transistors per processor. The microelectronics industry has therefore become nanoelectronics named after the Greek for a dwarf "nanos." This article will review the silicon nanoelectronic field and discuss how far the silicon MOSFET can be scaled down. At some point this scaling must end and therefore for the continued growth of the electronics industry, an alternative technology must be found. The remainder of the article will review other nanoelectronic devices and circuits presently in the research field that may eventually be able to take over the transistor mantle or open up new electronic applications.

II. SILICON NANOELECTRONICS AND ULTIMATE CMOS

All semiconductors such as silicon are insulators until impurities are added to the system. Silicon has four outer shell electrons that form tetrahedral bonds to four other silicon atoms in a crystal, and therefore all the electrons are bound and the material is insulating. If a group V atom such as phosphorus is added with five electrons in the outer shell, then there is an additional electron after four electrons are used to bond to the surrounding silicon atoms, which allows electronic conduction when a voltage is applied. The silicon is said to be n-type as the conduction is through electrons. If a group III atom such as boron replaces a silicon atom, then the boron can bond to three silicon atoms in the lattice but there is a hole left where the fourth bond used to be. Therefore bonding electrons from the outer shell of other silicon atoms can move to fill this hole, and they leave a hole behind from where they came from in the crystal. Therefore the silicon can conduct as electrons move into holes in the system. This is p-type silicon. If p-type silicon is placed beside an n-type piece of silicon, then a depletion layer forms where electrons and holes annihilate each other and electrons cannot flow between the two regions unless a large voltage is applied between them. Therefore a p-n junction stops electrons from being transported through the system.

The basic operation of a silicon MOSFET is shown in Fig. 1. It is a p-silicon substrate with an insulator (silicon dioxide) and a metal gate on top. It is basically two back-to-back p-n junctions so that with no gate voltage applied no current flows from the source n-type contact to the drain n-type contact. As a voltage is applied to the gate, electrons are attracted to the gate and form a thin layer near the surface of the p-type silicon. Electrons can now be transported from the source to the drain and a current flows between the source and drain. The gate therefore switches the current between the source and drain off and on.

Figure 2 shows the gate length and density on memory and logic chips found inside present computers along with predictions from the Semiconductor Industry Association (SIA) Roadmap for the next thirteen years. Already in 2001 dynamic random access memory chips (DRAM) have 270 million bits per cm^2. Each DRAM bit consists of a capacitor and a transistor with a gate length of 180 nm. The microprocessing units (MPU) such as Pentium and Power PC chips have 19.7 million transistors per cm^2 with gate lengths of 100 nm. Most of these transistors are in fact memory in the form of static random access memory (SRAM). A single SRAM cell consists of six transistors. Attempts to integrate the more compact DRAM memory cells onto MPU chips resulted in substantially poorer logic transistors, while the six SRAM transistors are identical to those used in logic and therefore are easy to integrate. The MPU chips use a complementary metal oxide semiconductor (CMOS) architecture that incorporates two MOSFET transistors, one that uses electrons as the conducting layer and a second that uses

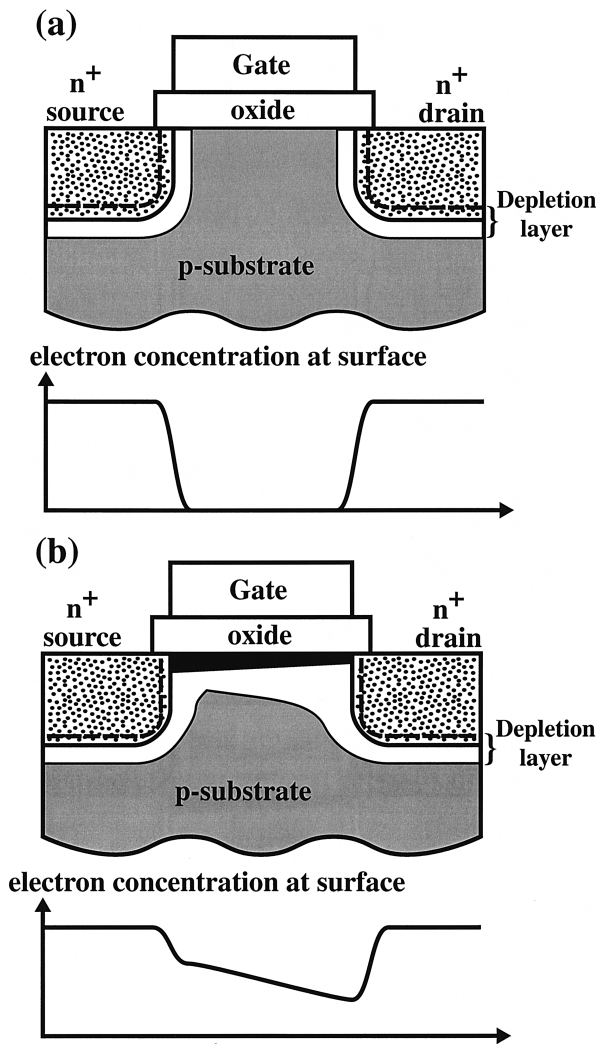


FIGURE 1 The operation of a FET. (a) The transistor switched off with no voltage applied to the gate. The p-region prevents electrons from traveling from the source to the drain. (b) The transistor switched on with a voltage applied to the gate which attracts electrons to the gate and therefore the electrons may travel from the source to the drain.

holes. The advantage of this architecture is that power is only dissipated (apart from that resulting from any leakage currents in the transistor) when the circuit is switched. This substantially reduces the power consumption of the chip. As the integration density of a chip is limited by the power dissipation, the ability to scale to smaller transistors aids the integration density from both the size and power requirements.

While the SIA Roadmap of Fig. 2 suggests that CMOS MPUs may be scaled to 20-nm transistor gate lengths in 2014, the SIA Roadmap also contains a list of requirements that have to be met to allow the scaling of silicon to continue to this level. Many of these requirements have no present or known solutions, and therefore the ability to scale CMOS to 20 nm is not guaranteed. In the scaling process, all the parts of the transistor are reduced to try and keep all the electric fields in the device constant. In reality this cannot be achieved in short channel devices and the electric field is increased, resulting in a number of problems, the most significant being drain-induced barrier lowering (DIBL). The p-region in Fig. 1 produces a barrier between the source and drain n-type contacts. If the gate length is reduced, then the source and drain n-regions deplete out carriers in the p-type region and the barrier between the source and drain is reduced, i.e., drain induced barrier lowering. The way around this is to increase the p-type dopant density under the gate, but this can only be achieved up to a certain level. Eventually electrons can quantum mechanically tunnel through the barrier between the source and drain when they are close enough together. Therefore the transistor can never be switched off and ceases to be a switch. NEC has, however, demonstrated a transistor that has a gate length of only 8 nm. This demonstrates that transistors can be made to operate down to this gate length, although due to the DIBL the gain or amplification of the transistor is much lower than that predicted by direct extrapolation from larger gatelength transistors.

A second problem is that electrons can quantum mechanically tunnel through the gate oxide when it becomes too thin. Therefore the oxide can only be scaled down to about 0.8-nm thick. Since transistors in 2001 already have a 2-nm thick oxide, this value is very close. One potential way around this problem is to find a new material for the gate oxide with a much higher dielectric

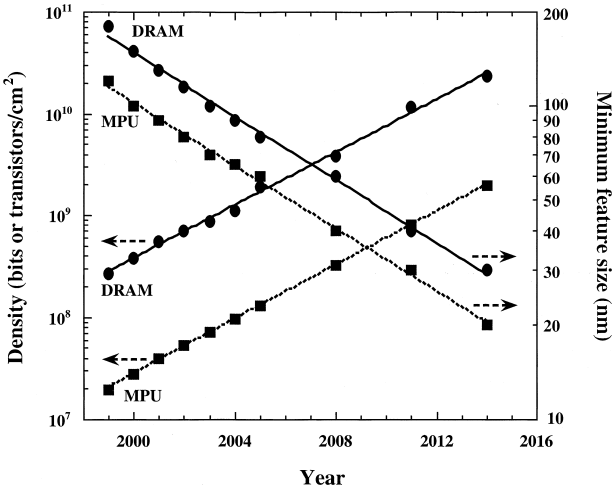


FIGURE 2 Moore's law for the future as predicted from the SIA 2000 Roadmap. DRAM is dynamic random access memory used in a computer and MPU is microprocessing unit—the central processor of a computer.

constant. A number of materials including Ta_2O_5 are being researched, but to date the results are not adequate for manufacture.

A third major problem is that the fabrication process uses optical lithography, which is a parallel process. The Rayleigh resolution criteria for optics states that the linewidth that can be achieved is proportional to the wavelength of the light and inversely proportional to the numerical aperture of the lens in the optical system. For the 100-nm gate-length transistors, a wavelength of light of 248 nm is used along with photosensitive resists that become more soluble when exposed to photons and can be selectively dissolved away. By shining the light through a mask with transparent and opaque regions, the pattern on the mask can be replicated in the resist and this can be etched into the underlying material. There are only a limited number of known radiation sources with wavelengths below 248 nm, and the appropriate lens and resists either do not exist or require substantial development before they can be used. While electrons can be used, almost all electron beam lithography systems are serial in nature and therefore much slower that the parallel lithographic processes. The serial processes therefore cannot be used on the present 300 mm wafers for manufacture. The radiation at these wavelengths is close to X-rays and therefore the photons are highly energetic. The lenses in the systems cost about $10 million (United States dollars) and must be replaced annually as the energetic radiation damages the lens materials. The reduction in wavelength is not guaranteed. One possible method is to use phase-shifting technology. By incorporating sections on the mask that rotate the phase of the photons, the light interferes and it is possible to beat the Rayleigh resolution criteria. Intel have recently demonstrated 30-nm gate length transistors and CMOS SRAM cells using such phase-shifting technology with 248-nm wavelength resists and optical lithography systems. This at least demonstrates that optical lithography with tricks can produce the required linewidths to at least 2011, although other problems such as layer overlay and mask testing (metrology) have yet to be resolved.

A corollary to Moore's law is that to increase the yield and therefore reduce the cost per transistor on a chip by scaling down the transistor size, the cost of the semiconductor fabrication equipment doubles every four years. While the cost of a DRAM bit is now 7.6 microcent/bit and of a transistor on an MPU is now 123 microcent/transistor, the cost of semiconductor fabrication plants announced at the start of 2001 is about $2.5 US billion (Fig. 3). At some point these plants may become too expensive and no return for the investment can be achieved by sales. Many in the semiconductor industry therefore predict that the economics will end Moore's law before any physical limitation does (Fig. 3).

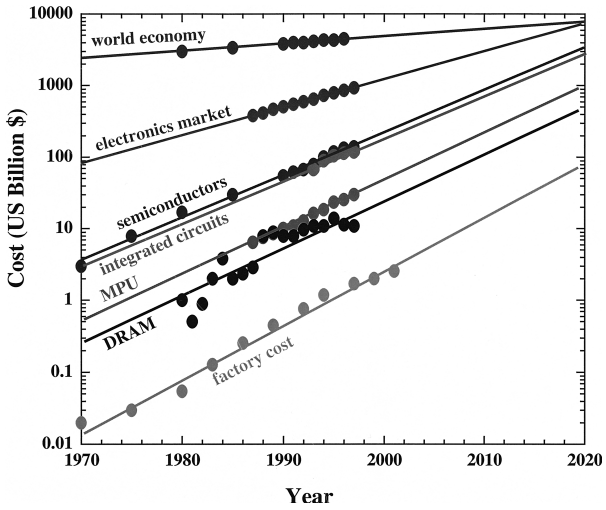


FIGURE 3 The economics of the semiconductor industry with data taken from Semiconductor International.

III. SINGLE ELECTRON DEVICES

While CMOS transistors now have 100-nm gate lengths, the number of electrons used in a switching operation are still tens of thousands. If this could be reduced to a situation where only one electron is used, then the energy required to switch should be much lower. This is the basic philosophy of single electron transistors (SETs). There are a number of different types of SETs. The original type relies on Coulomb blockade while a second type is literally miniature flash memory where the addition of a single electron to the gate-memory node results in a large change to the current in the measuring transistor.

A. Coulomb Blockade

Coulomb blockade involves a small island of charge situation between two electrodes. If the island is small enough and has N electrons, then an energy gap opens up between the energy of the last (Nth) electron and the first empty electron state ($N+1$). This gap equals the square of the electron charge e^2, divided by the capacitance of the island C. Therefore if the island is small enough so that this energy gap is larger than the thermal energy in the system (k_BT where k_B is Boltzmann's constant and T is the temperature), then electrons cannot quantum mechanically tunnel through the system since the only free states that electrons may tunnel onto the island are above the energy of the electrons in the electrodes (Fig. 4). If, however, a gate is used to electrostatically move the islands energy states with respect to the electrodes then the $N+1$th electron free state can be moved below that of the left electrode and electrons can quantum mechanically tunnel through the island one at a time [Fig. 4(b)], hence the name single

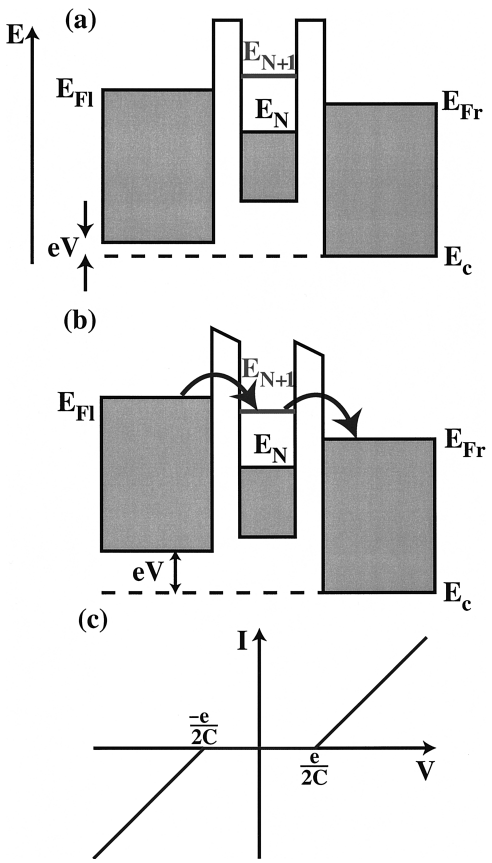


FIGURE 4 (a) The Coulomb blockade regime. An energy gap forms between the chemical potential E of the Nth and the $N+1$th electrons. (b) Single electron tunneling. (c) The current voltage characteristics.

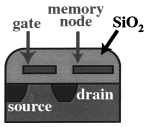
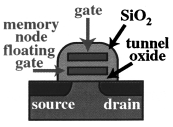
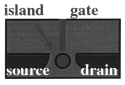
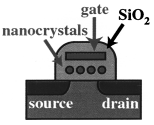
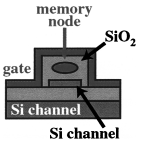
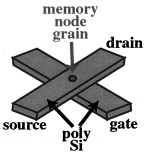
electron transistor. Therefore the current-voltage characteristics have zero current until the applied gate voltage corresponds to $\pm e/2C$ [Fig. 4(c)].

A large number of demonstrators of such transistors have been fabricated in a number of different material systems. For room temperature operation of such devices, the island must be below about 10 nm in diameter. Numerous SETs using silicon and silicon dioxide that operate at room temperature have been demonstrate. SETs are more likely to be used for memory applications because they have no gain (i.e., amplification). It is therefore difficult to create logic circuits where the gain in a transistor or logic device overcomes the losses in the circuit and interconnects (i.e., resistance). While a number of logic architectures have been proposed and a few have been demonstrated, it is questionable about the scalability of the circuits to the levels of present CMOS MPUs (Section 1).

B. Miniature Flash Memory

The second type of SET memory is really just a miniature version of the conventional CMOS flash memory found in mobile phones and Moving Picture Experts Group (MPEG) music stick players, for example. The structure is that shown in Table I, where the addition of a single electron to the memory node results in a substantial change to the electron current through the transistor channel. One potential problem of this approach is the robustness of the memory node to stray charge and fluctuations since one electron is enough to switch the device between memory states.

TABLE I A Comparison of the Performance of Different SET and Conventional Memory Technologies

	Conventional memory		Quantum dot memory			
				Nano-flash		
	DRAM	Flash	SET	Multidot	Single dot	Yano type
Device structure						
Read time	~6 ns	~6 ns	1 ns	~10 ns	~10 ns	~20 μs
Write time	~6 ns	1 ms	1 ns	~100 ns	<1 μs	~10 μs
Erase time	<1 ns	~1 ms	<1 ns	~1 ms	<1 ms	~10 μs
Retention time	250 ms	~10 years	~1 s	~1 week	~5 s	~1 day
Endurance cycles	Infinite	10^6	Infinite	10^9	10^9	10^7
Operating voltage	1.5 V	10 V	1 V	5 V	10 V	15 V
Voltage for state inversion	0.2 V	3.3 V	<0.1 V	0.65 V	0.1 V	0.5 V
Electron number to write bit	10^4	250	1 (Excluding no to change gate potential)	10^3	1 (Excluding no to change gate potential)	2 (Excluding no to change gate potential)
Cell size	8.5 F^2/bit	~9F^2/bit	9–12F^2/bit	9F^2/bit	9F^2/bit	2F^2/bit

A second approach to this concept is to use a number of Si nanocrystals as nodes in the oxide rather than one. This approach has the advantage that it is more robust to single electron fluctuations in the system.

C. Yano-Type Memory

The next type of SET memory is that demonstrated by Yano at the Hitachi Central Research Laboratories. It involves the fabrication using standard CMOS fabrication lines of two crossed poly-Si wires. The poly-Si consists of small grains of single-crystal silicon with grain boundaries between. This type of memory device uses the grains as memory nodes. One major problem is that conduction in the channel is the result of a percolation path through a large number of poly-Si grains. This is a random process and is difficult to control. Therefore the ability to mass produce such memory devices with the required control of properties may be the major problem with this approach. Hitachi has demonstrated a 128 Mbyte memory chip with the technology, although only half the devices operated.

D. Comparisons of Single Electron Devices

Table I summarizes the experimental results in SETs with the present production memory of DRAM and flash produced using CMOS processing lines. Some of the performance is comparable to present CMOS DRAM and flash, although noise because of the single electron nature is still a major barrier to manufacturable devices.

IV. QUANTUM MECHANICAL TUNNEL DEVICES

Quantum mechanical tunnel devices rely on the ability of electrons to tunnel through thin barriers when the electron wave function can penetrate through the barrier. The transmission coefficient T for an electron of energy E, impinging on a single barrier of height H and thickness b, is given by

$$T \approx \frac{16E}{H} e^{-2\alpha b}$$

The barrier thickness must therefore be below about 10 nm to allow an adequate amount of tunneling current. There are two main types of devices that use quantum mechanical tunneling to produce a negative differential resistance (NDR) current-voltage characteristic that may be used in a number of circuit applications.

A. Resonant Tunneling Diodes (RTDs)

The resonant tunneling diodes involve a device with two electrodes with two tunnel barriers between the electrodes (Fig. 5). The quantum well created by the confinement of

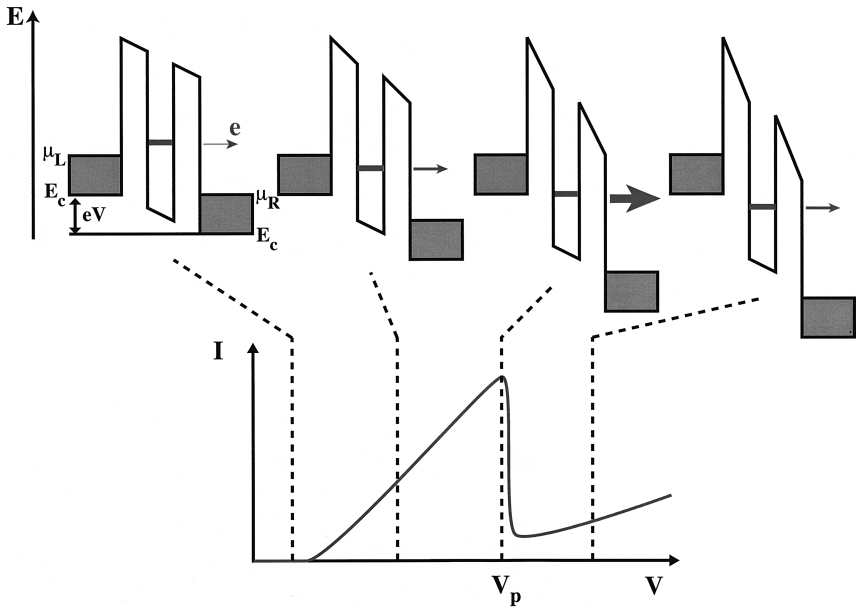


FIGURE 5 A schematic diagram of the operation of a resonant tunneling diode. The μ_R is the chemical potential for the right electrode and μ_L, for the left—electron fills states from the top of the conduction band E_c, to the chemical potential. The grey areas are filled electrons states in the heavily doped electrodes that provide reservoirs of electrons for tunneling. Only when the subband in the central quantum well has the same energy as an electron in one of the electrodes can electrons tunnel through the system to the other electrode.

TABLE II Comparisons of Different RTD Parameters and Predictions from Circuit Simulations of Required Device Parameters for Optimized Performance

Parameter	High-speed InP RTD logic	Predicted high-speed logic	Low-power InP RTD memory	Predicted low-power memory	Nanometer scaled RTDs	SiGe RTD	SiGe interband tunnel diode
Peak to valley current ratio	4	3	2	3	3	3.3	5.45
Peak current density (A/cm^2)	40 k	10 k	0.16	0.1	10 k	25 k	8 k
Minimum feature size	2 μm	200 nm	500 nm	200 nm	50 nm	14 μm	14 μm
Peak voltage	0.35	0.16	0.20	0.20	0.20	1.5	0.28
Maximum clocking frequency	12.5 GHz	6.25 GHz	592 kHz	56.8 MHz	6.25 GHz	No circuit	No circuit
RTD time constant	0.02 ns	0.04 ns	422 ns	4.4 ns	0.04 ns	$\sim$0.02 ns	$\sim$0.31 ns
Investigator	Dortmund 1998	Dortmund 1998	Raytheon 1998	Dortmund 1998	Dortmund 1998	Cambridge 2001	Stuttgart 2000

the electron wave function between the two barriers produces a discrete set of allowed electron energy states in the quantum well. Only when an electron from the electrode has an energy that corresponds to the allowed state in the quantum well can it quantum mechanically tunnel through the two barriers and quantum well, and reach the right electrode (Fig. 5). Therefore since each electrode has a reservoir of electron states above the conduction band but below the chemical potential in the electrode (μ_L for the left electrode and μ_R for the right), a peak forms in the current when the two are aligned after the application of a voltage between the electrodes V, which corresponds to lifting one electrode by eV with respect to the other electrode. Once a further voltage is applied, the electrons do not have allowed states in the quantum well to tunnel through and the current drops. If there is a second subband, then a second current peak will form when the reservoir of electrons has the same energy as the second subband.

The best results from RTD devices are in the III–V system and in particular the InP/InGaAs/InAlAs heterosystem, which includes a large number of circuit demonstrators. Silicon is, however, the dominant material in electronics and is substantially cheaper to manufacture. There has therefore been interest in reproducing the III–V based results in a silicon system. Silicon-based diodes have had much poorer peak current densities and peak-to-valley current ratios (the two most important parameters of a RTD for circuit applications), although recent Si/SiGe have improved to the stage of having good enough performance to be used in circuits. Typical RTD performance is summarized in Table II. The RTD is the fastest transit time device and an InAs/AlGaAs oscillator operating at 712 GHz with a 0.3 μW output power has been demonstrated. This power is at present too low for many applications, but there are very few other devices that can operate at this speed.

B. Esaki Diodes

While the RTD uses the one type of charge carrier (i.e., either electrons or holes), the Esaki diode is an interband device where electrons tunnel from the conduction band to hole states in the valence band. The operation is described in detail in Fig. 6. Esaki diodes have been around since Leo Esaki discovered and analyzed the effect in 1958 using Ge diodes. All the devices produced in the early 1960s were fabricated by diffusing one species of doping to create the p-i-n diode. This results in poor uniformity and poor control of the performance. Recent epitaxial growth and the inclusion of SiGe intrinsic layers has allowed much better performance results (Table II) and good uniformity.

C. Tunnel Diode Circuits

To date, tunnel diodes are one of the few quantum devices that has successfully demonstrated room temperature circuit operation. InP-based tunneling static RAM memory cells using two RTDs and one FET have demonstrated

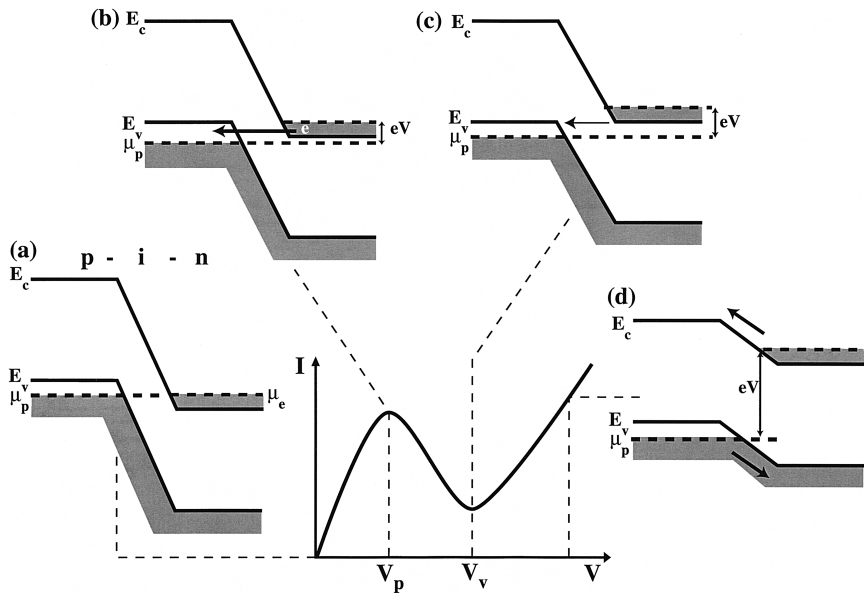


FIGURE 6 The operation of the Esaki interband diode which has p-i-n layers. The μ_p is the chemical potential in the p-region and μ_e is the chemical potential in the n-region. Grey areas correspond to states filled with electrons (in the valence band this corresponds to states without holes). (a) At zero bias electrons from the conduction band of the n-region cannot tunnel to the valence band because there are no hole states at the same energy. (b) A peak in the current voltage occurs when the reservoir of electrons in the conduction band of the n-region has the maximum overlap in energy with the reservoir of holes in the valence band. (c) The valley in the I-V occurs when the electron reservoir in the n-region becomes higher in energy than all the hole states in the valence band of the p-region. (d) Once the n- and p-states are misaligned, current can only flow due to second-order processes by scattering and tunneling through defect states and deep levels in the bandgap and thermally excited processes.

50 nW power dissipation with a 150-mm^2 cell footprint. This power dissipation is a factor of 200 lower than an equivalent six-transistor GaAs static RAM, which has a footprint three times larger. The RTD can therefore be used to reduce power consumption, circuit complexity, and the real estate usage on a chip.

In logic circuits there has also been a substantial number of demonstrators. Generic logic that can change functionality between NAND, NOR, or NOT just by changing the relative sizes of the transistors in the circuit has demonstrated 12-GHz circuit operation using 20-μm minimum feature sizes. As a comparison, CMOS requires 180-nm minimum features to achieve 1-GHz operation. Therefore the scaling of this technology to nanometer sizes will result in substantially faster performance than CMOS. Double current peak RTDs have also been used to demonstrate multivalued logic, which allows a substantial reduction in device count for logic chips. Monolithic 4-bit analogues to digital converters have shown 2 gigabit per second sampling and static binary frequency dividers operating at 40 GHz have all been demonstrated. There are also circuit demonstrators of photodetectors, clock circuits, and shift registers which all operate at far faster speeds than equivalent CMOS circuits and with reduced component counts.

RTD devices have therefore be used to demonstrate a large number of circuits that either provide higher speed or lower power dissipation than conventional CMOS or III–V FET technologies. At present all these demonstrators are in III–V technology and no silicon RTD or Esaki diode circuits have been demonstrated. One major problem of the III–V system is that since the tunneling current depends exponentially on the barrier thickness, only a two-atom difference across a wafer can have a substantial change in the tunneling current. This limits the yield at present of the circuits and must be improved if RTD circuits are to compete with other technologies in the market. In addition for high frequency oscillations, the power must be increased to able to be used. One method that is being used is to build arrays of RTDs, and this may allow oscillators to operate above 600 GHz with mW of power.

V. SPIN NANOELECTRONICS OR SPINTRONICS

While most nanoelectronic devices rely on charges being transferred around a device or circuit, the electron spin may also be used for memory storage or logical computing. The hard disks in computers are already a good

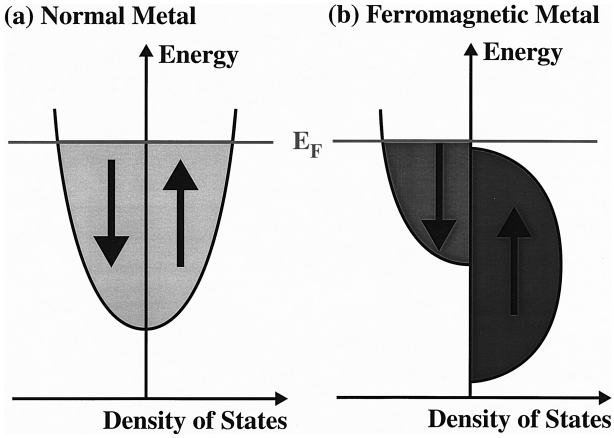


FIGURE 7 The difference between a metal and a ferromagnet. (a) In a metal, the number of up spins equals the number of down spins around the Fermi surface. (b) In a ferromagnet, there is a larger number of one type of spin around the Fermi surface.

example of how spins may be used to store information. Most new spintronic devices rely on spin injection rather than the measurement of a ferromagnetically polarized particle.

The basic operation of a spin device depends on the ferromagnetic material having a larger number of one spin close to the Fermi energy or surface as shown in Fig. 7. If the ferromagnet is put into a circuit, then only (e.g., from Fig. 7) the down spins can be transported through the system. Therefore switches can be built up by designing structures with different ferromagnetic and normal metallic layers.

Figure 8 shows the basic operation of a spin valve. By having two ferromagnetic layers placed on either side of a metal, if both ferromagnets are polarized in the same direction then the majority spins in the system will have a low resistance path through the device. If, however, the ferromagnets have opposite polarization, then both the majority and minority spins find a high resistance path. Therefore the resistance of the system corresponds to the on and off states required for switching. The application of a magnetic field provides giant magnetoresistance (GMR), which is now the basis of all the devices for reading hard disks.

A similar effect can be used to produce a spin tunnel device. If two ferromagnetic contacts are placed on either side of a metal (or semiconductor), then electrons of one particular spin will be injected into the metallic layer if a bias is applied to the system [Fig. 9(a)]. If the second contact to the metal also has the same spin polarization as the first, then the electrons can pass into the second contact and the circuit has a low resistance. If, however, the second contact is polarized opposite to that of the first

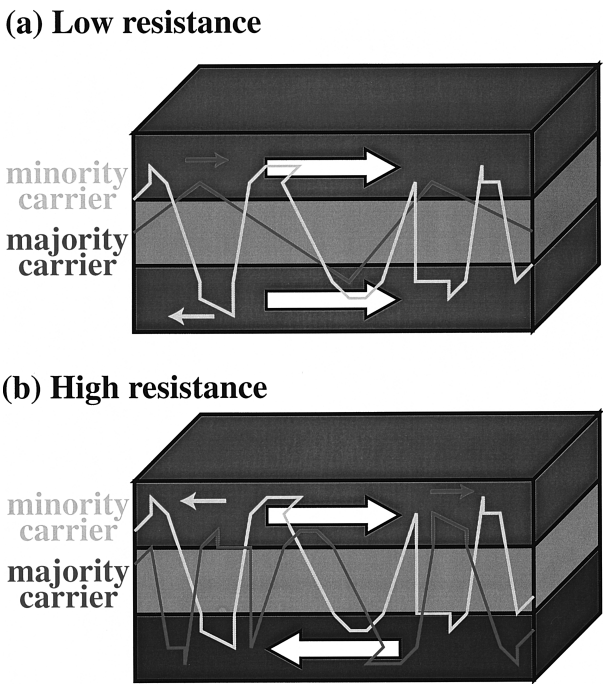


FIGURE 8 The concept of the spin valve switch. (a) If the two ferromagnetic layers have the spins aligned, then the majority-spin carrier has a low resistance path through the device. (b) If the ferromagnetic layers have opposite spins, then both spin carriers see a higher resistance path through the device.

contact [Fig. 9(b)], no free states exists for the electrons to tunnel into close to the Fermi surface and no current can flow in the system. If the metal is replaced by a semiconductor with a large Rashba term, then the application of an electric field to the semiconductor will rotate the spin of the electrons as they are transported across the semiconductor layer and the device can be switched on and off using the electric field from a gate. These effects can be used to form a number of different spintronic devices, including switches and spin-polarized filters.

The most successful magnetic nanoelectronic device to date is the magnetic random access memory (MRAM). This is shown in Fig. 10 and corresponds to a GMR element that is written and read using the magnetic fields generated by currents in the word and bit lines. It is therefore a very compact memory along with being nonvolatile and radiation hard. A substantial number of demonstrators have been produced with sizes above 1 MB and access times of about 50 ns using 0.25-μm dimensions. Such devices can be easily scaled to smaller sizes where the performance is predicted to increase substantially.

To date there have only been a few demonstrations of spin injection into semiconductors; in particular, a spin-polarized light emitting diode (LED) has been demonstrated where the output photons are polarized depending

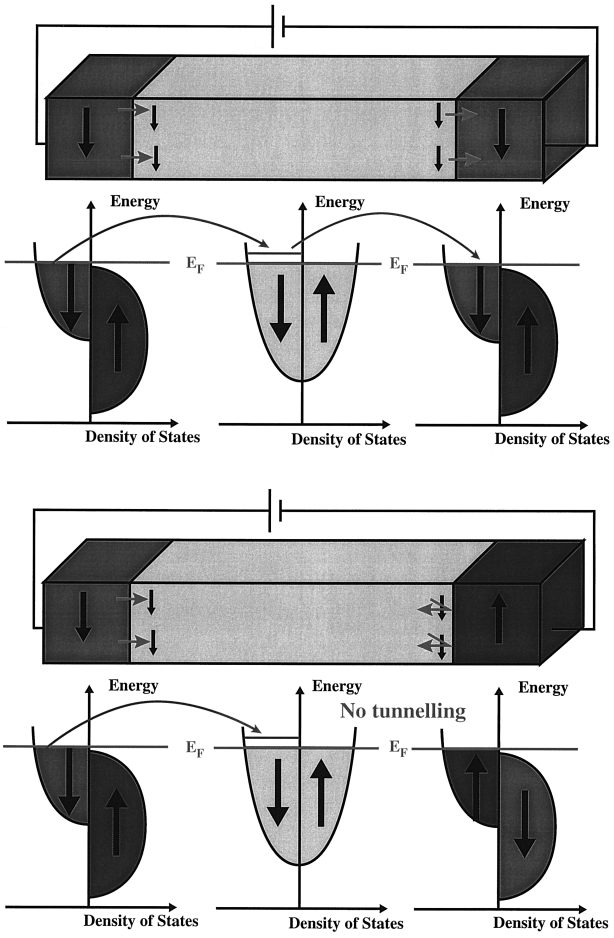


FIGURE 9 The basic principle behind a spin tunnel junction device. (a) If the two contacts to a metal or semiconductor have the same spin polarization then electrons can transfer around the circuit and a low resistance is measured. (b) If, however, the contacts have opposite spin polarization, then the electrons cannot enter the second contact as there are no free states near the Fermi energy.

on the spin of the electron and holes in the system. Significant spin injection can only be achieved from magnetic semiconductors into normal semiconductors. Results from metallic ferromagnets into semiconductors has to date been very poor.

Spintronics is therefore at a relatively mature stage of research, and potentially may allow both memory and logic functions in much smaller and robust systems than Si CMOS. In particular, the magnetic systems may be able to be designed so that the electrical resistance of small interconnects does not impede the ultimate performance.

VI. FABRICATION TECHNOLOGY

At present all CMOS wafers are fabricated using a top-down approach where deep ultraviolet photons are shone

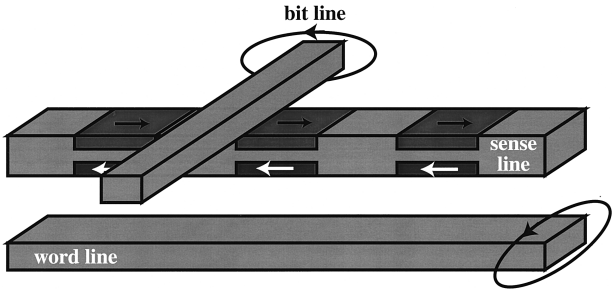


FIGURE 10 A schematic representation of a MRAM device where a GMR element is magnetized by the magnetic fields created by currents in the word and bit lines.

through a patterned mask made of glass (or quartz) and chrome. Where the light is transmitted through the mask, it reacts with an optical resist on the wafer, and this area may be dissolved in a solvent to leave the mask pattern in the resist. This pattern is then transferred to the underlying substrate by etching. By blanket deposition of insulators or metallic layers along with lithography and etching, complex circuits are patterned onto chips.

The smallest feature that may be produced using this lithographic technique is given by the Rayleigh resolution criteria, $k\lambda/NA$, where k is a constant, λ is the wavelength of light, and NA is the numerical aperture of the lens in the system. At present 248 nm KrF excimer laser sources are predominantly used and 197 nm ArF lasers are available. To produce smaller features, the wavelength must be reduced and the numerical aperture of the lens in the optical system must be increased. By using phase-shifting technology (i.e., parts of the mask also rotate the polarization of the light), interference effects can be achieved that produce smaller features than the Rayleigh criteria. The major problem is that such phase shifting technology is extremely complicated to design and difficult to fabricate the appropriate masks.

Other lithographic systems at the research stages include extreme ultraviolet sources, X-ray, and electron beams. In all cases these systems use smaller wavelengths to increase resolution. It must be stated that sources, masks, optics, and resists are all required for a lithographic fabrication process. These are not always available; for instance, X-ray optics are extremely poor at present. Another problem is that the speed of exposure must be high enough to allow mass production (Fig. 11). All early electron beam lithography systems used a single focused beam with a spot of a few nanometers. The pattern is then built up by rastering the beam over the chip area, but for large complex chips this is very slow and impracticable. Vector beam systems have been developed where a much larger standard feature (e.g., a rectangle or a repeated feature) can be exposed at one time, substantially increasing the throughput.

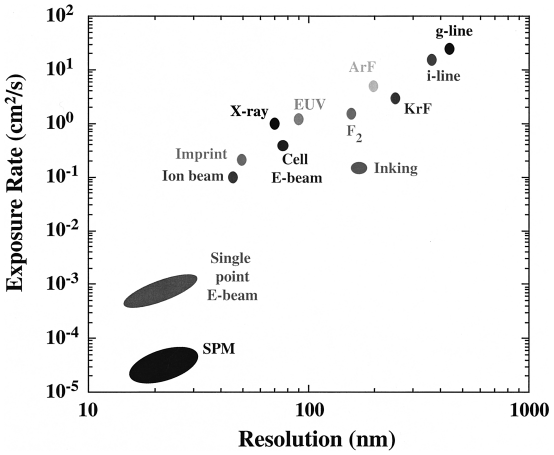


FIGURE 11 A comparison of the exposure rate and hence throughput of different lithographic techniques compared to the resolution of the techniques. The g-line, i-line, KrF, ArF, and F_2 are all sources for optical lithography. EUV is extreme ultraviolet.

Imprint lithography is a recent development where a master mold is imprinting into a polymer resist through pressure and sometimes additional heating. When the mold is removed the polymer typically has thick and thin parts in the appropriate pattern that, after a descum to remove the thin parts of the polymer, may be transferred to the underlying substrate using etching. This technique can potentially be applied on a wafer scale and has demonstrate small features, the present limit of which appears to be more related to the ability to fabricate small features on the master mold rather than through the imprint technique itself. Features as small as 10 nm have been demonstrated, although for high throughput, features are typically 50 nm or above (Fig. 11). The major problem with such a technique is that it is difficult to align and calibrate the mold to different layers. As a first stage lithographic process this is adequate but for large-scale nanofabrication, many different layers are required. The optical techniques easily allow scaling of the pattern to match to other layers through focusing, but a mold is fixed. An additional potential problem is that the mechanical contact is prone to dirt and defects, and may be damaged over time. A derivative of the imprint technique is inking, where the polymer or ink is applied to the mold so that only raised features have any ink on them. The pattern is then transferred to the substrate through contact. These techniques show good promise to scaling to features of tens of nanometers but may ultimately be limited by their inability to be flexible to other layers in real structures.

One of the major problems of these top-down lithographic approaches is that they require very clean environments, otherwise dust and particulates can mask part of the exposed area. It is the lithography that drives the cleanliness aspects in semiconductor production. For high yields and hence low costs, clean rooms must have particulate densities extremely low and at sizes much smaller than the lithographic minimum feature size. The cost of lithography is also increasing at an exponential rate as the feature size decreases and may eventually stop the progress of microelectronics.

VII. SELF-ASSEMBLY

The top-down approach of lithography described in the previous section is one of the expensive stages in fabrication of CMOS circuits and requires high levels of cleanliness to produce high yields. An alternative concept is to use a bottom-up technique where the devices self-assemble into circuits. This is at a relatively early stage of research but if successful should be substantially cheaper and more robust to "dirty" environments than lithographical techniques.

Much of the early self-assembly was in the area of metallic and semiconductor dots (small particles), which have improved optoelectronic properties for detectors, LEDs, and lasers. Most of these systems are not appropriate for electronic systems as no interconnects are available. In the last ten years a number of biological, organic, and inorganic systems have been shown to be able to either self-assemble or to be used as a template for other functional groups.

One of the best examples was the use of DNA to selectively attach conducting particles (in the first instance, silver particles). Since the DNA helix is held together by base pairs made up of 4 different bases, each of which can only bond to a reciprocal base through the Watson–Crick base-pairing interactions, single-stranded or double helix DNA can be designed to either selectively attach to other DNA molecules or selectively attach functional units (Fig. 12). An adenine (A) on one strand pairs preferentially with a thymine (T) on another strand while a guanine (G) pairs with cystosine (C). Single strands of DNA at the end of a double helix with a specific set of bases called a sticky end allows this end to only attach to another sticky end with the appropriate bases, thereby potentially allowing small, functional, hybrid DNA molecules to be self-assembled into much larger circuits (Fig. 13).

Numerous other examples of self-assembly exists including Langmuir–Bloggett films, S-bacterial layers, self-assembled monolayers (SAMs), and antibody–antigen recognition. All these techniques are still at a very early stage, and substantial research is required to get any of the techniques to a manufacturing level.

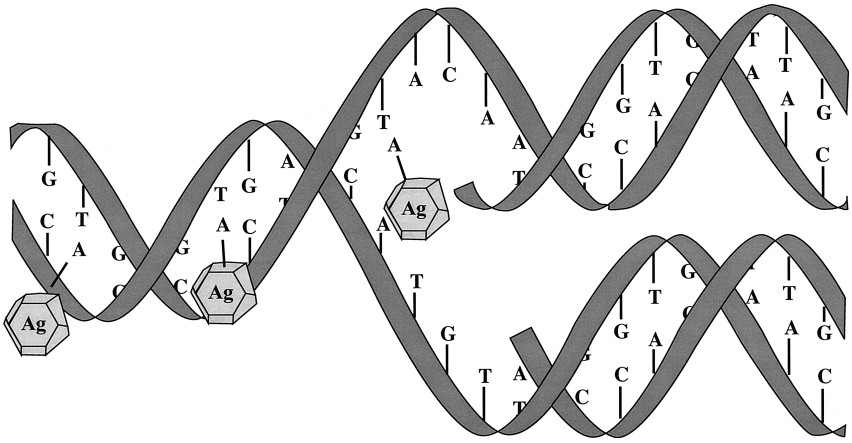


FIGURE 12 A schematic diagram of DNA templating where silver particles with an attached adenine (A) selectively bonds to thymine (T). Guanine (G) bonds to cytosine (C).

VIII. MOLECULAR NANOELECTRONICS

The first discussion of molecular nanoelectronics was the proposal by Aviram and Ratner in 1974 to produce a rectifier from organic molecules. The first example of a single molecular electronic device did not appear until 1990, the major problem relating to the difficulty of making individual electrical contacts to molecules which may only be a few nm in size. The development of the scanning tunneling microscope (STM) basically enabled the first measurements in this field to be started and has remained one of the major tools in electrically characterising single molecules.

Some of the first demonstrations of electronic properties of single molecules by Purdue University included Coulomb blockade and Coulomb staircase when a STM tip measured the conduction through gold nanoparticles self-assembled with SAMs. A second set of experiments at IBM, Zurich, demonstrated a STM tip deforming a C_{60} bucky ball. The resulting mechanical deformation modifies the resonance tunneling bands of the molecule and produces electromechanical amplification. Hitachi demonstrated a molecular abacus where a STM tip was used to move 0.25 nm high C_{60} molecules along monoatomic steps and then counted by imaging with the STM tip. While all these devices demonstrate functionality that may be used in circuits, the scaling to the level of present CMOS circuits would be impossible.

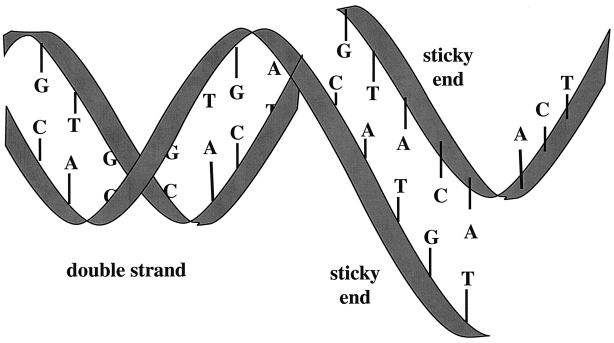


FIGURE 13 The ideal behind sticky ends that may be used to selectively attach functional DNA double or single strands to other DNA or substrates or electrodes with appropriate sticky ends.

Another approach to molecular electronics is the use of organic molecules. A collaboration between Yale and South Carolina Universities demonstrated the conduction through a benzene molecule attached to two gold electrodes using thiol groups to bind the molecule to the gold. Benezene rings have delocalized Π-electrons out of the plane of the molecule through which electrons can be transported when an appropriate bias is applied across the molecule. Carbon–carbon double and triple bonds provide similar orbitals out of the plane and therefore combinations of these polyphenylene molecules create conducting wires now known as Tour wires. Molecules could then be designed with conducting sections along with barriers created by methyl groups that do not have delocalised electron orbitals. Therefore organic chemists can design molecules where the high occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) through which electronic transport is controlled can be manipulated in a similar fashion to the band structure engineering in semiconductors (Fig. 14). Molecular RTDs operating at room temperature have been demonstrated. The low temperature properties of some of the RTD designs demonstrate PVCRs of over 1000 at low temperatures while the room temperature properties require improvement before they can be used in circuits.

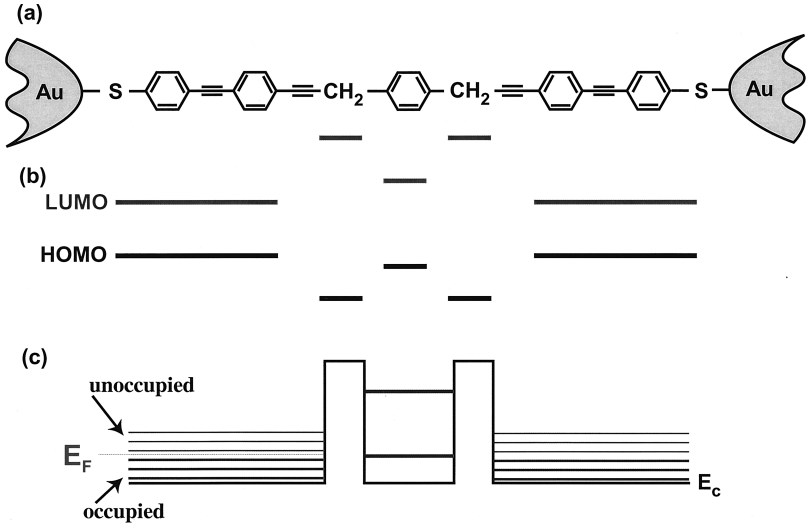


FIGURE 14 (a) The chemical structure of an organic RTD. (b) The HOMO and LUMO levels responsible for the electronic transport through the molecule. (c) The equivalent band structure for a semiconductor RTD.

Work at the Mitre Corporation has investigated possible architectures using organic molecules. A number of designs have been proposed based on diode logic using the Tour wires and diodes. AND, OR, and XOR gate designs are given along with an adder. The major problem with such organic systems is that the conductivity is relatively poor through the interconnecting Tour wires. The RC time constants of most of the devices is likely to be relatively large and some rough "back-of-the-envelope" calculations suggest that typical speeds will be up to tens of megahertz rather than gigahertz as presently available with CMOS. The major limitation is the conductivity, and unless better conductors or architectures for which the performance does not depend on resistance can be found, the organic systems will always be much slower than silicon.

Carbon nanotubes have also been used to conduct electrons and demonstrate reasonable conductivities for their size. A large number of groups have published results in the field. Both single and multiple walled tubes can be created that may potentially reduce resistance. Conductivities as high as 2000 Sm^{-1} that corresponds to a resistance per millimeter of 200 Ω have been measured, substantially better than organic polyphenylene molecules. A transistor was demonstrated, although the gate was a silicon substrate and the carbon nanotube had been placed across two metal electrodes fabricated on top of the thermally oxidized silicon substrate. Metal-nanotube rectifiers have also been demonstrated. To date, no switch or three-terminal device has been produced that is a basic requirement for most logical architectures.

Numerous groups have demonstrated measurements on nanocrystals; many have used cadmium selenide. CdSe nanocrystals can be prepared with sizes down to about 2 nm and with attached linker molecules that may bind to numerous surfaces including gold. A single electron transistor has been demonstrated where in a similar experiment to the carbon nanotube transistor, an oxidized silicon substrate is used as a gate to a CdSe nanocrystal placed between two gold electrodes on the SiO_2 surface.

DNA self-assembly was discussed in the previous section. There have also been a substantial number of papers measuring the conductivity of different types of DNA. Results range from highly insulating behavior to semiconducting and even metallic behavior. The results come from many different structures and types of DNA, and as yet no systematic investigation of varying specific parameters and the resulting changes in conductivity has been present. It would appear that the only consistent results are from DNA with self-assembled metallic nanoparticles. The best results to date correspond to Pd nanoparticles attached to a DNA strand, which has demonstrated 100 Sm^{-1} conductivity.

There are a large number of additional ideas presently around in the literature. To date, however, there has not been any demonstration of an all molecular transistor or any structure with gain. As will be discussed in the next section, gain is crucial to the distribution of information over large circuits. For a switch or transistor, a voltage corresponding to the difference in HOMO and LUMO energies will be required to switch conduction on or off. In most of the proposed devices this corresponds to about 1 eV, and hence 1 V must be applied to switch a device on or off. If the resistance of the molecule is about 200 Ω

and there is 10^{10} molecules, then the circuit will consume 10^8 W! Molecular nanoelectronics is at a very early stage and potentially offers very cheap self-assembly fabrication routes, but a substantial break through is required if the technology is ever to come to fruition.

IX. ALTERNATIVE ARCHITECTURES

A. Fault Tolerant

The main reason for the high cost of CMOS fabrication plants is that the top-down architecture required from lithographic fabrication techniques is not defect tolerant. Therefore any transistor or interconnect that fails on the circuit potentially destroys the whole chip. Some redundancy can be built in, but this is expensive and not ideal. Almost all nanoelectronic devices still rely on architectures that have no fault tolerance. This problem is substantially worse for applications specific integrated circuits (ASICS), where each application has a different circuit design, unlike the mass repetition of memory or processor manufacture. Testing has become a substantial cost (>60% of the total chip cost) as the new circuit design must be tested to check that no mistakes in circuit design or mask fabrication has taken place. With up to 20 mask levels in some chips and with up to 40 million transistors in designs, the potential for errors at some point in the design and fabrication process is substantial.

If an architecture could be found for which not all the transistors or interconnects are required to correctly function for the operation of a complete chip then this would substantially reduce costs. Most molecular ideas may require a fault-tolerant architecture if they are ever going to be successful.

One of the few defect tolerant architectures that has appeared is the Teramac architecture. This concept is shown in Fig. 15 and depends on the ability to have a large number of interconnects in a system so that some path may always be found around defects and nonfunctional parts of a circuit. Before the system is run, a map of all the defects is found and the system is configured so that the defects can be circumvented. Therefore the Teramac paradigm is to build a cheap and defective computer, find the defects, configure the resources with software, compile the program, and then run the computer.

The major disadvantage with the Teramac concept is that the biggest problem with both CMOS and molecules is the interconnects. With CMOS as interconnects are reduced in size, the conductivity is constant with the resistance increasing faster than the decrease in capacitance from reduced size. Therefore the RC time constants increase for reduced interconnect size and the speed of the circuit correspondingly is reduced. High quality interconnects are a serious problem in many of the molecular schemes. Therefore using larger interconnect bandwidth to provide defect tolerance is not an option with most of the known nanoelectronic schemes.

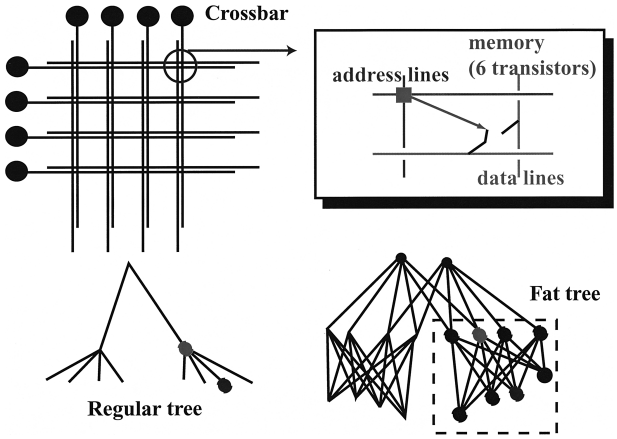


FIGURE 15 Schematic diagrams of the Teramac concept. Normal architectures use a regular tree structure where there is only one or a few paths between any points in the circuit. The Teramac architecture supplies much larger levels of connectivity through a fat tree structure so that if one interconnect fails, another can be used. The overall structure employs a crossbar addressing scheme with numerous connections to each memory node.

B. Gainless Architectures

To date, no molecular transistor with gain has been demonstrated. In single electron transistors, for every electron put into the transistor one gets almost one electron out of the transistor. These devices therefore have no gain. Gain serves two major purposes in architectures. First it provides isolation between the input and output. Therefore if one has some voltage at the output then that is the answer from the inputted data. If there is no isolation, then the logic gate could potentially run backward so that the output could be defining the input and therefore the whole logic gate oscillates between different values and any output cannot be believed.

The second major reason for gain in a circuit architecture is related to transmitting information around a large circuit. All electrical interconnects have a resistance that corresponds to a loss in the system. Therefore if you send a small current across a large circuit, it will only reach its destination if the losses are small. The way around this is to amplify the signal and that requires gain. Many circuits also connect one output from a logic gate to a number of new inputs to many logic gates (fan-out as it is known). The problem with a gainless system is obvious if one considers

the case of one electron exiting the first logic gate output. It can only be transmitted through one of the new logic gates. Therefore for fan-out in any circuit, a large signal is required that will be divided between a number of paths in the circuit. Since this is repeated many times in large logic circuits, there must be gain if any signal is to pass through the large number of gates.

For all the known architectures that are used in electronics, gain is therefore an essential requirement.

C. Quantum Cellular Automata

The quantum cellular automata (QCA) provides an interconnectless method of information exchange and computation. It therefore has gained a substantial amount of interest. The basic concept depends on four (or five) quantum dots being produced with only two electrons in the system (Fig. 16). The Coulomb repulsion of the electrons forces them to occupy diagonally opposite sites and therefore the arrangement dictates whether the cell is in a "1" or "0" state. As long as the electrons are forced to remain in this cell, then other cells can be placed beside the first cell. If a field is applied to change the arrangement of the electrons in the first cell, then this information is passed to neighbor cells and therefore is transported through the system.

By appropriate arrangements, logic calculations can also be achieved. Figure 17 shows a majority voting gate where the output of the gate equals the value of at least two of the inputs that have the same value. This sort of gate will only function with an odd number of inputs so that either "1" or "0" can dominate. The major problem with the QCA architecture can be easily demonstrated when a similar logic gate is constructed with an even number of inputs, half of which are in a "1" state and the remainder in the opposite "0" state. The output is therefore not predictable and the architecture breaks down. The second problem with QCA is that it is also a gainless system, and therefore if another cell applies a force to the output then the system may run backward, and the whole circuit becomes unstable and can oscillate between values.

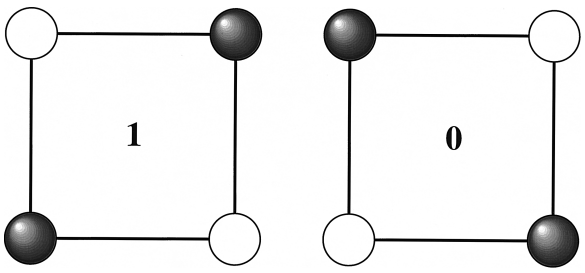


FIGURE 16 The two stable states of a quantum cellular atomata cell providing the "1" and "0" states.

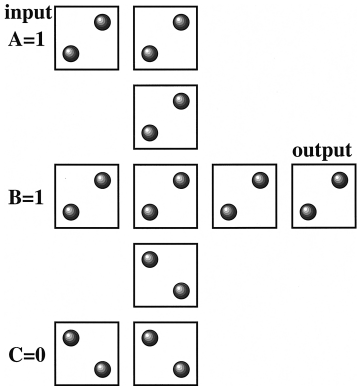


FIGURE 17 An example of a QCA logic gate. This is a majority voting gate where the output will equal the majority of the inputs, in the case above the output = 1.

So far only a single cell has been demonstrated experimentally and no logic gates have been demonstrated. Calculations suggest that Si quantum dots of less than 2 nm size are required to produce cells that may operate at room temperature. There is some belief that molecular self-assembled quantum dot systems may be the ideal system to produce QCAs, but much work is required. The QCA therefore is an interesting concept as it is an interconnectless architecture, but substantial problems suggest that it may not be realizable. Calculations at University College London have also shown that it may be much slower than conventional CMOS circuits at relatively small integration densities.

D. Quantum Computing

Quantum computing is a radically different architecture that uses the interference properties of entangle quantum mechanical particles to allow each bit of quantum information (called a qubit) in the computer to be intimately linked to every other qubit in the system. Quantum computing is therefore a massively parallel architecture at a level that is impossible in any classical architecture.

The important pieces of quantum mechanics for quantum computing is the Einstein–Podolsky–Rosen (EPR) paradox and the Bell inequality. If two particles are entangled and taken to the opposite ends of the universe, then quantum theory determines that if one measures the properties of one, you automatically then know the properties of the other since they are intimately linked. An example is the spin of a photon where if one has a right-hand polarization then the other after entanglement must have a left-hand polarization through quantum theory. Moreover, the state of one after a measurement automatically defines the state of the other. Before the measurement it is impossible to predict exactly which state each particle may be

in. This is as if information has traveled instantaneously across the enormous distance of the universe, which is impossible from Einstein's special relatively theory.

A bit is a two-state system from a physical point of view. In quantum mechanics, if a bit has two distinguishable states, then it may also exist in a superposition of these two states called a qubit. If each of these superpositions are entangled so that there is a further superposition of each qubit, then the state of each qubit depends on the states of every other qubit. Gate operations on these qubits provide a time evolution of the system where every entangle particle is affected and so a massive parallel computation results. This enormous parallelism allows classically numerically intensive problems such as the traveling salesman problem (known as NP problems in mathematics) to be solved in real time that cannot be solved with 100 years of present supercomputing time. Potential applications are factorization for cryptography, fast database searching, and modeling of quantum systems.

To produce a quantum computer, there are five criteria called the DiVincenzo checklist that must be adhered to:

1. A scalable physical system with well characterized qubits.
2. Initialization of simple fiducial states. This basically is setting the qubits into similar quantum states before any calculation can proceed. An example is setting all the spins in one system to be identical.
3. Long relevant decoherence times, much longer than the gate operation. As the decoherence time dictates the length of time the qubits can be entangled without loss of any information, any computation must be finished before the qubits loose information.
4. A universal set of quantum gates. For a quantum computer, two types of gate operation are required to produce a universal computer that may be designed and programmed to complete any computation task. These are the single- and two-qubit operations, and from these a computer for factorization, database searching or quantum system simulation may be built and operated. These two-gate operations can be achieved by different techniques that depend on the specific two-state system used for each qubit.
5. A qubit-specific measurement capability. Once a calculation is complete, the information must be outputed from the qubits.

At present only a few systems with up to 5 or so qubits have been demonstrated. For most applications such as factorization of large numbers for cryptography, 30,000 qubits or so are required, although even 50 qubits will solve problems quickly that are impossibly slow on classical computers. Ion traps and nuclear magnetic resonance systems have so far demonstrated the largest number of qubit entanglements. For large integration, however, solid state quantum computers potential provide the best platforms. Examples for the two-state systems required for qubits include semiconductor quantum dots, superconducting device, the nuclear spins of donors in silicon, the electron spins of donors in SiGe heterostructures, and the use of surface acoustic waves with low dimensional structures in GaAs heterostructures.

Quantum computing is presently at an immature experimental stage but if many of the problems can be overcome, the potential power of the technique will allow many numerically intensive calculations to be completed that are presently impossible with classical computational machines.

X. SUMMARY AND FUTURE PROSPECTS

The major conclusion of any review of nanoelectronics must be that the silicon transistor through CMOS is the dominant technology and will remain so for the foreseeable future. Only a break in the relentless pursuit of Moore's law in silicon may provide a chance for other technologies to try to compete. Such a break is more likely to be the result of economics rather than technological problems.

RTDs and single electron memories have serious possibilities of niche markets, and at least the RTDs are already being sold for specialist high-speed digital to analogue converter applications. The spin and magnetic devices, in particular MRAM, have very clear markets and are already close to production. Molecules and self-assembly are much further from the marketplace, and with present results, substantial breakthroughs are required if they are ever going to reach the marketplace.

A second significant conclusion of any nanoelectronics review will always be that the architectures for nanoelectronics are all the same ones used for CMOS, and new architectures are required for most new nanoelectronic concepts and devices if they are to succeed. Quantum computing is a very new architecture but it only provides niche applications and is not applicable to most nanoelectronic systems. It is also a field with much theory but little experimental demonstration to date. Much of the technology for quantum computing is way beyond present fabrication technology but this may change as the nanoelectronics research field progresses. The realization of a successful fault-tolerant architecture would bring enormous benefits to all nanoelectronic fields including CMOS and molecular electronics. The research field is therefore ripe for major advances in the future.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ARCHITECTURE • ELECTRON SPIN RESONANCE • FIELD-EFFECT TRANSISTORS • INTEGRATED CIRCUIT MANUFACTURE • MOLECULAR ELECTRONICS

BIBLIOGRAPHY

Bennett, C. H., and DiVincenzo, D. P. (2000). "Quantum information and computation," *Nature* **404:** 247–255.

Compañó, R., Molenkamp, L., and Paul, D. J., Eds. (1999). "Technology Roadmap for Nanoelectronics," European Commission, Brussels (http://www.cordis.lu/esprit/ist/fethome.htm).

de Boeck, J., and Gorghs, G. (1999, April). "Magnetoelectronics," *Physics World* **12**(4), 27–32.

Ellenbogen, J. C., and Love, J. C. (1999). "Architectures for molecular electronic computers: 1. Logic structures and an adder built from molecular electronic diodes," *Proc. IEEE* **87,** 386–426.

Heath, J. R., Kuekes, P. J., Snider, G. S., and Williams, R. S. (1998). "A defect-tolerant computer architecture: opportunities for nanotechnology," *Science* **280,** 1716–1721.

Lent, C. S., and Tougaw, P. D. (1997). "A device architecture for computing with quantum dots," *Proc. IEEE* **85,** 541–557.

Reed, M. A. (1999). "Molecular-scale electronics," *Proc. IEEE* **87,** 652–658.

Semiconductor Industry Association. (2000). "International Technology Roadmap for Semiconductors 2000 Update," SEMATECH, CA.

van der Wagt, J. P. A. (1999). "Tunneling-Based SRAM," *Proc. IEEE* **87,** 571–595.

Yano, K., Ishii, T., Sano, T., Mine, T., Murai, F., Hashimoto, T., Kobayashi, T., Kure, T., and Seki, K. (1999). "Single-electron memory for giga-to-tera bit storage," *Proc. IEEE* **87,** 633–651.



Semiconductor Alloys

T. S. Kuan
State University of New York, Albany

I. Long-Range Order in III–V Ternary Alloys
II. Origin and Stability of Ordered Phases
III. Optical and Electrical Properties of Ordered Phases

GLOSSARY

Epitaxial growth Continuing growth of a crystal where the placement of new atoms is influenced by the lattice structure of the crystal.

Lattice Orderly arrangement of atomic sites within a crystal.

SEMICONDUCTOR ALLOYS made of elements from column III (Al, Ga, In) and column V (P, As, Sb) of the periodic table are used extensively to fabricate heterostructure optoelectronic or high-speed devices. These III–V-type alloys in various binary (AB), ternary ($A_xB_{1-x}C$), and quaternary ($A_xB_{1-x}C_yD_{1-y}$) solid solution systems offer a rich variety of energy band gaps, which can be exploited in many heterojunction device applications. These III–V semiconductor alloys are produced in thin-layer materials grown on single crystal substrates by metal-organic vapor-phase epitaxy (MOVPE) or molecular beam epitaxy (MBE) techniques at high temperatures. It was recently discovered that in these epitaxially grown III–V alloys, the arrangement of column-III or column-V atoms, if not inhibited by growth kinetics, can exhibit a long-range order. Until these discoveries, the occupancy of column-III atoms or column-V atoms in their sublattices had always been thought to be completely random.

Three types of ordered structures have been identified in these III–V alloys, each with ordering occurring along the ⟨100⟩, ⟨210⟩, or ⟨111⟩ direction, respectively. First-principles calculations indicated that atomic ordering along the ⟨210⟩ or ⟨100⟩ direction lowers the alloy's strain energy by accommodating dissimilar bond lengths more effectively than in the random alloy. The ⟨111⟩ ordering, on the other hand, does not represent a lower-energy, preferred arrangement in the bulk. This ordered structure is induced by the surface growth process.

The ordered III–V alloys have interesting optical and electrical properties different from those in the corresponding disordered phases. Long-range order leads to significant bandgap reductions. The degree of order is thus an important controlling parameter for bandgap modulation in III–V materials. Ordered III–V alloys are being explored for use in novel devices and for improving the performance of devices.

I. LONG-RANGE ORDER IN III–V TERNARY ALLOYS

A. Ordered Structures and Their Diffraction Patterns

Most III–V alloys crystallize in the zincblende structure, which consists of two face-centered cubic (fcc) sublattices displaced from each other by one-quarter of a body diagonal. In a binary alloy one fcc sublattice is occupied by column-III atoms and the other occupied by column-V atoms (see Fig. 1). In an isovalent ternary alloy $A_xB_{1-x}C$ ($0 < x < 1$), A and B atoms belonging to the same column share one sublattice. Until recently, the arrangement of A and B atoms in this sublattice was always believed to be completely random, and phase diagrams of III–V ternary systems were all based on disordered configurations. New diffraction studies and theoretical calculations, however, indicated that A and B atoms in these ternary alloys can actually favor an ordered arrangement under certain conditions (i.e., certain sites in the sublattice are preferred by A atoms and certain sites are preferred by B atoms).

The presence of long-range order in these alloys was detected through the appearance of superstructure reflections between the Bragg reflections in their electron diffraction patterns. These superstructure reflections, forbidden in a disordered phase, arise when a long-range order is present. The arrangement of the preferred A and B sites in the sublattice can be determined from the distribution of these superstructure reflections in the diffraction pattern. Several ordered structures have been observed in III–V alloys. Most of these ordered structures fall into three categories: the CuAu-I type, the chalcopyrite type, and the CuPt type. Other possible ordered structures, such as the Cu_3Au or $CuPt_3$ type, have not been found in III–V alloys.

In the CuAu-I-type ordered structure the preferred A and B sites in the fcc sublattice are arranged as a monolayer superlattice along the ⟨100⟩ direction. This superlattice can be along the [100], or the equivalent [010] or [001] direction. These three orientations constitute the three possible variants of the ordered phase. The positions of the preferred A and B sites in the [001] variant viewed along the [$1\underline{1}0$] and along the [001] directions are shown in Figs. 2a and 2b, respectively. According to the kinematic diffraction theory, the Bragg reflections from the zincblende structure (Fig. 1) have unmixed hkl indices (i.e., h, k, and l are either all even or all odd). A monolayer superlattice along the [001] direction (the [001] variant) would give rise to extra reflections at h, k, $l + 1$ between two neighboring Bragg reflections h, k, l and h, k, $l + 2$. Typical diffraction patterns of this variant observed in MOVPE-grown $Al_xGa_{1-x}As$ layers are shown in Figs. 2c and 2d. In these patterns, weak reflections observed at 001, $00\underline{1}$, $\underline{1}10$, $1\underline{1}0$, 110, $\underline{11}0$, ... are superstructure reflections from the ordered structure. From diffraction patterns one can also tell how many variants of the ordered phase are present, since different variants give rise to different sets of superstructure reflections. For instance, superstructure reflections at $h + 1$, k, l come from the [100] variant, and those at h, $k + 1$, l come from the [010] variant.

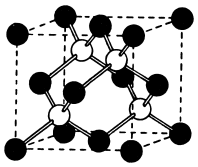


FIGURE 1 Arrangement of atoms in a unit cell of a III–V compound: (●) represent the column-III atoms and (○) the column-V atoms.

The chalcopyrite structure in a tetragonal unit cell with a c axis twice as long as a (or b) axis is shown in Fig. 3a. This structure consists of a bilayer superlattice when viewed along the [201], [$\underline{2}01$], [021], and [$0\underline{2}1$] directions. This bilayer superlattice configuration can be easily recognized if the structure is projected along the [010] direction, as in Fig. 3b, or along the [100] direction. The bilayer superlattice would give rise to superstructure reflections at $h + 1$, k, $l + \frac{1}{2}$, and h, $k + 1$, $l + \frac{1}{2}$, where h, k, and l are all even or all odd. The chalcopyrite structure can have three variants, each with its tetragonal c axis oriented along the [001], [010], or [100] direction, respectively.

The chalcopyrite structure is very closely related to the CuAu-I-type structure. Every two consecutive layers of the chalcopyrite structure along the tetragonal c axis is equivalent to the CuAu-I structure. As shown in Fig. 3b, layers 1 and 2, 2 and 3, 3 and 4, ... are equivalent, respectively, to the [100] variant, [010] variant, [100] variant, ... of the CuAu-I-type structure. The chalcopyrite structure is common among II–IV–V or I–III–VI alloys, but is only recently observed in isovalent III–V alloys together with the CuAu-I-type structure. Figure 3c is a diffraction pattern obtained from a MOVPE-grown $GaAs_{0.5}Sb_{0.5}$ layer, which contains a mixture of the chalcopyrite and CuAu-I-type structures. In this diffraction pattern, the chalcopyrite reflections 0, 1, $\frac{1}{2}$; 0, $\underline{1}$, $\frac{1}{2}$; ... are from the [001] variant, and 0, $\frac{1}{2}$, 1; 0, $\frac{1}{2}$, $\underline{1}$; ... are from the [010] variant. The extra reflection 0, 0, 1, and 0, 1, 0, are from the two variants of CuAu-I-type structures.

In the CuPt-type ordered structure, the preferred A and B sites in their sublattice are arranged as a monolayer superlattice along the ⟨111⟩ direction. This structure can have four variants, each oriented along one of the four equivalent ⟨111⟩ directions. The arrangement of A and B

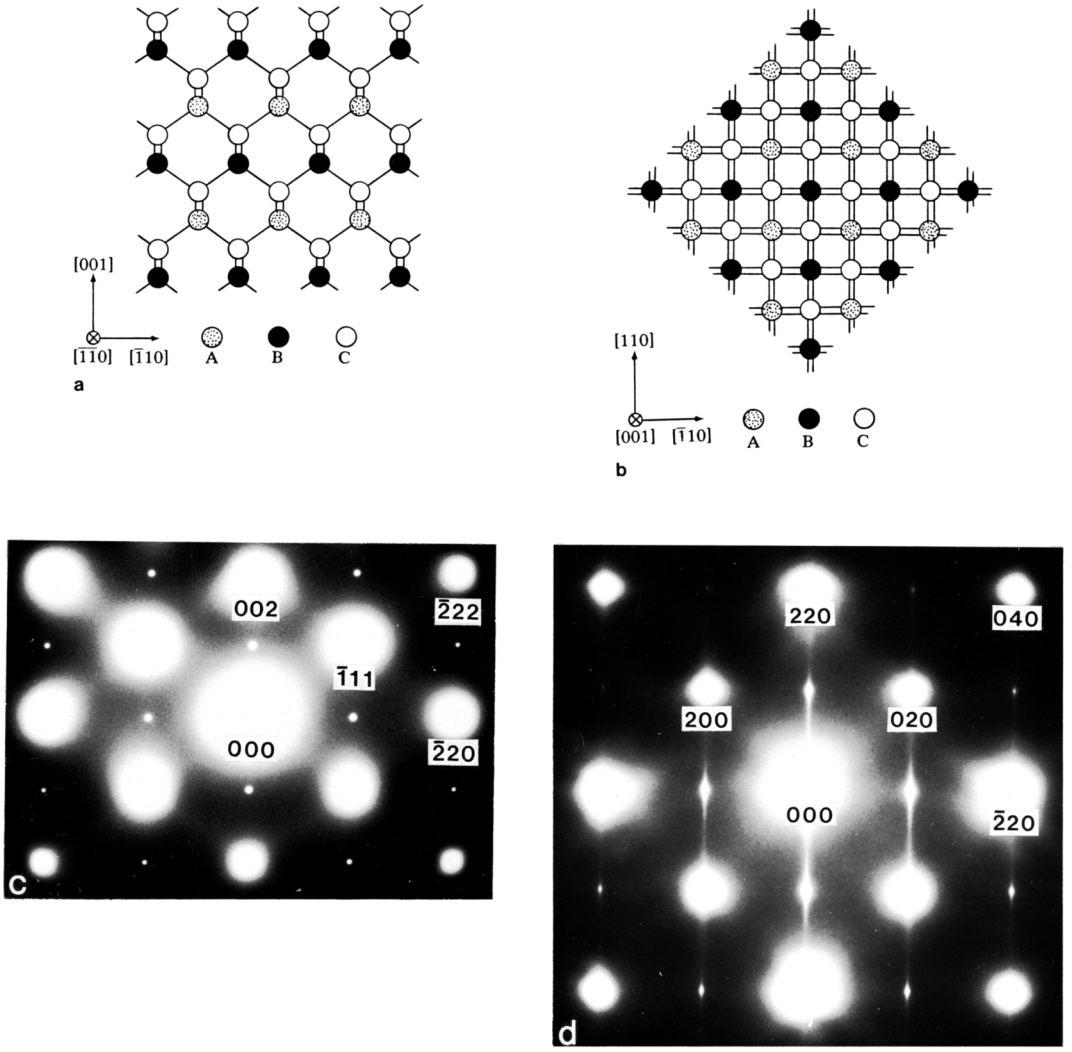


FIGURE 2 (a) Arrangement of the preferred A and B sites in the CuAu-I-type ordered structure projected along the [1$\underline{1}$0] direction. (b) The same ordered structure projected along the [001] direction. (c) Diffraction pattern taken with an incident electron beam along the [1$\underline{1}$0] direction in an $Al_{0.75}Ga_{0.25}As$ layer grown on a (110)-oriented GaAs substrate by MOVPE at 700°C. The layer contains the CuAu-I-type ordered structure depicted in (a). (d) Diffraction pattern from the same layer obtained with an incident electron beam along the [001] direction. The intensity streaks and the broadening of the superstructure reflections suggest the presence of a high density of antiphase boundaries parallel to the (110) growth surface. [Diffraction patterns from T. S. Kuan, T. F. Kuech, W. I. Wang, and E. L. Wilkie (1985). *Phys. Rev. Lett.* **54,** 201.]

sites in the [1$\underline{1}$1] variant is shown in Fig. 4a. The CuPt-type ordered structure can also occur in a binary system such as Ge_xSi_{1-x}. In this case the CuPt-type ordering occurs in both fcc sublattice as shown in Fig. 4b. A (111) monolayer superlattice in the CuPt-type ordered structure would give rise to superstructure reflections at $h+\frac{1}{2}, k+\frac{1}{2}, l+\frac{1}{2}$ between two neighboring Bragg reflections h, k, l and $h+1$, $k+1$, $l+1$. In Figs. 4c and 4d are diffraction patterns obtained from two $GaAs_{0.5}Sb_{0.5}$ layers grown on different substrates by MBE at different temperatures. Strong superstructure reflections observed at positions $h+\frac{1}{2}, k-\frac{1}{2}$, $l+\frac{1}{2}$, and $h+\frac{1}{2}, k+\frac{1}{2}, l-\frac{1}{2}$ in Fig. 4c indicate that the layer grown at a higher temperature contains two variants of highly ordered CuPt-type structure. The superstructure reflections in Fig. 4d obtained from the layer grown at a lower temperature are weaker and more diffused, indicating the presence of short-range order of similar type.

A variation of the CuPt-type ordering with a periodicity of $3d_{111}$ (as compared to the periodicity of $2d_{111}$ shown in Fig. 4a) has occasionally been observed. In this case the superstructure reflections occur at $h+\frac{1}{3}, k+\frac{1}{3}, l+\frac{1}{3}$,

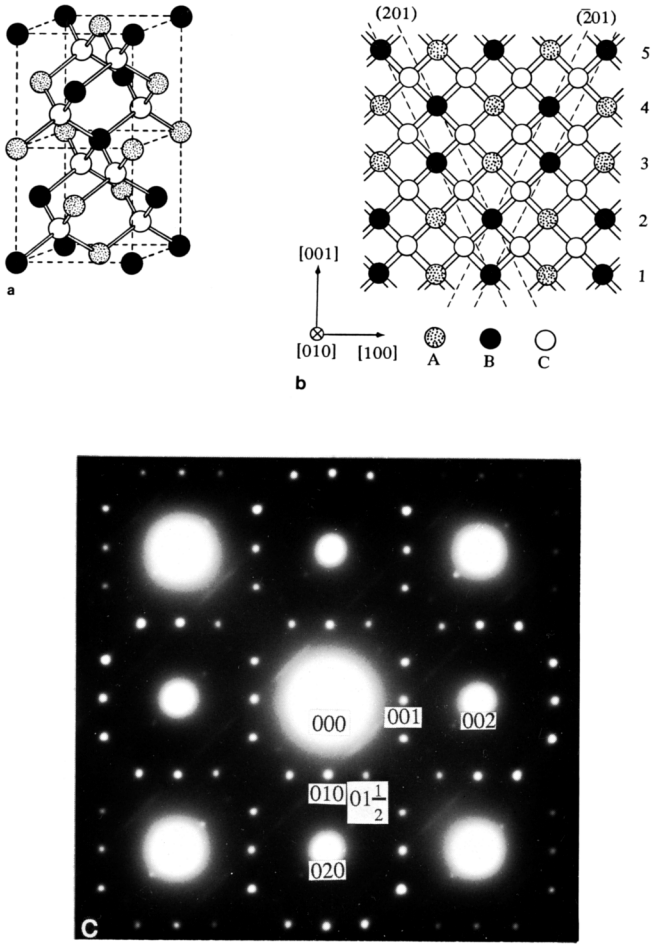


FIGURE 3 (a) Arrangement of the preferred A and B sites in the chalcopyrite structure in a tetragonal unit cell. (b) Projection of the chalcopyrite structure (the [001] variant) along the [010] direction. (c) Electron diffraction pattern from a $GaAs_{0.5}Sb_{0.5}$ layer grown on a (100)-oriented InP substrate at 600°C. This pattern contains superstructure reflections from two variants of the CuAu-I-type structure and two variants of the chalocopyrite structure. [Diffraction pattern from H. R. Jen, M. J. Cherng, M. J. Jou, and G. B. Stringfellow (1987). *In* "Ternary and Multinary Compounds, Proceedings of the 7th International Conference, Materials Research Society," p. 353.]

and $h + \frac{2}{3}, k + \frac{2}{3}, l + \frac{2}{3}$ between two neighboring Bragg reflections h, k, l and $h + 1, k + 1, l + 1$.

The symmetry, the number of atoms per unit cell, and the square of structure factors of the three most commonly observed ordered structures are listed in Table I. It is noted that the long-range order observed in III–V alloys, regardless of composition, is rarely perfect. Only part of the preferred A and B sites are correctly occupied. This partial order is described by an order parameter S, which can be expressed as

$$S = r_A + r_B - 1, \tag{1}$$

where r_A and r_B are fractions of A and B sites that are occupied by the preferred atom. The parameter S can have value ranging from 0 to 1 depending on the degree of order. For a perfectly ordered structure $S = 1$, which is possible only in stoichiometric $A_{0.5}B_{0.5}C$ compounds, and for a completely random arrangement $S = 0$. Kinematically, the intensity of superstructure reflections is proportional to $|F|^2$, which is proportional to S^2 in these structures. However, for electron diffraction, the relation between the degree of order and superstructure intensity is not as straightforward, since the observed intensity is often strongly affected by multiple scattering and the presence of defects, such as antiphase boundaries, in the ordered structure.

B. Occurrence in Epitaxially Grown Layers

It has been demonstrated for some time by Gossard *et al.* that monolayer superlattices of $(AlAs)_1(GaAs)_1$ along the [001] direction (the CuAu-I-type ordered structure) can be synthesized by alternate depositions of monolayers of

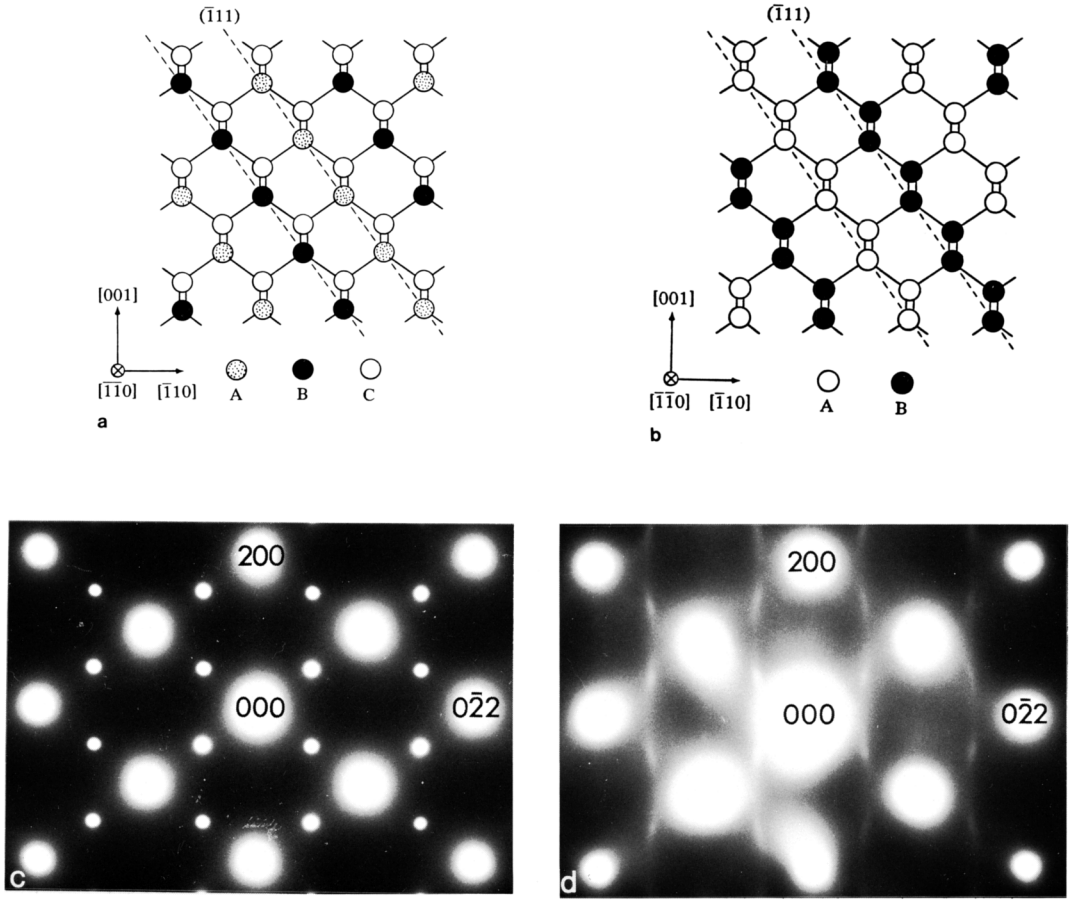


FIGURE 4 (a) Arrangement of the preferred A and B sites in the CuPt-type ordered structure (the [$\underline{1}11$] variant) projected along the [$1\underline{1}0$] direction. (b) Similar CuPt-type ordered structure in a binary system. (c) Electron diffraction pattern from a $GaAs_{0.5}Sb_{0.5}$ layer grown on a (100)-oriented GaAs substrate at 580°C. This pattern contains superstructure reflections from the [$1\underline{1}1$] and [$11\underline{1}$] variants of the CuPt-type ordering. (d) Diffraction pattern from a $GaAs_{0.5}Sb_{0.5}$ layer grown on a (100)-oriented InP substrate at 520°C. The weak and diffused superstructure reflections indicate a shorter range of order. [Diffraction patterns from Y.-E. Ihm, N. Otsuka, J. Klem, and H. Morkoç (1987). *Appl. Phys. Lett.* **51,** 2013.]

AlAs and GaAs using the MBE technique. This would require repeated manipulation of shutters that control the flux of Ga and Al atoms during the growth. Monolayer superlattices of $(InAs)_1(GaAs)_1$ along the [001] direction have also been grown by MOVPE using similar alternate monolayer deposition technique by Fukui and Saito. In contrast with these artificially grown ordered structures, the ordered phases discovered in III–V alloys were produced spontaneously in a continuous codeposition of A and B atoms without any artificial imposition of special growth sequences.

The first spontaneous long-range order was discovered by Kuan *et al.* in the $Al_xGa_{1-x}As$ alloy system. They found from the appearance of superstructure reflections that the CuAu-I-type ordered phase could be grown over a wide range of compositions ($0.15 < x < 0.8$) and growth temperatures (600°–800°C) by MOVPE. Furthermore, the growth along the ⟨110⟩ direction produced much stronger

TABLE I Ordered Structures in III–V Ternary Alloys

Order type	Symmetry	Unit cell	Number of variants	Structure factor $\|F\|^2$
CuAu-I	$P\underline{4}m2$	Tetragonal (2A, 2B, 4C)	3	$4S^2(f_B - f_A)^2$
Chalcopyrite	$I\underline{4}2d$	Tetragonal (4A, 4B, 8C)	3	$8S^2(f_B - f_A)^2$
CuPt	$R\underline{3}m$	Trigonal (4A, 4B, 8C)	4	$16S^2(f_B - f_A)^2$

ordering that along the ⟨100⟩ direction. Nakayama and Fujita found in $Ga_{0.5}In_{0.5}As$ layers grown by liquid-phase epitaxy (LPE) a larger set of superstructure reflections, which they thought was due to a famatinite-type ordered structure. The presence of famatinite structure in $Ga_{0.5}In_{0.5}As$ was not certain, however, since the same set of superstructure reflections can also result from a mixture of CuAu-I- and chalcopyrite-type structures. Jen *et al*. subsequently observed exactly the same set of superstructure reflections from a MOVPE-grown $GaAs_{0.5}Sb_{0.5}$ layer, and their analysis indicated clearly that the alloy contains a mixture of CuAu-I- and chalcopyrite-type structures. In a $GaAs_{0.5}Sb_{0.5}$ layer grown on a (110) InP substrate, Jen *et al*. found only the CuAu-I-type structure. In a $Ga_{0.5}In_{0.5}As$ layer grown by MBE on a (110) InP substrate, Kuan *et al*. observed also the presence of CuAu-I-type structure. In a II–VI alloy, $Zn_{0.5}Fe_{0.5}Se$, grown by MBE, Park *et al*. also detected the presence of the CuAu-I ordering.

The CuPt-type ordered structure was first observed in the Ge_xSi_{1-x} binary system by Ourmazd *et al*., and subsequently in MBE-grown $GaAs_{0.5}Sb_{0.5}$ by Murgatroyd *et al*., and in VLE (vapor levitation epitaxy)-grown $Ga_{0.5}In_{0.5}As$ and $Ga_{0.4}In_{0.6}As_{0.8}P_{0.2}$ by Shahid *et al*. The CuPt-type structure was also observed in MBE-grown $GaAs_{0.5}Sb_{0.5}$ by Ihm *et al*. and in MOVPE-grown $Ga_{0.5}In_{0.5}P$ by Gomyo *et al*. In $Ga_{0.5}In_{0.5}P$ grown at lower temperatures Gomyo *et al*., found a ⟨111⟩ ordering with a larger periodicity ($3d_{111}$). CuPt-type ordering was also found in MOVPE-grown $Al_{0.5}In_{0.5}P$ and $Al_{0.25}Ga_{0.25}In_{0.5}P$ by Yasuami *et al*. and in $GaAs_{1-x}P_x$ and $InAs_{1-x}Sb_x$ by Jen *et al*. These alloys exhibiting CuPt-type ordering were all grown on (001) surfaces, and in most cases only two variants out of the possible four were found. The same CuPt ordering was also observed in a LPE-grown II-VI alloy, $Hg_{0.6}Cd_{0.4}Te$, by Chang and Goo.

All these pioneer observations of alloy ordering are summarized in Table II. It seems likely that long-range order can occur in any possible III–V ternary alloy. It is noted that different ordered phases can be grown in the same alloy system by different growth techniques or under different growth conditions. To date all three ordered structures have been observed in the $GaAs_{1-x}Sb_x$ and $Ga_xIn_{1-x}As$ systems.

II. ORIGIN AND STABILITY OF ORDERED PHASES

A. Stability of Ordered Phases in Bulk III–V Semiconductors

The presence of order–disorder transition in metal binary alloys, such as Cu–Au, Cu–Pt, or Cu–Zn, implies a negative enthalpy of mixing (i.e., the interaction energy favors the bonding between pairs of unlike atoms more than between like atoms). The long-range order in III–V ternary alloys involves atoms that are second nearest neighbors to each other. Nevertheless, its occurrence is still surprising, since most of these semiconductor alloys are known to have positive enthalpies of mixing. According to Stringfellow's model, the enthalpy of mixing ΔH is always positive for all disordered ternary alloys $A_xB_{1-x}C$ with end compounds AC and BC:

$$\Delta H^{(D)} = H^{(D)}(A_xB_{1-x}C) - xH(AC) - (1-x)H(BC) > 0. \quad (2)$$

Consequently, the disordered alloy is unstable within a miscibility gap present in the phase diagram at temperatures below T_c. ($T_c = \Delta H^{(D)}/\Delta S^{(D)}$, $\Delta S^{(D)}$ is the entropy of mixing of the disordered alloy). In alloys such as $Al_xGa_{1-x}As$ with a small lattice constant difference between two end compounds, the miscibility gap occurs only at low temperatures, and the alloy is completely miscible at all compositions at growth temperatures. In alloys such as $GaAs_{1-x}Sb_x$ with a large lattice constant difference between two end compounds, a wide miscibility gap can extend all the way up to the solidus curve. However, unstable ternary alloys can usually be grown inside the gap using the MBE or MOVPE techniques, since the kinetics of phase separation can be hindered by limited surface diffusion during the growth and by an even slower rate of bulk diffusion after cooling to lower temperatures.

Srivastava *et al*., and more recently, Dandrea *et al*. and Bernard *et al*., based on first-principles total energy minimization calculations, have looked into the possibility of stable ordered phases in bulk ternary III–V alloys. They calculated the enthalpies of mixing of all possible ordered phases, $\Delta H^{(O)}$, including the CuAu-I, chalcopyrite, and CuPt types, in a wide variety of III–V alloys. They found that except for the chalcopyrite structure in certain alloys, none of the other ordered phases are fundamentally stable against phase separation. Nevertheless, certain ordered phases are more stable than the disordered phase, i.e.,

$$\Delta H^{(D)} > \Delta H^{(O)} > 0. \quad (3)$$

Among all ordered phases, the chalcopyrite and famatinite structures in alloys with dissimilar A–C and B–C bond lengths are the most stable, followed closely by the CuAu-I-type ordering. The relative stability of these ordered phases stems from their ability to accommodate more effectively the dissimilar bond lengths and reduce the alloy's microscopic strain energy. These calculated results predicted that chalcopyrite or CuAu-I-type ordering can happen in alloys such as $GaAs_{1-x}Sb_x$ and $Ga_xIn_{1-x}As$ with large bond length differences. The observed ordering in $Al_xGa_{1-x}As$ alloys with almost equal bond lengths (within ~0.1%), however, is not expected. The origin of ordering in this alloy system remains to be determined.

TABLE II Long-Range Order Observed in III–V, IV–IV, and II–VI Alloys

Alloy	Order type in sublattice	Growth method and substrate	Growth temperature (°C)	Variants observed
$Al_xGa_{1-x}As$ $(0.15 < x < 0.8)$	CuAu-I	MOVPE, MBE (110)GaAs (100)GaAs	600–800	One (110) Two (100)
$Ga_{0.5}In_{0.5}As$	Famatinite (?)	LPE (001) InP	630	Two
$Ga_{0.5}In_{0.5}As$	CuAu-I	MBE (110)InP	440–530	One
$Ga_{0.5}In_{0.5}As$	CuPt	VLE, MOVPE (001)InP	600–700	Two
$Ga_{0.4}In_{0.6}As_{0.8}P_{0.2}$	CuPt	VLE, MOVPE (001)InP (001)GaAs	600–700	Two
$GaAs_{0.5}Sb_{0.5}$	CuAu-I	MOVPE (110)InP	550–600	Two
$GaAs_{0.5}Sb_{0.5}$	CuAu-I; chalcopyrite	MOVPE (001)InP (001)GaAs	550–680	Two
$GaAs_{0.5}Sb_{0.5}$	CuPt	MBE (001)InP (001)GaAs	480–580	Two
$Ga_{0.5}In_{0.5}P$	CuPt	MOVPE (001)GaAs (001)GaP	550–750	Two
$Al_{0.5}In_{0.5}As$	CuPt (triple period)	MBE (001)InP	415–460	Two
$Al_{0.5}In_{0.5}As$	CuPt	MBE (001)InP	520–570	Two
$Al_{0.5}In_{0.5}P$	CuPt	MOVPE (001)GaAs	610–740	Two
$Al_{0.25}Ga_{0.25}In_{0.5}P$	CuPt	MOVPE (001)GaAs	650–700	Two
$GaAs_{1-x}P_x$ $(0.4 < x < 0.8$	CuPt	MOVPE (001)GaAs	640–680	Two
$InAs_{1-x}Sb_x$ $(0.2 < x < 0.9)$	CuPt	MOVPE (001)InSb (001)InAs	375–480	Two
$InAs_{0.6}Sb_{0.4}$	CuPt	MBE (001)GaSb	425	Two
Ge_xSi_{1-x} $(0.4 < x < 0.5)$	CuPt	MBE (100)Si	390–550	Four
$Zn_{0.5}Fe_{0.5}Se$	CuAu-I	MBE (001)InP	330	Two
$Hg_{0.6}Cd_{0.4}Te$	CuPt	LPE (111)CdTe	500	Four

The bulk thermodynamics argument cannot explain the CuPt-type ordering, which is the most widely observed ordered structure among the III–V alloys. Here the first-principles calculations indicated that the CuPt-type ordering is even less stable than the disordered phase, i.e., $\Delta H^{(\mathrm{CuPt})} > \Delta H^{(\mathrm{D})} > 0$. Experimental results also concluded that the CuPt-type ordering is not stable in the bulk; the ordering in GaInP and GaInAsP alloys was eliminated after a prolonged high-temperature (>700°C) anneal or after Zn diffusion into the alloys. We now believe that the CuPt-type ordering is induced by the surface reconstruction during the growth process.

B. Surface-Reconstruction Induced Ordering

Murgatroyd *et al.* suggested that the (2×4) surface reconstruction may play an important role in the formation of CuPt ordering and variant selection. This suggestion

was based on the similarity between electron diffraction patterns of the ordered $GaAs_{1-x}Sb_x$ alloys and the reflection high-energy electron diffraction (RHEED) patterns of the (001) substrate surface. LeGoues *et al*. showed that the appearance of four variants of CuPt-type ordering in Ge_xSi_{1-x} is tied to the growth on a (2×1) reconstructed surface. They suggested that local stain field associated with surface dimers can produce site preferences for Si and Ge atoms leading to the CuPt-type ordering. The role of surface dimers was further established by Gomyo *et al*., who found that the (2×1) surface reconstruction produces the CuPt-type ordering in $Al_xIn_{1-x}As$, while the (2×3) reconstruction produces a different triple period ordering in the same alloy. Murata *et al*. have recently verified the direct correlation between the density of surface dimers present during the MOVPE growth and the degree of order achieved in $Ga_xIn_{1-x}P$ alloys.

First-principles calculations of Zhang *et al*. indicated that the CuPt phase indeed represents a lower energy and stable configuration within a few atomic layers next to a reconstructed surface. Figure 5 shows a most commonly observed (2×4)-β reconstructed (001) surface with rows of column-V dimers along the [$\underline{1}$10] direction. The dimer bonds put the column-III site A under compression and site B under tension. Given sufficient near-surface mobility during growth, site A will be preferred by smaller atoms and site B by larger atoms, leading to the CuPt-type ordering (compare Fig. 5 to Fig. 4a). The strain field associated with dimer formation provides a strong driving force toward ordering, since the preferred occupancy has a much lower energy (~150 meV/atom for $Ga_{0.5}In_{0.5}P$) than the reversed configuration. The dimer-induced site selection allows only two variants with ordering direction normal to the dimer rows to be formed, in agreement with observations.

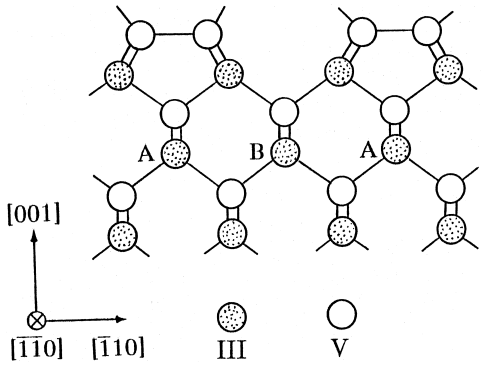


FIGURE 5 Schematic of a (2×4)-β reconstructed (001) surface projected along the [$\underline{1}$10] direction. Column-V atoms cover 75% of the top layer and form rows of dimers along the [110] direction. Site A directly under a dimer prefers a smaller atom because the dimer bond compresses it. Site B is under tension and therefore prefers a larger atom. The A and B site preferences during growth give rise to the CuPt-type ordering shown in Fig. 4a.

Surface dimers may assist or even be responsible for the formation of other types of ordering, e.g., chalcopyrite and CuAu-I. This prospect remains to be explored. Surface steps on the growth surface also affect the degree of order and variant preference. Many models have been proposed to explain the role of surface steps, but the involved atomic mechanisms are still not well understood.

C. Effects of Growth Kinetics

The onset of ordering requires high surface and subsurface mobility of the depositing species. Growing $Al_xGa_{1-x}As$ on a (110) GaAs surface produces much more enhanced ordering than on a (100) surface, since surface mobility is generally higher on the atomically flat (110) surface. At low growth temperatures, ordering is prohibited by lack of sufficient surface mobility or by changes in surface structure. At high growth temperatures, the degree of order also drops, because surface reconstruction ceases to exist, or the entropy of the alloy favors disordered phases even at the surface. The ordered phases can therefore be grown only within a narrow temperature range (see Table II).

Since the column-V dimers on a (001) growth surface are responsible for the onset of CuPt-type ordering, growth at a high column-V partial pressure can enhance the degree of order. Growth rate can also affect the degree of order. Cao *et al*. found in their growth of $Ga_{0.5}In_{0.5}P$ by MOVPE a marked drop in the degree of order for growth rate higher than 4 μm/hr. The ordering ceased to exist at a growth rate of 12 μm/hr. The growth rate effect suggests that the time constant of the ordering process occurring on the surface during growth is of the order of 0.1–1 sec.

As discussed in Section II.B, the direction of dimer rows dictates the selection of the allowed variant. For instance, the (2×4) reconstruction with dimer rows along the [110] direction produces two "B-variants" [ordering on ($\underline{1}$11) and (1$\underline{1}$1) planes], and the (2×2) and (2×3) reconstructions with dimer rows along the [$\underline{1}$10] direction produce the other two "A-variants" [ordering on (111) and ($\underline{1}\underline{1}$1) planes]. For the CuAu-I- and chalcopyrite-type structures, only those variants with their tetragonal c axis parallel to the growth surface are formed. The exact reason for the preference of these variants is not clear.

III. OPTICAL AND ELECTRICAL PROPERTIES OF ORDERED PHASES

Earlier success in synthesizing (100) $(GaAs)_1(AlAs)_1$ monolayer superlattices by Gossard *et al*. using an alternate monolayer deposition technique had stimulated wide interest in the unique optical and electrical properties

of monolayer superlattices. New band structure and interesting phonon properties and zone folding effects were predicted. Fukui and Saito found a smaller bandgap in their MOVPE-grown (100) $(InAs))_1(GaAs)_1$ monolayer superlattices than in disordered $Ga_{0.5}In_{0.5}As$. The CuPt-type long-range order in $Ga_{0.5}In_{0.5}P$ was found by Gamyo *et al*. to yield a significant bandgap reduction of ~80 meV. Similar bandgap reductions have also been observed in ordered $Ga_{0.5}In_{0.5}As$ (65 meV) and $InAs_{0.6}Sb_{0.4}$ (45 meV). Theoretically, the bandgap narrowing is proportional to S^2 [see Eq. (1)], and Wei *et al*. have calculated and predicted bandgap reductions of 300–400 meV in perfectly ordered ($S = 1$) $Ga_{0.5}In_{0.5}P$ or $Ga_{0.5}In_{0.5}As$ alloys. The observed bandgap values thus reflect a degree of order $S \sim 0.5$ achieved so far in these alloys.

Bandgap energy controls the wavelength in light-emitting or light-absorbing devices. When a wide bandgap is desired to achieve short wavelengths, CuPt-type ordering should be avoided. However, in alloys such as $InAs_{1-x}Sb_x$, bandgap shrinkage due to ordering can extend the wavelength further into the 8 to 12-μm atmospheric window for IR detector applications. Bandgap reduction due to ordering, if large enough, can transform an alloy from an indirect-gap material to a direct-gap material. Alloy ordering changes the band alignment at heterojuctions. The noncubic symmetry of ordered phases (see Table I) also leads to a splitting of the valence bands and possible improvement in laser performance.

Ordered alloys are expected to have transport properties superior than disordered alloys due to the absence of disorder scattering. Nakata *et al*. observed that the CuAu-I ordering leads to a higher two-dimensional electron mobility in GaInAs/AlInAs modulation-doped heterostructures grown on (110) InP substrates. The dependence of bandgap energy on alloy ordering provides a new method to modulate the bandgap during growth without changing the alloy composition. By controlling the growth temperature, abrupt order/disorder transitions can be achieved without introducing lattice defects such as dislocations. Heterostructures consisting of order/disorder layers are now being actively explored for device applications.

SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • CRYSTAL GROWTH • EPITAXIAL TECHNOLOGY FOR INTEGRATED CIRCUIT MANUFACTURING • MOLECULAR BEAM EPITAXY, SEMICONDUCTORS • PHOTONIC BANDGAP MATERIALS • SUPERLATTICES

BIBLIOGRAPHY

Bernard, J. E., Dandrea, R. G., Ferreira, L. G., Froyen, S., Wei, S.-H., and Zunger, A. (1990). *Appl. Phys. Lett.* **56,** 731.

Cao, D. S., Reihlen, E. H., Chen, G. S., Kimball, A. W., and Stringfellow, G. B. (1991). *J. Cryst. Growth* **109,** 279.

Chang, K. T., and Goo, E. (1992). *J. Vac. Sci. Technol. A* **10,** 1549.

Dandrea, R. G., Bernard, J. E., Wei, S.-H., and Zunger, A. (1990). *Phys. Rev. Lett.* **64,** 36.

Fukui, T., and Saito, H. (1984). *Jpn. J. Appl. Phys.* **23,** L521.

Gomyo, A., Suzuki, T., and Iijima, S. (1988). *Phys. Rev. Lett.* **60,** 2645.

Gomyo, A., Makita, K., Hino, I., and Suzuki, T. (1994). *Phys. Rev. Lett.* **72,** 673.

Gossard, A. C., Petroff, P. M., Wiegmann, W., Dingle, R., and Savage, A. (1976). *Appl. Phys. Lett.* **29,** 323.

Ihm, Y.-E., Otsuka, N., Klem, J., and Morkoç, H. (1987). *Appl. Phys. Lett.* **51,** 2013.

Jen, H. R., Cherng, M. J., and Stringfellow, G. B. (1986). *Appl. Phys. Lett.* **48,** 1603.

Jen, H. R., Ma, K. Y., and Stringfellow, G. B. (1989). *Appl. Phys. Lett.* **54,** 1154.

Jen, H. R., Cao, D. S., and Stringfellow, G. B. (1989). *Appl. Phys. Lett.* **54,** 1890.

Kuan, T. S., Kuech, T. F., Wang, W. I., and Wilkie, E. L. (1985). *Phys. Rev. Lett.* **54,** 201.

Kuan, T. S., Wang, W. I., and Wilkie, E. L. (1987). *Appl. Phys. Lett.* **51,** 51.

LeGoues, F. K., Kesan, V. P., Iyer, S. S., Tersoff, J., and Tromp, R. (1990). *Phys. Rev. Lett.* **64,** 2038.

Murata, H., Ho, I. H., Su, L. C., Hosokawa, Y., and Stringfellow, G. B. (1996). *J. Appl. Phys.* **79,** 6895.

Murgatroyd, I. J., Norman, A. G., and Booker, G. R. (1986). Presented at Materials Research Society Spring Meeting, Palo Alto, CA, April 15–19.

Murgatroyd, I. J., Norman, A. G., and Booker, G. R. (1990). *J. Appl. Phys.* **67,** 2310.

Nakata, Y., Ueda, O., and Fujii, T. (1991). *Jpn. J. Appl. Phys.* **30,** L249.

Nakayama, H., and Fujita, H. (1986). *In* "GaAs and Related Compounds 1985," Institute of Physics Conference Series No. 79, p. 289, Hilger, London.

Ourmazd, A., and Bean, J. C. (1985). *Phys. Rev. Lett.* **55,** 765.

Park, K., Salabanca-Riba, L., and Jonker, B. T. (1992). *Appl. Phys. Lett.* **61,** 2302.

Shahid, M. A., Mahajan, S., Laughlin, D. E., and Cox, H. M. (1987). *Phys. Rev. Lett.* **58,** 2567.

Srivastava, G. P., Martins, J. L., and Zunger, A. (1985). *Phys. Rev.* **B31,** 2561.

Stringfellow, G. B. (1974). *J. Crystal Growth* **27,** 21.

Stringfellow, G. B. (1999). "Thin Films: Heteroepitaxial Systems," World Scientific, River Edge, NJ, p. 64.

Wei, S.-H., Franceschetti, A., and Zunger, A. (1996). *Mater. Res. Soc. Symp. Proc.* **417,** 3.

Yasuami, S., Nozaki, C., and Ohba, Y. (1988). *Appl. Phys. Lett.* **52,** 2031.

Zhang, S. B., Froyen, S., and Zunger, A. (1995). *Appl. Phys. Lett.* **67,** 3141.

Zunger, A., and Mahajan, S. (1994). "Handbook on Semiconductors," Vol. 3, Elsevier Science B. V., Amsterdam, p. 1399.



Solid-State Imaging Devices

Albert J. P. Theuwissen
Philips Semiconductors Image Sensors
Delft University of Technology

I. Solid-State Physics: A Summary
II. Photon Sensing
III. Signal Reading
IV. Device Architecture
V. Color Imaging
VI. Light Sensitivity
VII. Noise
VIII. Conclusions

GLOSSARY

Active Pixel Sensor An image sensor based on pixels with internal amplification, fabricated in CMOS technology.

CCD Semiconductor technology and device architecture, based on series and parallel connection of capacitors, used mainly for solid-state imaging.

CMOS Standard semiconductor technology, based on the combination of n-channel and p-channel transistors, used mainly for digital and analog circuitry.

Frame-transfer CCD A two-dimensional image sensor based on the CCD principle.

Interline-transfer CCD A two-dimensional image sensor based on the CCD principle.

Passive Pixel Sensor An image sensor based on photodiodes, without internal amplification, fabricated in CMOS technology.

SOLID-STATE IMAGE SENSORS replaced the classical imaging tubes in video applications in the mid-1980s. Although the first papers on solid-state imaging were published in the 1960s, it took almost 2 decades before solid-state imagers could make any inroads into the consumer market. The main reason for this long design-in period was the lack of a mature technology to fabricate the imaging devices. Photographic equipment has recently gone through the same process as the imaging tubes did in the past. The cost-quality ratio of solid-state image sensors is becoming so attractive that integrated circuits are replacing the classical silver–halide film in consumer photography. This is turning chemical photography into electronic or digital photography.

Today's image sensors are based on two technologies: CCDs (Charge-Coupled Devices), which are made in a dedicated semiconductor process, and CMOS imagers (Complementary Metal–Oxide-Semiconductor), which are produced in a standard semiconductor process.

This chapter will deal with both device types. The main difference between the two types is the device architecture and the fabrication process. Both devices make use of a silicon substrate in which the incoming photons are converted into an electrical signal. This means that several of the imaging properties are very similar for both CCDs and CMOS imagers, and it does not really matter which one consider.

I. SOLID-STATE PHYSICS: A SUMMARY

Mother nature has provided the electrical engineers with various materials, which can be subdivided into three categories depending on their electrical conducting potential:

- Insulators do not conduct electrical current at all.
- Metals conduct electrical current very easily.
- Semiconductors can conduct electrical current under certain circumstances.

This particular subdivision is based on the chemical construction of the atoms and molecules of the various materials. Atoms and molecules without any free electrons in their outer shells have no carriers available for conduction. To free these electrons, one has to supply quite a lot of (thermal) energy to these charge carriers. In semiconductor physics' terms: the valence band of the material under consideration is completely filled with electrons, and the conduction band is completely empty. The energy gap between the valence band and the conduction band for electrons is too large and cannot be bridged by thermal stimulation at room temperature. Electrical conduction is not possible due to the lack of free electrons in the conduction band, and these materials are therefore categorized as insulators. Examples are silicon–oxide SiO_2 and silicon–nitride Si_3N_4.

The opposite is true for metals. Metal atoms do have a lot of free electrons present in their outer shells, and these carriers can contribute to electrical conduction. In semiconductor physics' terms: the conduction band of a metal is partly filled with electrons, or the conduction band and the valence band overlap each other, so that electrons can simply "jump" from the valence band into the conduction band (the energy gap between both is negative). Even at room temperature, enough free electrons are available to conduct the electrical current. Examples of metals are aluminum (Al), tungsten (W), and copper (Cu).

Insulators have an energy gap that is too wide; metals have no gap at all. Semiconductors are in-between these two. They do have an energy gap between the valence band and the conduction band (the bandgap), but it can be overcome by electrons due to thermal stimulation, even at room temperature. Examples of semiconductors are silicon (Si) and germanium (Ge).

Another way to overcome the bandgap in semiconductors is by optical absorption: impinging photons can be absorbed by the semiconductor substrate. The energy of the photon is transferred to the semiconductor by the generation of an electron–hole pair. The electron is swept into the conduction band and a hole is left in the valence band.

Semiconductors physics and semiconductor technology are based on the ability to "tune" the number of electrons (or holes) in the conduction (valence) band, by "doping" the base material with donor (acceptor) atoms. The doping material adds a single energy level within the bandgap of the doped substrate, in such a way that the remaining energy gap (toward the conduction band) for the electrons supplied by the doping material is very small. Electrons can freely jump from their energy level (introduced by the dopant) into the conduction band of the silicon substrate. The electrical conductivity and the electrostatic properties of semiconductor materials can be accurately controlled using doping. When doping a silicon substrate with phosphorous, free electrons are added to the crystal lattice, and the substrate is called "n-type" silicon. The opposite is true for boron doping, resulting in "p-type" silicon.

It is not only doping atoms that can create extra energy levels in the semiconductor bandgap. Crystal defects and impurities can also add extra energy levels, and if these energy levels are somewhere around the middle of the bandgap, they can have a negative effect on the characteristics of the semiconductor devices. For instance, electrons from the valence band can jump relatively easily toward this intermediate energy level, and from there they can then jump into the conduction band of the semiconductor substrate, resulting in a leakage current. The exact number of impurities or defects in a substrate is very hard to control, and so is this two-step process of generating extra electrons in the conduction band. The generation of leakage currents should therefore be avoided.

This brief overview of semiconductor physics summarizes the most important concepts used in this chapter: how electron–hole pairs are generated by photon absorption, how leakage current is generated, and how the characteristics of a semiconductor can be tuned using doping.

II. PHOTON SENSING

Solid-state imaging is based on the physical principle of converting light (photons) into a measurable quantity (electrical voltage, electrical current). The link between the photons at the input of an imager and the voltages at the output of the device are the electrons. The collection and transport of those electrons are of great importance in this

chain. This section pays particular attention to all the items involved in the basic operation of solid-state imaging.

Energetic species (photons) falling onto and penetrating into a semiconductor substrate can transfer part of their energy to the substrate by generating electron–hole pairs. For an n-type semiconductor, if the energy content of the photons is high enough, electrons can be released from the valence band and swept into the conduction band, leaving behind a hole in the valence band. To generate an electron–hole pair, the energy of the photons has to be larger than the bandgap of the semiconductor substrate.

If this principle is applied to solid-state imaging using a silicon substrate, photons can create electron–hole pairs in silicon if their energy is higher than 1.1 eV, or if their wavelength is shorter than about 1100 nm. After the generation of the charge carriers in the bulk of the semiconductor, the negative electrons have to be separated from the positive holes. The easiest way to do is by applying an electrical field by which the electrons are collected and the holes are drained (or vice versa, depending on the type of substrate). If the separation is successful, however, it is almost impossible to transport a single electron or hole to an output stage, and to convert it into a measurable quantity: its energy content is much too small. This problem can be avoided by integrating: instead of transferring each single electron or hole, the charge carriers are integrated locally into a charge packet, over a certain period of time. In this way, the number of charge carriers in such a packet can be large enough to represent an energy which is easily detectable by the output of the device.

The separation of the electrons from the holes and the integration of charge carries, both actions in the bulk of the semiconductor, can be done using a small capacitor. Two alternatives are shown in Fig. 1: a reverse biased metallurgical junction (photodiode) and an externally induced junction (photogate). The separation of the electrons and the holes is realized by the electric field across the junctions: electrons are collected and holes are drained to the p-type substrate. The integration of the electrons takes place on the reverse biased and electrically isolated junction capacitances. In both cases, the voltage across the capacitance decreases with the generation and collection of electrons. Consequently, the width of the depletion region around the junction also decreases, depending on the number of electrons that are integrated or collected.

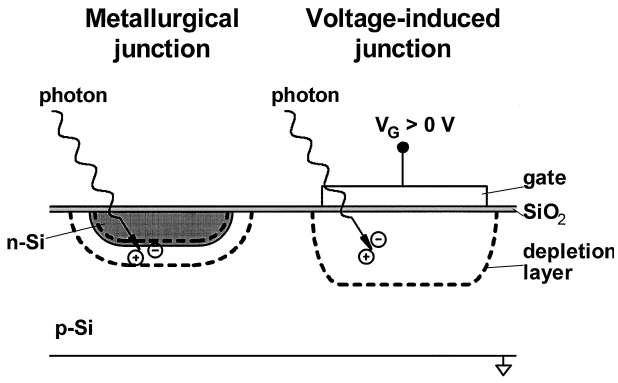


FIGURE 1 Collecting photon-generated electrons using a reverse biased metallurgical junction (photodiode) and an externally induced junction (photogate).

III. SIGNAL READING

The next link in the imaging chain is the transport of the integrated charge packets from the integrating sites towards the output of the device. Again, two alternatives are shown in Fig. 2: an MOS transistor acting as a switch with a sensing line, and a shift register based on the CCD principle. In both cases, the imaging cell or pixel is a photodiode built on a p-type silicon substrate. The choice between the MOS switch combined with the sensing line and the CCD shift register is application dependent, both solutions having their advantages and disadvantages.

A photodiode or a photogate in combination with a MOS switch and a sensing line, forms the basic architecture of a CMOS imager, while a CCD is made up of photodiodes or photogates combined with CCD shift registers. The fabrication technology for the MOS switch with the sensing line is relatively simple, but the MOS switch connects the small capacitance of the pixel to the relatively large capacitance of the sensing line. This is why the signal-to-noise ratio performance of this construction is rather poor. The technology for the CCD shift register, on the other hand, is somewhat more complicated, and the transfer of the charge packet from the pixel toward the output node needs to be appropriately clocked. The

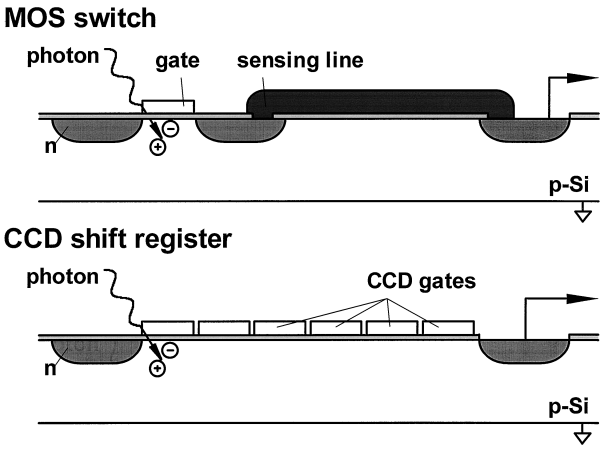


FIGURE 2 Reading the video signal using a MOS transistor and a sensing line, or using a CCD shift register.

charge packet taken from the small pixel capacitance is transferred to the small capacitance of the output amplifier. This can result in a relatively high signal-to-noise ratio performance when the imagers are made using a CCD shift register as the charge transport medium. Figure 2 only shows the combination of a metallurgical junction with the readout schemes. An MOS capacitor could also be combined with the two alternatives.

IV. DEVICE ARCHITECTURE

The operation of a single imaging cell or pixel was explained in the previous section. In practical applications, images can be built up one dimensionally (e.g., facsimile), but most of them are constructed in a two-dimensional configuration (e.g., home video, digital still applications). In this section, only the following two-dimensional imager architectures are discussed: frame-transfer CCD, interline-transfer CCD, and CMOS imagers with passive and active pixel.

A. Frame-Transfer CCD

All frame-transfer devices are built using MOS capacitors or photogates as sensing elements. However, all CCD cells are also composed of MOS capacitors, thus enabling the designers to combine both functions in one structure, by using the CCD shift register itself as a linear solid-state imager. A two-dimensional imaging array can be created by placing several of these linear devices close to each other, as depicted in Fig. 3. Each light-sensitive CCD register is extended by a CCD shift register of the same length, but which is shielded from incoming light. These shielded CCD lines can act as a temporary memory for the information which is shifted into them from the light-sensitive lines above.

The complete operation of a frame-transfer device can be summarized as follows: all CCD cells in the light-sensitive top part of the device or image section are biased in the integrating mode. Part of the CCD phases or gates are connected to a high DC level, and part of them to a low DC level, thus creating individual potential wells in the silicon substrate. In this situation, charges generated by impinging photons are collected in the induced potential wells. At the end of a defined integration time, the CCD shift registers transfer their charge packets to the corresponding CCD lines in the light-insensitive memory or storage section. This transfer from image to storage area has to be done as quickly as possible, to avoid any disturbance caused by extra charge generation from the information already available in the imaging section's CCD shift registers. Once the complete information from the image section has been transported to the storage section, the

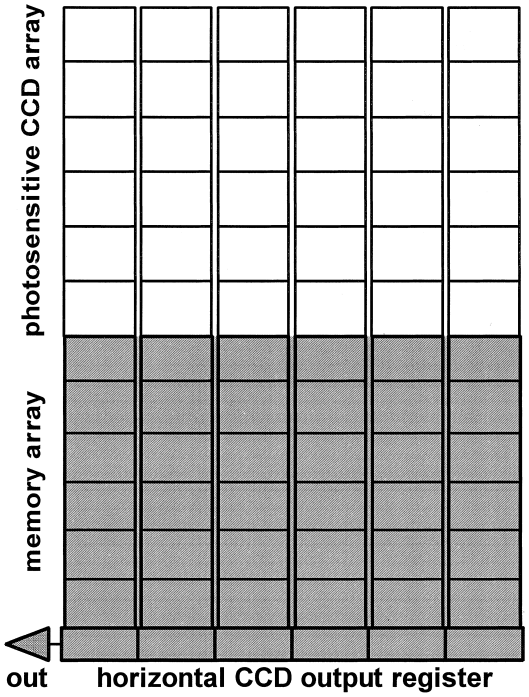


FIGURE 3 The basic architecture of a Frame-Transfer CCD image sensor.

transport of the charge packets toward the output structure of the device can start. During this action, the charge packets located on a horizontal line (and therefore belonging to different vertical CCD shift registers) are transferred to a single horizontal output CCD register. Once the charge packets have been shifted into the horizontal output register, the final transfer to the output stage can take place, charge packets can be converted to an electrical voltage, and the video signal can be processed further.

During the transfer of the video information from the light-sensitive area to the memory area, photons continuously impinge on the sensor (unless a shutter is used). This effect results in a spurious signal called "smear." To minimize the smear, the charge packets need to be transferred as quickly as possible. While transporting the video information from the storage area to the output stage, the CCD cells of the image section are again biased in the integration mode, which allows the next image to be integrated.

Frame-transfer CCDs with a complete storage area are used in standard video applications, but a derivative of the frame-transfer CCD without storage area (or even with a reduced storage area) is attractive for applications in combination with a shutter, such as digital still photography.

B. Interline-Transfer CCD

The architecture of the interline-transfer device is based on photodiodes for the light converting parts, in combination

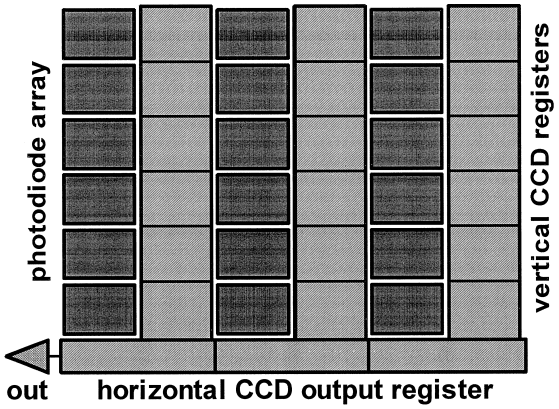


FIGURE 4 The basic architecture of an Interline-Transfer CCD image sensor.

with CCD shift registers to transfer the video information to the output amplifier. The typical construction of an interline-transfer device is shown in Fig. 4. Its basic operation can be summarized as follows: the integration of charge takes place in the photodiodes. At the end of the integration time, the charge packets are shifted from the pixels into the nearby vertical CCD shift registers. These shift registers are shielded from light and act as a memory for the information coming from the pixels. After this transfer, a new integration period can start, and the shift of the video information of the image just built-up can also begin. The vertical shift registers are emptied into the horizontal output register, line by line, in the same way as with the frame-transfer device. The serial information in the horizontal output register is transferred to the output, and is consequently converted into an electrical voltage or current.

Strangely enough, the interline-transfer CCD also shows some disturbance from the video signal. The origin of this noise component is totally different when compared to the frame-transfer imagers. In the case of the interline-transfer devices, electrons can be generated directly into the vertical CCDs, or if they are generated deeply in the semiconductor bulk, they can diffuse into the vertical registers instead of into the photodiodes.

With an analog signal back to the frame-transfer device, the interline-transfer CCD imager can now integrate a new image during the storage of the previous one in its memory part: the vertical CCD shift registers.

C. CMOS Image Sensor with Passive Pixels

A CMOS XY-addressable imager is a matrix of photodiodes, each of which is provided with a MOS transistor acting as a switch. This setup is shown schematically in Fig. 5. The basic structure of a unit cell is very similar to the one depicted in Fig. 2, where the photodiode is connected to the output by means of a sensing line. To convert the

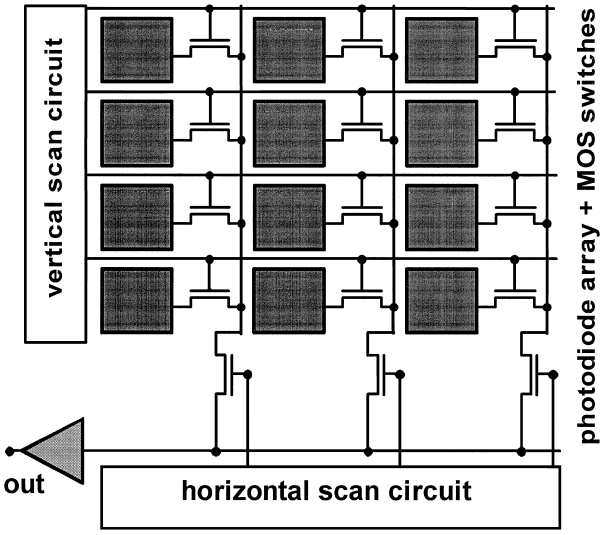


FIGURE 5 The basic architecture of a CMOS XY addressable imager with an array of passive pixels.

two-dimensional spatial information into the serial stream of electrical signals, electronic scan circuits are added to the device to address all pixels in a sequential mode and to read out their information. This addressing feature forms the basic operation of the device. At the beginning of a new field, the vertical scan circuit is activated. Suppose that the first row of pixels is selected. This is done by setting a high DC voltage on all gates of the MOS switches of this first row. Next, the horizontal scan circuit selects the pixels on one particular column by scanning its own outputs using a single high DC output, while all others are at a low level. This combination of one output of the vertical scanner and one output of the horizontal scanner at a high level and all the others low, selects one single pixel from the two-dimensional matrix. This pixel can be emptied and can dump its information into the output stage. Immediately after this action, the pixel can be reset and restart an integration. The neighboring pixel will then be addressed and read out.

The XY-addressed imager with passive pixels (the simple photodiode) is limited in its performance due to the relatively high levels of noise and fixed-pattern noise. These are partly caused by the small capacitance of the photodiode which is connected to the large readout capacitance of the sensing line.

D. CMOS Image Sensor with Active Pixels

A small amplifier can be added to every pixel to overcome the previously mentioned drawback of noise and fixed-pattern noise. A source-follower stage (amplification in the charge domain) is constructed within every single pixel, and with appropriate correlated-double

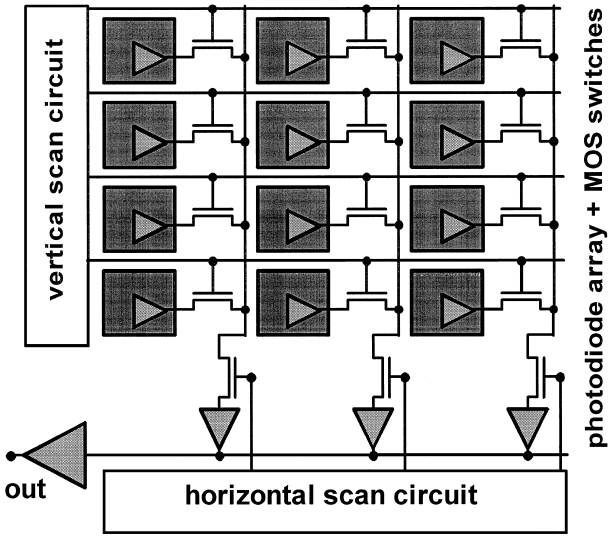


FIGURE 6 The basic architecture of a CMOS XY addressable imager with an array of active pixels.

sampling circuitry (see Section VII.D) on the column buses, the noise and the fixed-pattern noise can be limited.

The new architecture of the Active Pixel Sensor (APS) is very similar to the Passive Pixel Sensor (PPS), except for the pixel design, of course. Figure 6 shows the basic architecture of the APS. The driver of the source follower stage (the amplifier in the pixel) is implemented in the pixel itself; the load of the source follower is common for all pixels on a column.

Compared to the PPS, the overall noise and fixed-pattern noise performance of the APS sensors is improved by at least one order of magnitude.

V. COLOR IMAGING

In the case of imaging for consumer and TV applications, the camera should be able to capture color images, and has to deliver a color coded signal to the video system. The most straightforward way to do this is to provide the camera with three sensing devices, one for each main color: red, green, and blue. This technique is only applied in systems where a very high performance is required of the camera. Such an optical system built around three sensing devices is much too expensive for consumer applications. For this reason, the color separation of the incoming optical information is done using a single chip and a color filter is applied to it. Color filters used to be fabricated on a separate glass plate and glued to the sensor; nowadays, all single-chip color cameras are provided with an imager which has the color filter processed on-chip and not as a hybrid.

The color filters for a single-chip color camera can be organized in different structures, and even the colors themselves can change between manufacturers. The filters can be defined on-chip in stripes or in a mosaic configuration. The colors can be primary: red (R), green (G), and blue (B); complementary: cyan (Cy), yellow (Ye) and magenta (Mg); or even a combination of both, keeping in mind that:

$$Cy = B + G \tag{1}$$

$$Ye = R + G \tag{2}$$

$$Mg = B + R. \tag{3}$$

One of the alternatives is selected depending on the application: usually complementary colors in a mosaic configuration for video applications, and primary colors in a mosaic configuration for digital still photography. So, mosaic-filter arrangements are typically used with solid-state image sensors. The reason lies in the fact that even the most simple stripe filter construction reduces the bandwidth of the color signals to at least one-third of the bandwidth of the imager without color filters. To maintain a high resolution in the horizontal direction for color sensors with stripe filters, the number of pixels per line has to be increased considerably. This can cause severe problems. Mosaic-filter constructions only reduce the horizontal color resolution to half the resolution of the monochrome sensor resolution, but they also do the same to the vertical resolution. Stripe filters do not alter the vertical resolution.

Figure 7 shows an example of complementary colors (Yellow, Green, Cyan, Magenta) in a mosaic configuration for video applications (with interlacing). The different pixels belonging to line n and $n+1$ contain the following information (independent of the field):

$$Ye + G \quad Cy + Mg \quad Ye + G \quad Cy + Mg \tag{4}$$

$$Ye + Mg \quad Cy + G \quad Ye + Mg \quad Cy + G \tag{5}$$

With the interlaced concept, two vertically adjacent colors are always mixed together. Subtracting two horizontal neighbors yields:

$$(Mg + Cy) - (G + Ye) = 2(B - G/2) = 2(B - Y) \tag{6}$$

$$(Mg + Ye - (G + Cy) = 2(R - G/2) = 2(R - Y). \tag{7}$$

Taking these relationships into account, generation of the corresponding video signals $(R-Y)$ and $(B-Y)$ is quite straightforward. This color filter configuration with its easy video processing results in a very attractive technique for use in almost all up-to-date consumer camcorders.

A very popular application in digital photography is that of a mosaic filter configuration based on the so-called Bayer pattern, with primary colors. In applications where

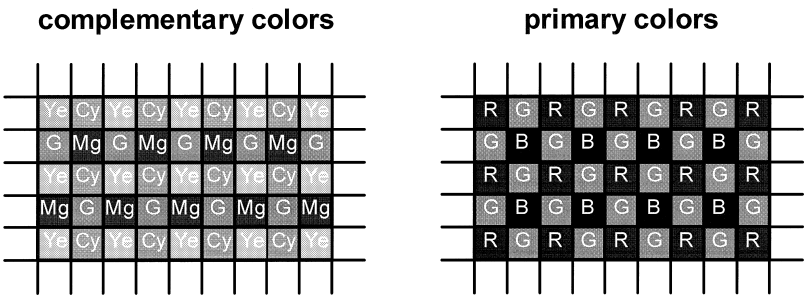


FIGURE 7 An example of color filter configurations in a mosaic pattern: typically used in standard video applications with complementary colors, and in digital photography with primary colors.

color fidelity is of prime concern, primary colors are almost a must. Every sub-matrix of two by two pixels of the monochrome sensor is covered by one red, one blue and two green filter dots. The basic filter architecture is also shown in Fig. 7.

The challenge with primary colors in a mosaic configuration is the reconstruction of the three colors in every pixel, since each single pixel only contains the information of a single color. A potential interpolation scheme is as follows:

- The missing green information (at a "blue" or "red" pixel) can be reconstructed simply by an interpolation of two horizontal or vertical neighbors.
- The missing red information (at a "green" or "blue" pixel) can be reconstructed by an interpolation of its four neighbors.
- The missing blue information (at a "red" or "green" pixel) can be reconstructed by an interpolation of its four neighbors.

The final quality of the reconstructed color picture very much depends on the algorithm which is used in the reconstruction mentioned above. The previous example of how to reconstruct all missing information is too simple to yield high-quality pictures, and is only used to illustrate the issue. The actual algorithms used are very complicated and are the main subjects of the intellectual property rights of a digital still camera.

As mentioned in the introduction of the color imaging section, the ultimate solution is an optical system with three image sensors, one for each main color. The incoming optical information is split into the three components using a color splitting prism. The camera head of a professional color camera is shown in Fig. 8: the prism is placed behind the lens of the camera, with the three imaging devices for the blue, the green and the red information glued to it.

These image sensors are electronically driven, fully in parallel, and the full black and white resolution of each device is available in each color component. After signal pre-processing of the CCD output signals (using Correlated Double Sampling, for example), the signals are fed into a signal processor to generate the chrominance and luminance signals. Luminance can be generated as

$$Y = (0.11 * B) + (0.59 * G) + (0.30 * R). \qquad (9)$$

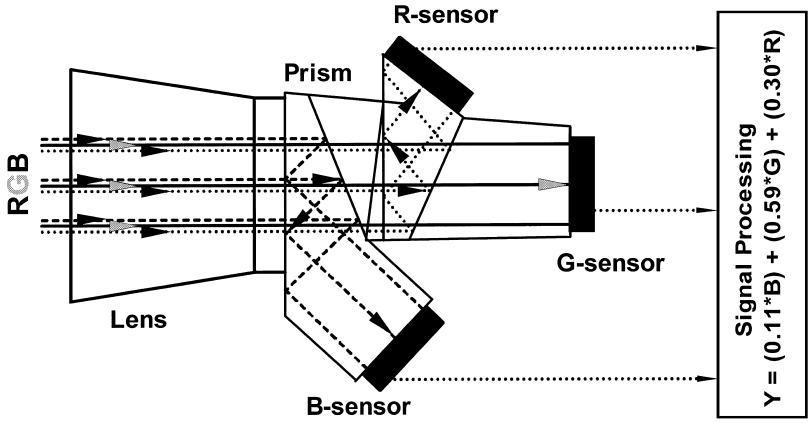


FIGURE 8 Color imaging for professional applications using three sensors glued to an optical beam-splitting prism.

VI. LIGHT SENSITIVITY

Research is continuing to increase the performance of solid-state image sensors. Improving fabrication technology, optimizing design, and creating new device architectures all have the same goal: increasing the signal-to-noise (S/N) ratio of the sensor for a given application or a given resolution. Resolution is not currently an issue as far as imaging is concerned for consumer TV applications, but is important for applications such as High-Definition TV and digital still cameras. In this section, various techniques applied to increase light sensitivity are described. Decreasing noise levels is the subject of the following section.

A great deal of effort is being put into the development of solid-state imagers to further increase the S/N ratio. This figure of merit is simply defined as being the ratio of the number of signal electrons to the number of noise electrons.

The number of signal electrons can be specified in terms of the number of photons impinging on the sensor $_0/(h.)$(=photon flux/energy of a single photon), the pixel area A_{pixel}, the quantum efficiency and the integration time T_{int}:

$$\#signal_electrons = \frac{\Theta_0}{h.\nu} * A_{pixel} * \eta * T_{\text{int}}. \qquad (10)$$

The number of noise electrons can be split up into three main noise sources:

- Photon shot noise, equal to the square root of the number of signal electrons, given by (10)
- Noise electrons generated in the CCD channels (e.g., incomplete transfer, shot noise on dark current, fixed-pattern noise, ...), all summed in n_{sens}
- Output amplifier noise: n_{out}

or the number of noise electrons can be written as

$$\#noise_electrons = \sqrt{\frac{\Theta_0}{h.\nu} * A_{pixel} * \eta * T_{\text{int}} + n_{sens}^2 + n_{out}^2}. \qquad (11)$$

If all these contributions are put in the definition of S/N, this yields

$$\frac{S}{N} = \frac{\frac{\Theta_0}{h.\nu} * A_{pixel} * \eta * T_{\text{int}}}{\sqrt{\frac{\Theta_0}{h.\nu} * A_{pixel} * \eta * T_{\text{int}} + n_{sens}^2 + n_{out}^2}}. \qquad (12)$$

Figure 9 illustrates the noise levels involved in a typical image sensor as a function of the light level input. This graph holds for both CCD imagers and CMOS imagers.

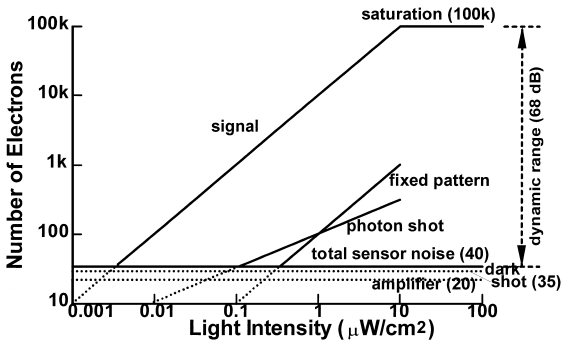


FIGURE 9 Signal and various noise levels of an imager sensor as a function of light input.

Of course, amplifier noise and sensor noise are independent of this input signal (they are strongly dependent on readout frequency, chip temperature and integration time), but the photon shot noise is proportional to the square root of the number of generated electrons. Another important noise source is the light fixed-pattern noise. This noise is static and is related to the non-uniformities between the various pixels of the imager. This fixed-pattern noise is proportional to the light input. A typical value for this noise source is 1%. Figure 9 also includes the definition of the dynamic range of the imager: the ratio of a full well to the minimum signal level at the output stage, being the noise floor of the sensor and the output amplifier.

For high levels of light input (Θ_0 high), the photon shot noise becomes dominant, and the S/N ratio becomes proportional to $\eta^{0.5}$. For low levels of light input (Θ_0 low); however, the photon shot noise becomes negligible compared to the amplifier noise, and the noise from the sensor itself, the S/N ratio, becomes proportional to η. In general, the S/N ratio is proportional to at least the square root of the quantum efficiency, but, in many situations, it is even directly proportional to the quantum efficiency. It is therefore very important to increase the quantum efficiency of the devices in order to further optimize the S/N ratio.

The quantum efficiency of the devices can be increased by optimizing the aperture ratio of the pixels, by increasing the light transmission of the multilayered structures on top of the pixels, or by illuminating the solid-state imagers from behind.

A. Aperture Ratio of the Pixels

The aperture ratio of a pixel is defined as the ratio of the light-sensitive area of a pixel to the total area of that pixel. The basic design concept of a frame-transfer pixel means that their aperture ratio is relatively high, as almost the entire pixel is sensitive to light. In contrast to

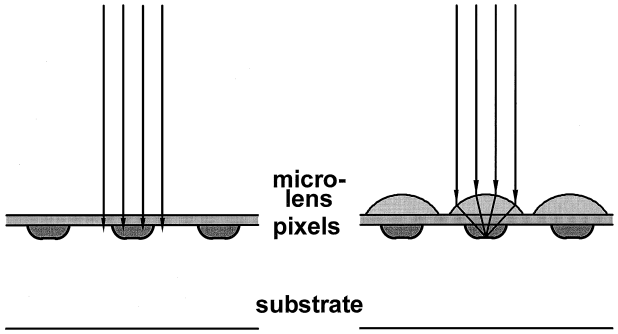


FIGURE 10 The application of microlenses on an interline-transfer pixel and a CMOS pixel.

frame-transfer imagers, the aperture ratio of an interline-transfer pixel is quite low. A CCD cell of the vertical shift register, which consumes silicon but is not light sensitive, is included next to each sensitive pixel. The same is true for the CMOS pixels: every pixel contains a light-sensitive photodiode, but also a switching transistor, a reset transistor and even a source-follower in the case of active pixel sensors. Typical values for the aperture ratio are between 25% and 45% for sensors dedicated to consumer applications. This relatively low figure can be increased by providing the pixels with microlenses. This is shown in a schematic cross-section of an interline-transfer pixel or a CMOS pixel in Fig. 10. Each pixel has its own microlens, made from a base resin layer. In this way, part of the light impinging above the light-shielded CCD register or MOS transistors can also be focused onto the sensitive part of the pixel.

B. Light Transmission of the Multilayered Structure

The aperture ratio optimization is a typical issue for interline-transfer devices and CMOS imagers since, as mentioned earlier, the aperture ratio for a frame-transfer device is relatively high. A typical issue involving the frame-transfer principle is the limited light transmission of the CCD gates on top of the pixels. Almost every interline-transfer CCD or CMOS imager is provided with photodiode pixels, which are only covered by fully transparent silicon–oxide and/or silicon-nitride. This means that the pixels of these types of imagers can be more light sensitive, in the blue and near UV spectrum, when compared to the frame-transfer CCD imagers. A great deal of effort is also put into increasing the sensitivity of frame-transfer devices. The most common solutions to enhance the light sensitivity are all related to the minimization of the light absorption in the CCD gates. One way to do this is to cut light windows in the gates themselves. Figure 11 shows an example of a four-phase CCD in which the gates are provided with light windows to enhance the device's blue sensitivity. The number of gates is not changed, but the area of absorbing polysilicon is reduced. The absorption of photons in the CCD gates can also be minimized by reducing the thickness of the polysilicon gates. This technique has resulted in a reduction of gate thickness from 500 nm (standard) to 50 nm ("membrane gates").

VII. NOISE

One way of increasing the S/N ratio is to optimize the light sensitivity of the devices as discussed in the previous section. At extremely small light inputs, however, the detection limit of any signal is determined by the remaining noise floor. As shown in Fig. 9, the dynamic range is also partly defined by the noise level of the imager and the output amplifier. Further reducing the various noise sources is just as important as increasing the light sensitivity of the devices. Since most noise sources have different origins, it is not possible to tackle them all at once; several mechanisms are technology related, others have to do with the design of the device, some can be minimized by the processing of the video signal, and others cannot.

In addition to the various noise sources in the imaging part of the sensor itself, another important noise source is the output amplifier. The small analog circuitry used to convert the electrons into a voltage and to buffer the output voltage towards the outside world, adds some uncertainty to the signals in the form of noise electrons. The overall noise from the output amplifier can be split up into different sections, depending on the mechanisms by which they are generated. To study these items separately, Fig. 12 shows a schematic diagram of the most commonly used output amplifier connected to a CCD image sensor: a double or triple source-follower stage together with a reset transistor. It is interesting to notice that exactly the same amplifier configuration as used in the CCD of Fig. 12 is also used as the amplification stage in active pixel CMOS sensors (although an APS only has a single stage source follower). The noise characteristics of this kind of output

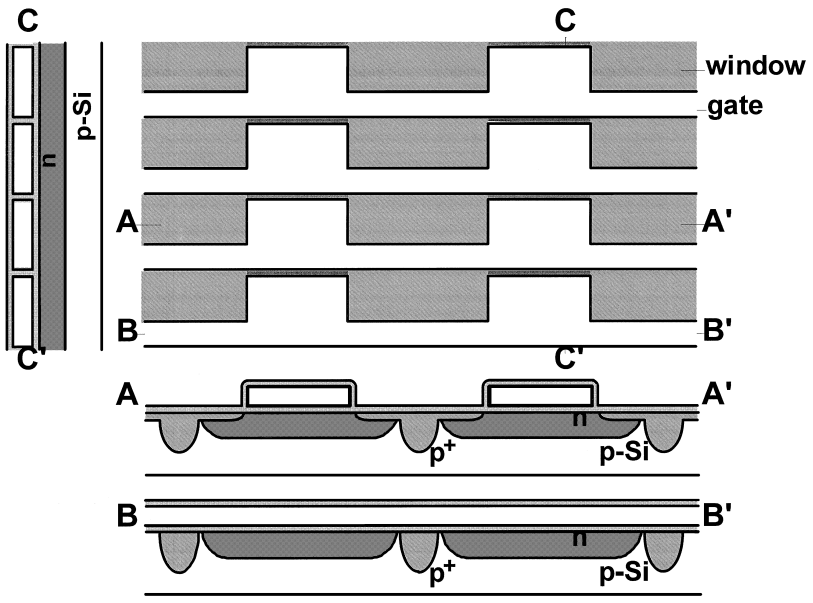


FIGURE 11 Sensitivity enhancement using cutting light windows in the active gate material of a frame-transfer CCD.

stages can be separated into: thermal noise, reset noise and 1/f noise.

A. Thermal Noise

The thermal noise of the complete system is mainly due to thermal noise generation in the inversion channels of the MOS transistors. An optimized design and a well-considered choice of the transistor areas and their width to length ratio, both for the drivers and the loads of the output stage, can keep the noise figure of the output amplifier as low as possible.

B. 1/f Noise

The 1/f noise is mainly generated in the first follower stage of the output amplifier. Interaction of the electrons in the inversion channel with the surface states, located at the Si–SiO_2 interface of the MOS transistor, cause fluctuations in the voltage at the output of the first follower stage. The interactions between the surface states and the charges can be kept to a lower level by using a depletion-type MOS transistor, since the inversion channel is pushed somewhat deeper in the silicon in these type of devices.

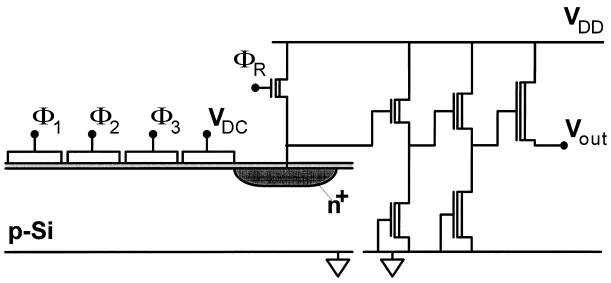


FIGURE 12 The output amplifier configuration of a CCD image sensor.

C. Reset Noise

Inherent to the reset action of the floating-diffusion capacitance is an uncertainty of kTC_{FD} on the voltage at the capacitor C_{FD}. It is very hard to minimize this reset noise, and is impossible to get rid of using an adapted design or by optimizing the technology.

An additional component of the reset noise is the partitioning noise: at the moment the reset transistor is switched off, the charges from the inversion layer underneath the gate of the transistor have to be pushed to either the source or the drain side of the device, to diminish the inversion channel. Which charges will move to the drain and which ones to the source or floating diffusion? This uncertainty causes some extra voltage fluctuation from one reset action to the next. The partitioning noise can be minimized by designing the reset transistor properly. Its channel should be funnel shaped, with the "funnel output" being the floating diffusion. During the switching-off action of the reset transistor, the narrow channel effect will then push the electrons out of the inversion channel in the direction of the widest side of the channel: the reset-drain side. The partitioning will be minimized and the partitioning noise on the floating diffusion will be as small as possible.

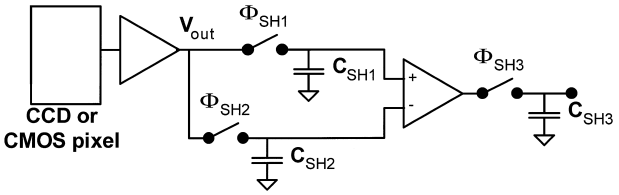


FIGURE 13 The working principle of a Correlated-Double Sampling circuitry.

D. Correlated Double Sampling

The only way to handle the reset noise is to "measure" its value and (electronically) compensate for it afterwards. This is done using a method known as "Correlated Double Sampling" (CDS), as shown in Fig. 13. The output signal of the sensor (or the CCD output, or the output of an APS pixel) is sampled twice: once for its reset-hold level (by the sample and hold pulse 1) and once for its actual video output (by the sample and hold pulse 2). The first sample takes a measure of the reset noise (C_{SH1} signal), as the reset-hold level is equal to a preset DC voltage with the reset noise kTC_{FD} added to it. The second sample takes a measure of the video signal which is superimposed on the reset noise (C_{SH2} signal). By subtracting both samples, the video signal remains, without the reset noise (OUT signal).

Including a CDS circuit in the pre-processing electronics (when using a CCD) or even in the pixel (when using an APS) also removes part of the 1/f-noise, especially in the low-frequency range.

VIII. CONCLUSIONS

Only a few aspects of solid-state imaging with CCDs and CMOS imagers have been discussed in this overview: the basic working principles of photon sensing, the different architectures of the imagers, some features of image sensors, and some recent developments to increase the device's performance. The rate of technology expansion and of device physics evolution has moved so fast since the invention of the CCDs, that it is impossible to cover all the details of charge-coupled devices in an overview of this size. Research and development work involving CCD and CMOS imagers is still going on today. These efforts deliver new ideas, new structures and new patents, as demonstrated at the many technical and scientific conferences: the International Electron Devices Meeting, the International Solid-State Circuits Conference, the Society of Photographic Institute and Engineers conferences, and the International Workshops on CCDs and Advanced Image Sensors.

SEE ALSO THE FOLLOWING ARTICLES

IMAGE PROCESSING • IMAGE RESTORATION, MAXIMUM ENTROPY METHOD • IMAGING OPTICS • LASERS, SOLID-STATE • PHOTOGRAPHIC PROCESSES AND MATERIALS • SMART PIXELS • SOLID-STATE CHEMISTRY

BIBLIOGRAPHY

Theuwissen, A. J. P. (1995). "Solid-State Imaging with Charge-Coupled Devices," Kluwer Academic Publishers, Dordrecht, The Netherlands.

Burke, M. W. (1996). "Image Acquisition," Chapman & Hall, London, UK.

IEEE Transactions on Electron Devices (1997). "Special Issue on Solid-State Image Sensors," IEEE Press, Piscataway, NJ.

Holst, G. C. (1998). "CCD Arrays, Cameras, and Displays," SPIE Optical Engineering Press, Bellingham, WA.



Superlattices

Michael J. Kelly
University of Surrey

I. Semiconductor Heterojunctions and Superlattices: New Physics and New Devices
II. Semiconductor Crystal Growth and Technologies
III. Physics and Applications of Single and Multiple Heterojunctions
IV. Physics and Applications of Superlattices
V. Mesoscopic Systems
VI. The Coming Decade

GLOSSARY

Electron volt (eV) Energy change of one electron whose potential changes by 1 V. It has a value of 1.6×10^{-19} J. Visible light has an energy of 1.5–3 eV per photon.

Epitaxy Growth of a crystal upon a substrate whose own crystal structure is continued in the overlayer.

Heterojunction Abrupt interface within a single crystal between two materials of different composition.

Homojunction Abrupt interface within a single crystal between two regions of different doping levels.

Metal–organic chemical vapor deposition (MOCVD) Cracking of gases over a substrate to form semiconductor overlayers, e.g., trimethyl-gallium and arsine to form gallium arsenide (and methane).

Molecular beam epitaxy (MBE) Deposition of semiconductor thin films under ultrahigh vacuum conditions using beams of constituent species, e.g., gallium and arsenic to form gallium arsenide.

Quantum well Thin layer of a narrower-gap semiconductor sandwiched between thicker layers of a wider-gap semiconductor, presenting a potential well for electrons in the conduction band and holes in the valence band. Quantum size effects are seen in the properties of these electrons and holes if the thin layer is only a few atomic layers thick. Multiple quantum wells refer to repeated thin layers where the intervening wide-gap semiconductor layers are relatively thick.

Superlattice A single crystal formed by the periodic repetition of a set of multilayers whose individual thicknesses range from a few to about 10 atomic layers.

Tunnel barrier A thin layer of a wider-gap semiconductor sandwiched between thicker layers of a narrower-gap semiconductor. Low-energy incident electrons can tunnel through the thin layer rather than be reflected as would happen classically.

THE DEMAND for ever-increasing performance from semiconductor devices has led to the development of

technologies to grow semiconductor crystals an atomic layer at a time. Note that one atomic layer is about 0.3 nm thick. Within each layer, the key semiconductor properties of the band gap (difference in energy between the top of the valence bands filled with electrons and the bottom of the empty conduction bands) and the level of doping can be controlled independently. Atomically abrupt interfaces between two semiconductors, called *heterojunctions*, have applications in lasers, transistors, and microwave and optoelectronic devices. In the limit, the periodicity of multilayers can be engineered during growth, resulting in the formation of artificial crystal structures, *superlattices*, with useful properties not found in naturally occurring semiconductors. Research over the last decade has concentrated on achieving ever-greater control over growth and also the formation of patterns within layers with nanometer-scale precision. Instead of thin layers, thin wires and dots have been engineered where quantum size effects are observed in two and all three spatial dimensions (as opposed to the quantum size effects in one spatial dimension in quantum wells).

I. SEMICONDUCTOR HETEROJUNCTIONS AND SUPERLATTICES: NEW PHYSICS AND NEW DEVICES

The electronic and optical properties of crystals are determined by the chemistry of the different atomic constituents, the crystal structure (i.e., the relative positions of the different atoms), and the distance over which a basic unit of the crystal structure is repeated. In semiconductors such as silicon and gallium arsenide these repeat distances are of the order of two atomic spacings, and the naturally occurring forms of these materials at room temperature are in the diamond and wurtzite crystal structures, respectively. It is now possible to grow semiconductor crystals atomic layer by atomic layer, controlling the chemical composition within each layer, and, for example, to interleave two semiconductor compositions with prescribed repeat distances (say repeats of two atomic layers of one and four of the other). The electronic and optical properties of such engineered crystals, called *superlattices*, now depend in part on the repeat distances, and the semiconductor device performance can be tailored. For example, while gallium arsenide lasers operate in the infrared, and aluminium arsenide does not lase at all, repeated multilayers of the two lase in the red. During the 1970s and 1980s, the physics and applications of epitaxially grown semiconductor multilayers came to dominate much of semiconductor science and technology. Even a single interface between semiconductors—a *heterojunction*—can be used to trap electrons and force them to move in two, rather than three, dimensions. The physics of such constrained motion is quite different from that of the motion in bulk semiconductors, and lower-noise, more temperature-stable devices have emerged.

The intrinsic quantum wavelength (the de Broglie wavelength) associated with an electron in a semiconductor at room temperature is of the order of 20 nm, or 60 atomic spacings. Once the dimensions of active regions of structures or devices become this small, one can anticipate quantum effects being important in determining the electronic and optical properties. Furthermore, the typical distance that an electron moves in a semiconductor device between scattering events is also about 20 nm: one might anticipate ballistic motion, i.e., motion without scattering, and recover vacuum valve-like behavior in small regions of a solid.

At the same time that multilayers are controlled to atomic layer precision, the lithography used to pattern surfaces is also reaching down to the nanometer scale. Clearly electronic devices cannot be miniaturized forever and still retain recognizable macroscopic behavior. Few electrons will be present if we make a device very small, and research has been active for a decade on structures which carry on average at most one current-carrying electron: their resistance at low temperatures can take only the value of $h/2e^2 = 12.91\ \mathrm{k}\Omega$ divided by an integer, and this is qualitatively different from anything macroscopic.

II. SEMICONDUCTOR CRYSTAL GROWTH AND TECHNOLOGIES

In the decades since the 1950s, the advances in semiconductor-based technologies (e.g., computation performance per unit area on a chip or per unit cost) has been exponentially rapid. This has been possible in part because of the ability to define the internal interfaces of device structures with ever-growing precision. The earliest p–n–p and n–p–n transistors of the 1950s were made by successive alloying of the dopants into the semiconductor, resulting in interfaces several micrometers (μm) thick and nonplanar (i.e., the interfaces followed the melt fronts). In the following decade, implantation of the dopants as high-energy ions, followed by annealing of the crystal to remove damage, resulted in planar interfaces, but not more than about 1.5 μm below the surface and with an interface thickness of $\sim$0.1 μm. Liquid-phase epitaxy was introduced in the 1970s. Here the growing crystal is pulled from a melt whose concentration of dopants can be changed. The interfaces are planar and of the order of 0.01 μm thick, and a much greater control and range of doping values can be achieved. Finally, in the 1980s, two forms of epitaxial

growth came to dominate the methods of fabricating sophisticated semiconductors crystals.

Molecular beam epitaxy (MBE) is basically an ultraprecise form of evaporation, in which atomic or molecular beams impinge on a heated semiconductor substrate in a very high vacuum, at such a rate as to produce monolayer coverage in about 1 sec. The most common material system with which MBE has been developed is the alloy $Al_xGa_{1-x}As$ (including the binary end points gallium arsenide and aluminium arsenide), which can be grown by controlling the relative fluxes of Al and Ga (in parallel with a flux of As_2 or As_4 molecules) to produce the desired alloy composition. The Al/Ga ratio can be altered in typically 0.1 sec, so producing an interface that is abrupt on the scale of atomic diameters. [In practice, the details of surface coverage of the molecular beams and the limited surface mobility of the added species means that the layers are not continuous over the surface but have regions differing by a monolayer with atomic steps between them. These steps can be frozen into growth, giving an effectively greater interface thickness for device applications. This produces problems for tunnel devices (see Section III.A).] One can also change the flux ratio continuously and produce linear and other composition gradings over prescribed distances. Fluxes of dopant species (Be, Si, etc.) can also be introduced for engineering the doping profiles. A schematic of an MBE chamber is shown in Fig. 1a.

In contrast, metalorganic chemical vapor deposition (MOCVD) is basically a highly controlled form of cracking, whereby $Ga(CH_3)_3$ and AsH_3 gases are carried in a stream of H_2 over a heated substrate to form GaAs and CH_4. The alloy and doping are produced with $Al(CH_3)_3$ and with $Zn(CH_3)_2$ and SiH_4 gases, respectively. The growth rates are typically 10 times faster than for MBE, and without extraordinary effects at gas flow control (which are possible); the interfaces between the layers of different alloy compositions are thicker, at two to six atomic monolayers, than the single monolayer from MBE. A schematic of an MOCVD reaction cell is shown in Fig. 1b.

The abruptness of the doping profiles is more problematic, as while, in principle, one can deliver dopants to a single atomic plane, the detailed thermodynamics and kinetics of growth in some cases mean that the dopant can rise with the growth surface. Remembering that dopant concentrations are typically at the parts per million level, for many applications this effect can be allowed for without degrading the device performance.

These two epitaxial techniques have approached the ultimate limits in tailoring multilayer materials. Early work concentrated on the growth of heterojunctions that were closely lattice matched, and the $Al_xGa_{1-x}As$ alloy system

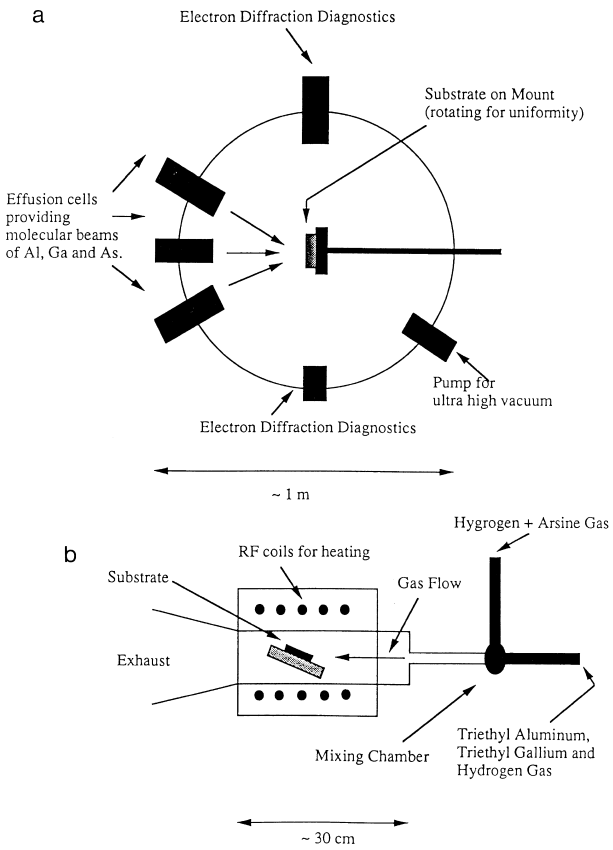


FIGURE 1 Schematics of (a) an MBE growth chamber and (b) an MOCVD reaction cell.

has been particularly convenient, as the lattice constant varies by only <0.2% between $x = 0$ and $x = 1$. Crystal growers have become increasingly bold, so that large strains (even higher than 10%) can be incorporated in very thin layers in semiconductor systems. In addition to the strained layer systems in the group III–V semiconductors, the 1990s saw many research advances in the group IV system Si/Si_xGe_{1-x} and in the group II–VI systems such as CdTe/ZnTe. There have also been efforts at growing metallic multilayers (in this case mainly by sputtering) and metal–semiconductor and semiconductor dielectric multilayers. Some attempts have also been made to combine elements of MBE and MOCVD to get the best from both.

III. PHYSICS AND APPLICATIONS OF SINGLE AND MULTIPLE HETEROJUNCTIONS

At room temperature, intrinsic gallium arsenide has a band gap of 1.42 eV separating filled valence and empty conduction bands. The alloy $Al_xGa_{1-x}As$ (at least for $x < 0.4$) has a direct band gap of $1.42 + 1.247x$ eV. For $x > 0.4$ the

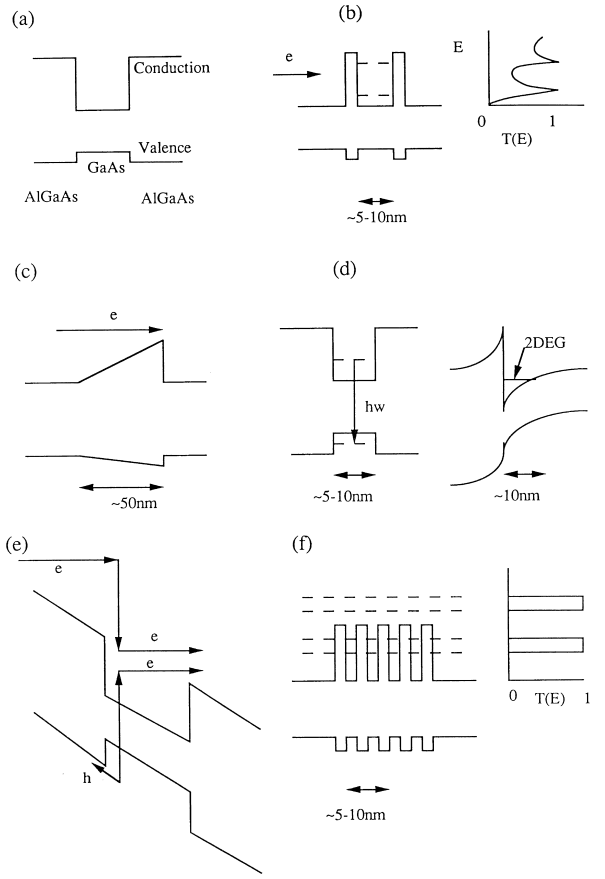


FIGURE 2 The physics from semiconductor multilayers. (a) Typical conduction and valence band profiles for the GaAs/AlGaAs materials system; (b) the double-barrier structure for resonant tunneling; (c) a graded gap layer for hot electron injection; (d) quantum confinement of electrons in a quantum well and at a heterojunction interface; (e) avalanching at a heterojunction in a multilayer structure; (f) energy filtering of electrons in a periodic multilayer (superlattice).

alloy is an indirect-gap semiconductor with more complicated energy bands, which also means that it is not optically active. At the abrupt interface between GaAs and $Al_xGa_{1-x}As$, about 65% of the band-gap discontinuity appears as a potential step in the conduction band. Thus for $x = 0.3$, we have a potential step of ~0.24 eV for electrons. This step can be used in several ways (see Fig. 2) to generate and exploit physics not available in bulk uniform semiconductors. Our discussion here is confined to the conduction bands and excess electrons in n-doped material: the equivalent physics applies to holes in valence bands but the energy band structure is much more complicated.

A. Tunneling

In classical mechanics, an incident electron is reflected by a potential barrier if its height is greater than the electron kinetic energy. In quantum mechanics, an electron has a small but finite probability of penetrating into the barrier. If the barrier is sufficiently thin (i.e., of the order of the relevant de Broglie wavelength in the barrier), the electron has a finite probability of emerging from the other side and the electron is said to have tunneled through the barrier. Tunneling has been studied widely for several decades: in many cases it is manifest as a type of leakage current through thin layers. Resonant tunneling is a more striking phenomenon. If two thin barriers are separated by a thin intervening layer (see Fig. 2b), there are discrete energies at which the electron transmission grows to unity: a careful analysis shows that there is a constructive interference between electrons traveling forward and those traveling backward in the central layer. If such a structure is made in the GaAs/AlGaAs materialsntem, the barrier layers typically have $x \sim 0.4$ and the barrier and central layers on either side are of the order of 3–10 nm thick. The GaAs layers on either side are doped to provide electrical contacts, and under an applied bias, electrons tunnel through the multilayer. When, under bias, electrons in the contacts have the same energy for the high transmission, the current rises (see Fig. 3). It falls for a further increase in bias, because the transmission falls, and so the structure provides a negative differential resistance, the basic prerequisite for oscillator circuits. Indeed these structures have operated at up to 420 GHz in the GaAs/AlAs materials system and up to 720 GHz in the InAs/GaSb system, which represents the fastest frequency achieved by a solid-state device. Variations on this simple structure have been used extensively to clarify the fundamentals of quantum transport of electrons and holes in semiconductors.

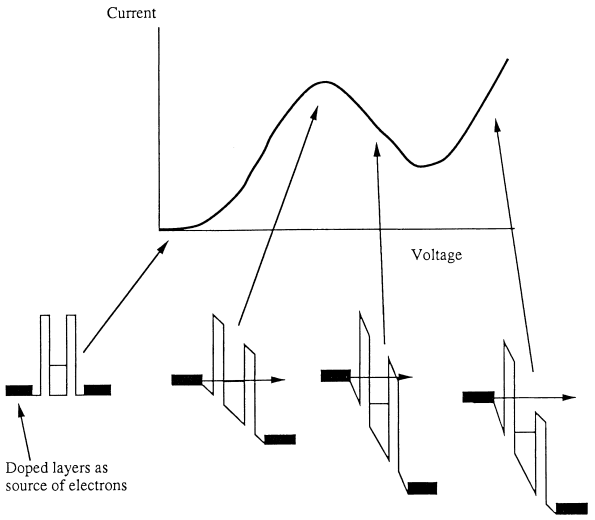


FIGURE 3 The current–voltage characteristics of a resonant tunnel diode.

If one of the two barriers in Fig. 3 is removed, we have an n–i–n structure with a tunnel barrier toward one end. This structure acts as a microwave detector diode (called an ASPAT diode) with a number of attractive properties including a low sensitivity to ambient temperature, a wide dynamic range, and low associated noise. However, this device depends very sensitively on the barrier parameters, and any roughness on interfaces contributes to an unacceptably wide spread in electrical performance.

B. Hot Electron Injection

With reference to Fig. 2c, which typically has the aluminium fraction rising from 0 to 40% over 50 nm, the tunneling probability through this barrier is very low. Under bias, an applied electric field can reduce the "quasi-electric field" represented by the rising conduction band edge (see Fig. 4). Even when the electrons are in the AlGaAs layers, they have the typical kinetic energy associated with thermal electrons, just as in regions away from the barrier. Once they cross the heterojunction they gain typically 0.3 eV as excess kinetic energy. Such electrons are called hot, as this excess energy is typical of that if they were in a solid at 3000 K. The potential energy the electrons receive as they move against the quasi-electric field is converted into forward kinetic energy, implying that the electron motion is strongly forward collimated. The transit time for such electrons over short distances is reduced by having crossed the heterojunction.

Heterostructures such as this have several device application. In heterojunction bipolar transistors the reduced transit time across the base makes for faster transistor action. The heterojunctions also allow heavier doping to be tolerated without unwanted minority carrier effects coming into play, and this also helps increase the speed. In the Gunn diode used as the source of microwave power, electrons are accelerated (heated) in a strong electric field so that peculiarities of the gallium arsenide band structure at $x > 0.3$ eV can be exploited. In a uniform piece of gallium arsenide, electrons with excess energy can lose that excess to the lattice. The structure in Fig. 4 allows the electrons to remain "cold" until they receive all their excess energy as they cross the heterojunction. The layer of graded composition alloy acts as a hot electron launcher and, in this mode of operation, has led to a new generation of Gunn diodes with greater efficiency, lower noise, and reduced sensitivity to operating temperature. The quest for a unipolar hot electron transistor has continued without success since the 1960s, and the hot electron launcher looked to be a promising candidate in the early 1980s. The way in which the hot electrons relax when injected into a heavily doped GaAs base has proved to be an interesting field of physics: in the process of such studies, the relaxation rate was found to be too fast for useful transistor action. Research has continued with other materials, particularly InAlAs/InGaAs, where high-speed transistor action and small-scale integration have been demonstrated: different band-gap offsets allow even higher-energy injection, and so more energy can be lost in the base while still being able to be collected.

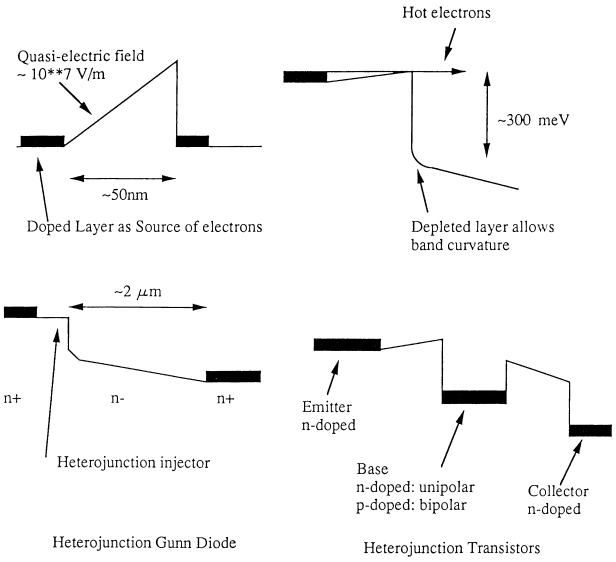


FIGURE 4 The hot electron injector as used in transistors and Gunn diodes.

C. Quantum Confinement

The two structures shown in Fig. 2d each form the basis of successful commercial devices, namely, the quantum well laser and the heterojunction field-effect transistor, which have superior properties by virtue of the quantum confinement of electrons into quasi-two-dimensional layers. The electron states in a thin layer sandwiched between layers of a wider band gap have a kinetic energy of confinement in the direction perpendicular to the layers of $\sim h^2/2m^*t^2$, where t is the thickness of the layer, m^* is the effective mass of the electrons in the semiconductor, and h is Planck's constant. This result is a simple application of the particle-in-a-box problem in elementary quantum mechanics. For $t = 10$ nm in gallium arsenide, this extra energy is of the order of 50 meV, an appreciable fraction of the potential step at the heterojunction. (There are added complications for thinner wells when this energy equals or exceeds the potential step, but these are understood.) For holes in the valence band, the appropriate hole mass must be used. Both electrons and holes are free to move in the plane of the layer, i.e., in two dimensions. Whereas in bulk semiconductors an electron and hole at the conduction and valence band edges could recombine and emit

a photon of light with an energy equal to the band gap of the semiconductor, here we see from the diagram that the extra kinetic energy of confinement for both electrons and holes must be added on. This was the first example of tailoring the optical properties of semiconductors, not by bulk crystal structure and chemistry, but by local structure via the thickness t. The additional energies are sufficient to decrease the wavelength of light emitted from a gallium arsenide layer from 0.86 to as little as 0.66 μm, i.e., from the infrared into the visible red. This principle and indeed other attractive, but technical, features of electrons in two dimensions have led to the current generation of quantum well lasers which have a greater efficiency, lower threshold currents, a reduced dependence on temperature, and shorter wavelengths than the more conventional lasers in gallium arsenide.

The second structure in Fig. 2d results from the growth of a wider-gap n-doped layer of semiconductor on a lower-gap layer of undoped semiconductor. Under thermal or optical activation, the donors give up their electrons as in bulk materials. Here it is possible for them to cross the heterojunction, beyond which they lose their ~0.25–0.3 eV of excess energy before coming into equilibrium. The potential step at the heterojunction prevents them from returning to their donor sites. At the same time the coulomb attraction between the positive donor ion and the electron remains. The electrons are thus trapped at the interface but are free to move parallel to it. The electrons are said to form a quasi-two-dimensional electron gas, which has been the basis of exciting physics arising from the strong reduction in coulomb scattering and the reduction of lattice scattering at low temperatures. Electrons at low temperatures can propagate macroscopic distances (~1 mm) without scattering: regarded as waves, they maintain phase memory over these distances. The low-temperature low-electric field mobility is as much as 1000 times greater than can be achieved in high-quality bulk semiconductors. The quantum Hall effect is seen in this system when a magnetic field is applied perpendicular to the layers: the Hall resistance (the lateral voltage induced by the magnetic field divided by the current) takes on the values of $h/e^2 i = 28.82\ldots/i$ kΩ, where h is Planck's constant, e the charge on the electron, and i an integer. This relation is held to a precision of 1 part in 10^8, so that it is now a resistance standard. A new and complex many-electron fluid is formed when very high magnetic fields are applied—the so-called fractional Hall regime. A range of quantum interference effects is also being discovered where electron currents are split and recombined all within a distance over which electron phase coherence is maintained. (See Section V.)

The same two-dimensional electron gas is at the heart of the high-electron mobility transistor (HEMT; see Fig. 5),

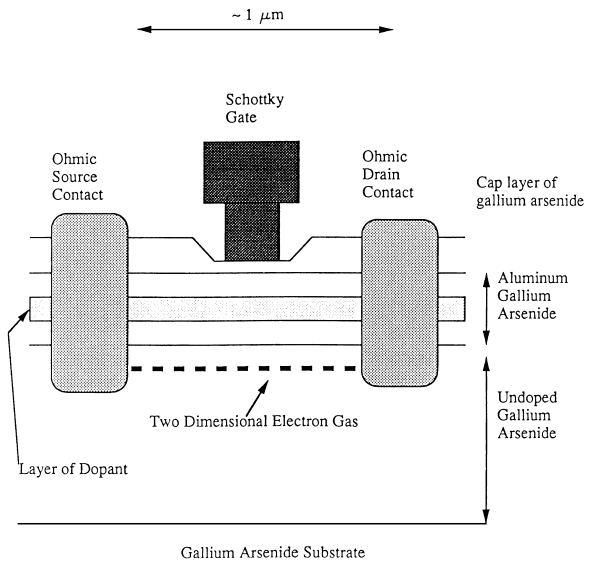


FIGURE 5 The high-electron mobility transistor.

which is a heterojunction version of the GaAs MESFET (metal–semiconductor field effect transistor, which uses a layer of bulk doped GaAs instead of the two-dimensional electron gas in Fig. 5). The process of placing the doping only in the wider-gap semiconductor and set away from the heterojunction is known as modulation doping. The new transistor has two advantages over the MESFET: (1) the parasitic resistances in the low-field regions not under the gate are lower, leading to higher-frequency operation if all other geometric features are the same; and (2) the noise of amplifier operation is much lower, due in part to the restricted final states into which two-dimensional electrons might be scattered. The HEMT is in widespread use as a low-noise amplifier.

In both the quantum well laser and the high-electron mobility transistor, the heterojunction between the two semiconductors is used as a new design feature to improve the device operation in terms of greater efficiency, lower noise, and reduced sensitivity to temperature.

IV. PHYSICS AND APPLICATIONS OF SUPERLATTICES

A. Multiple Quantum Wells

Quantum wells that are separated by more than about 15–20 nm are effectively noninteracting. A structure with typically 50–100 quantum wells over a thickness of 1–3 μm behaves as an unusual solid. For the optical absorption, the effects of each well add together. For transport we have effectively a two-dimensional solid, with easy conduction parallel to the wells but not perpendicular.

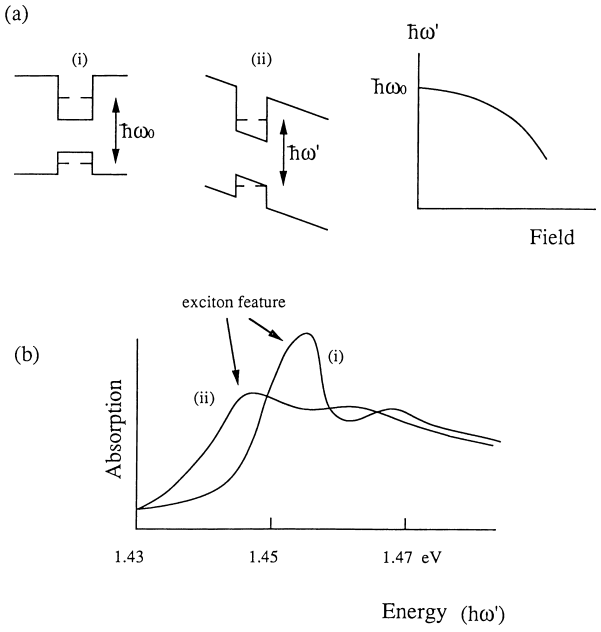


FIGURE 6 Optical properties of quantum wells in an electric field. (a) Effect of the field on electron and hole states and the principal absorption features; (b) schematic of actual absorption curves.

In multiple-quantum well lasers, the optical output is a complex trade-off between competing electronic effects associated with getting electrons and holes into the wells and optical effects: about five quantum wells is optimal. Under the application of an electric field normal to the layers, the optical absorption changes markedly in both strength and wavelength. This phenomenon, the quantum confined Stark effect, has important applications in optical modulation, and its origin is shown in Fig. 6. In elementary quantum mechanical terms, the applied electric field tilts the conduction and valance band edges as shown, and so the energy levels move. Since these levels depend on the effective mass (see Section III.C), and the hole and electron masses differ, the separation in energy levels for optical absorption will depend on the applied bias. If one works at a fixed wavelength, the absorption can be increased or decreased by the applied bais.

A second effect is also very important. An exciton is an excitation of a valence electron to the conduction band, leaving a positively charged hole behind. The coulomb attraction between the electron and the hole results in a bound-state energy lower than the conventional semiconductor band gap. Bound electron–hole pairs, excitons, share all the physics of a hydrogen atom scaled by the reduced effective mass of the electron and hole ($1/\mu = 1/m_e + 1/m_h$) and the dielectric constant (ε) of the semiconductor. In gallium arsenide, the exciton binding energy is only $E_b = \mu e^4/h^2\varepsilon^2 \sim 5$ meV, and its diameter is ~20 nm (compared with 13.6 eV and 0.1 nm for a hydrogen atom). Excitons are formed in quantum wells, but if the well is less than 20 nm thick, the electron and hole wave functions must fit within the quantum well: the electron and hole are constrained to be closer to each other than is the case in a bulk semiconductor, and the exciton binding energy increases by as much as a factor of four. Whereas thermal energy at room temperature is sufficient to ionize a bulk exciton, quantum well excitonic absorption is still strong. In practical applications it is the exciton absorption that is moved by an applied bias. Another attractive feature of excitons in quantum wells is that the heterojunctions confine the electrons and holes in the direction perpendicular to the layers. Quite modest electric fields applied to a bulk exciton result in field ionization, with the electrons and holes moving apart, but up to 40 times these fields can be applied to a quantum well exciton before such ionization occurs, and electrons tunnel out of the quantum well. Note that if the field is applied in the plane of the quantum well, ionization still occurs easily at low fields.

If a multiple-quantum well structure is built into the intrinsic (undoped) region of a p–i–n diode), an applied bias can be used to modulate the optical absorption. If this diode is placed in a feedback circuit with a series resistance, the excited electrons can in turn lead to a photocurrent which then alters the bias across the diode and so modifies the optical absorption. Under appropriate conditions, optical bistability between states of high and states of low absorption for the same incident light intensity and controlled switching between such states can be achieved. Neither effect occurs in bulk semiconductors. There are near-term applications in optical signal transmission and modulation using multiple quantum wells.

Under very strong electric fields, some electrons are excited into the conduction bands with sufficiently high energy that they can excite another electron across the band gap and still remain in the conduction band. This process is known as impact ionization and avalanching (because of the rapid multiplication over distance of the number of electrons and holes) and is the basis of both powerful solid-state microwave sources and sensitive photodetectors. The spatial localization of the ionization process is not controlled, and these devices have undesirable noise levels as a result. In Fig. 2e we show the impact ionization in a quantum well, where the sudden increase in electron energy at the heterojunction can result in local impact ionization. This localisation would reduce the noise: such effects are seen for photodetectors, although they are not as dramatic as had been hoped—the band-gap energy difference is only a small fraction of the excess energy required for ionization.

B. Superlattices

When the barriers between adjacent quantum wells are sufficiently thin, electrons in bound states in one well can tunnel into adjacent wells. For two identical wells separated by a barrier only two or three atomic layers thick, the interaction is strong enough to induce analogues of the hydrogen molecule, namely, symmetric and antisymmetric combinations of the wave functions from adjacent wells, high polarizability under electric fields, etc. With a periodic repetition of many GaAs wells and 1- to 5-nm AlGaAs alloy barriers, a completely new crystal structure, a *superlattice*, and electronic band structure results. The electron propagation in the plane of the quantum wells is unaffected, but in the direction of the multilayer growth it is very sensitive to the barrier and well parameters. Because the repeat distance of the crystal is several layers, the energy bands are generally narrow and are known as minibands. They range from 1 to150 meV. Viewed as a resonant tunneling system, the transmission of electrons through a superlattice is either near unity, if the electron is in a miniband, or zero otherwise (see Fig. 2f). Doped superlattices act like metals with partially filled energy bands. The characteristic distances in superlattices are comparable with the electron wavelength so there are electron analogues of antireflection coatings and optical filers can be realized. It was the possibility of new and tailored energy bands, and the prediction in 1970 of a novel form of negative differential resistance in superlattices under applied electric fields, that set off the whole scientific research enterprise described in this article. Only in the last decade has this effect been realized in practice, and it has proven to be rather delicate (i.e., susceptible to destruction by thermal vibrations above low temperatures or by imperfections in the crystals). The effect is very fast, operating at 1–3 tHz, 10 times faster then the resonant tunnel diode.

V. MESOSCOPIC SYSTEMS

The previous topics have concentrated on quasi-two-dimensional physics in multilayers. During the 1990s much effort has concentrated on defining patterns in these layers using electron beam lithography to achieve the same length scales. From quantum wells one progresses to quantum wires and quantum dots, where carriers are constrained in two and all three spatial dimensions, respectively. In such small structures, only a few electrons are present, and they are known as mesoscopic systems (meso = middle, i.e., between macroscopic and microscopic). These systems to date suffer from very large

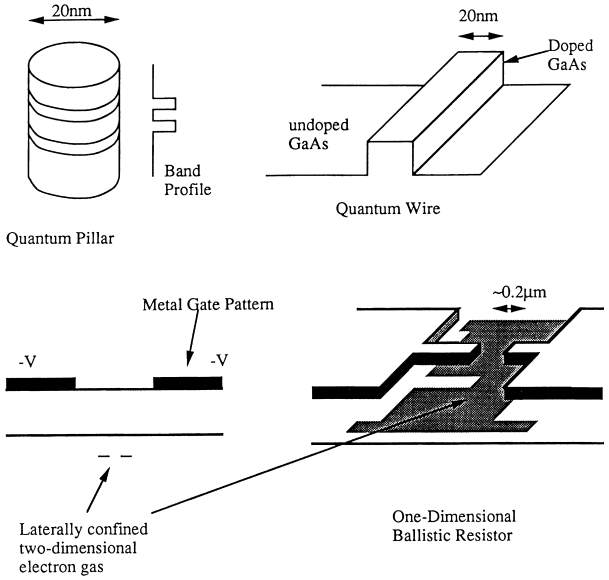


FIGURE 7 A collage of mesoscopic systems.

fluctuations in conductance, optical absorption, etc., as the detailed control over size is not adequate, and when structures get small there are intrinsic statistical variations of small numbers of atoms and electrons. The predicted advantages of quantum dots in laser structures are all lost if the size fluctuations are too great. With nanoampere currents in structures less than 0.1 μm in diameter and $\sim$1 μm long, there can often be less than one transport or optically excited electron present (time averaged) in these structures.

One fertile area of research is based on the so-called split-gate technology—a patterned gate (see Fig. 7) over a two-dimensional electron gas at a high-quality buried heterojunction. A negative bias on the gate depletes the carriers underneath, and where there is a split in the gate a quasi-one-dimensional conducting wire is formed. As electrons can move several microns without scattering, a short split-gate structure can give rise to ballistic motion of electrons in the wire. This is the ballistic resistor mentioned in Section I, taking the resistance value of $h/2e^2$ divided by an integer. This integer is given by the number of quasi-one-dimensional energy bands that are occupied. An increasing gate bias narrows the channel, increases the quantum energy of confinement in the lateral direction, reduces the number of occupied subbands, and increases the resistance in sharp steps. Under a longitudinal bias a new form of negative differential resistance has been predicted and observed. With more elaborate gate patterns it is possible to split an electric current in two and recombine it within 1 μm and see quantum interference phenomena (the so-called Aharonov–Bohm effect)

whereby a magnetic field perpendicular to the electron gas induces different phase differences in the two arms of current.

The quantum pillar in Fig. 7, made from resonant tunneling multilayers, has a capacitance of the order of 10^{-18} F, so that charging by even one electron requires a bias on the scale of a volt. This is the co-called coulomb blockade regime. There is now evidence of electrons passing through this structure in a correlated manner—the noise is below the shot noise levels that would apply if the electrons were not correlated.

VI. THE COMING DECADE

In recent years, the rate of discovery of qualitatively new phenomena associated with low-dimensional structures has slowed, and those that are discovered are more delicate (i.e., observed only at very low temperatures and under very high magnetic fields). The emphasis has shifted to extending the range of new materials under investigation. Gallium nitride crystals have now been grown that act as light-emitting diodes and lasers in the blue. This material has a higher dielectric strength (i.e., can withstand larger fields before avalanche breakdown) and the electrons have a higher saturated drift velocity (i.e., maximum speed under bias) than the other group III–V semiconductors. These are the materials parameters that determine transistor performance, and so faster and more power transistors are being developed. Metallic multilayers are giving rise to new magnetic structures and giant magnetoresistance effects that are exploited in magnetic recording. The range of mesoscopic systems under investigation is ever widening. Greater emphasis will be placed on control of these systems, either during fabrication or with clever circuit configurations, so that some applications may be realized. Finally, the advances in polymer-based electronic and optoelectronic devices are exploiting heterojunction concepts.

SEE ALSO THE FOLLOWING ARTICLES

CRYSTAL GROWTH • EPITAXIAL TECHNOLOGY FOR INTEGRATED CIRCUIT MANUFACTURING • EXCITONS, SEMICONDUCTOR • INTEGRATED CIRCUIT MANUFACTURE • MOLECULAR BEAM EPITAXY, SEMICONDUCTORS • SEMICONDUCTOR ALLOYS • THIN-FILM TRANSISTORS

BIBLIOGRAPHY

Kelly, M. J. (1995). "Low Dimensional Semiconductors," Clarendon Press, Oxford.

Luryi, S. A., Xu, J., and Zaslavsky, A. (eds.) (1999). "Future Trends in Microelectronics: The Road Ahead," Wiley, New York.

Sze, S. M. (ed.) (1998). "Modern Semiconductor Device Physics," Wiley, New York.



Superconducting Devices

Jukka Pekola
University of Jyväskylä

I. Introduction, On-Chip Superconducting Devices
II. SQUIDs
III. Thermal Detectors
IV. NIS Tunneling Devices
V. NIS Microcooling
VI. Superconducting Single Charge Devices

GLOSSARY

Cooper pair Current carrying condensate of a superconductor formed of pairs of electrons.
Josephson junction Insulating barrier separates two superconductors. Cooper pairs can tunnel through.
NIS Normal metal—superconductor—normal metal tunnel junction, which can be used for sensitive thermometry and for on-chip cooling.
RSFQ Rapid Single Flux Quantum devices for ultrafast superconducting digital electronics.
Single cooper pair devices New set of devices based on transport and storage of single charges. Possible digital, sensor, and quantum computing applications.
SQUID Superconducting Quantum Interference Device, a sensitive detector of magnetic flux and its derivatives.
TES Transition Edge Sensor. Usually a transition of superconductivity is used for sensitive energy and power measuring.

I. INTRODUCTION, ON-CHIP SUPERCONDUCTING DEVICES

In this article we deal with superconducting *on-chip* and mainly *planar* devices. Thus we do not discuss bulky, nonplanar components or instruments, such as cables, magnets, motors, etc.

One of the main ingredients in superconducting electronics and devices is the Josephson tunnel junction, and the superconducting quantum interference device (SQUID), with Josephson junctions forming essential parts of it. Thus, much of the history of superconducting electronics and a good fraction of this article also deal with SQUIDs, either as devices as such, or as sensitive detectors integrated in the device. Josephson junctions, forming the key details in the SQUIDs, have been in the focus of even large industrial programs to investigate and promote the possibilities of superconducting electronics.

In the present text we discuss, besides Josephson junctions and SQUIDs in their traditional framework, Josephson-junction-based rapid single flux quantum devices (RSFQs), thermal radiation detectors based on superconductivity properties, normal-insulator-superconductor tunneling in thermometry and in superconducting on-chip coolers, and future devices based on single Cooper pair tunneling in very small Josephson junctions.

II. SQUIDS

A. Josephson Junctions

There are principally two different kinds of SQUIDs: superconducting loops interrupted by either one or two Josephson junctions. These SQUIDs are called, mainly for historical reasons, an rf SQUID and a dc SQUID, respectively. Before detailing the SQUID operation, let us discuss the basic properties of a Josephson junction. A Josephson junction consists of two superconductors separated by a thin insulating barrier. The two superconductors both have their macroscopic wavefunctions with phases φ_i, $i = 1, 2$, and the phase difference $\varphi = \varphi_1 - \varphi_2$ determines the current I_s, i.e., the flow rate of Cooper pairs through the barrier as

$$I_s = I_c \sin\varphi, \tag{2.1}$$

where I_c is the critical current of the junction, which is the maximum supercurrent that the junction can sustain without dissipation. Upon increasing the current I through the junction from zero up to $I > I_c$, there is initially no voltage across the junction. But at about $I = I_c$ voltage V appears, and φ starts to evolve with time according to

$$\dot{\varphi} = 2eV/\hbar = 2\pi V/\Phi_0, \tag{2.2}$$

where $\Phi_0 = \hbar/2e \simeq 2.07 \cdot 10^{-15}$ Wb is the so called flux quantum.

To avoid the inevitable, and in usual operation harmful hysteresis in sweeping the current from zero up to above I_c and back to zero again (voltage does not drop back to zero at I_c on return but at current much below I_c), one typically shunts the junction by an external resistance R. In this case, the cross dynamics of the junction can be described according to the so-called "resistively and capacitively shunted junction" (RCSJ) model, as depicted in Fig. 1A. In this model an ideal junction [Eq. (2.1)], is connected in parallel with its self capacitance C and the shunt resistance R.

Including the noise current I_N, which arises from the shunt resistance, we can write the equation of motion of the junction as

$$I_c \sin\varphi + C\dot{V} + V/R = I + I_N, \tag{2.3}$$

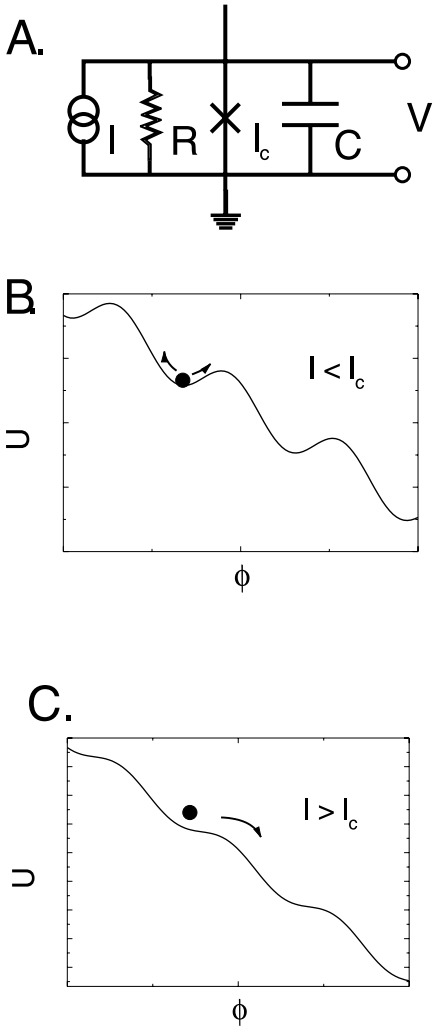


FIGURE 1 The resistively and capacitively shunted Josephson junction (RCSJ) model. The RSJ circuit is seen in (A), the cross depicts the Josephson junction itself. The tilted washboard model for $I < I_c$ in (B) and for $I > I_c$ in (C).

where now I is the total current through the junction-shunt system. Neglecting I_N and making use of Eq. (2.2), we have

$$\frac{\hbar C}{2e}\ddot{\varphi} + \frac{\hbar}{2eR}\dot{\varphi} = -\frac{2e}{\hbar}\frac{\partial U}{\partial \varphi}. \tag{2.4}$$

Here $U = -\frac{\Phi_0}{2\pi}(I\varphi + I_c\cos\varphi)$ acts as a potential for the variable φ. This model [Eq. (2.4] presents an analogy between the phase difference and a particle in the tilted washboard potential U, where the average slope of the potential is proportional to $-I$.

When $I < I_c$ (Fig. 1B), the "particle" is confined to one of the wells, and oscillates therein at the plasma (angular) frequency $\omega_p = \sqrt{2\pi I_c/\Phi_0 C}[1 - (I/I_c)^2]^{1/4}$. Then the average of $\dot{\varphi}$ and hence the average voltage across the junction are zero. When I exceeds I_c (Fig. 1C), the

"particle" starts to roll down the washboard, the average of $\dot{\varphi}$ and the average voltage across the junction are nonzero. This is the dissipative regime of the RCSJ. If the junction is nonhysteretic, the "particle" gets trapped when current drops below the critical value, and V drops to zero again. In order for this to happen we require $\beta_c \equiv (2\pi I_c R/\Phi_0)RC \equiv \omega_J RC \leq 1$. Here ω_J is the Josephson (angular) frequency at the voltage $I_c R$.

One of the consequences of Eqs. (2.1) and (2.2) is that when the Josephson junction is subjected to radiofrequency (rf) irradiation, there are so-called Shapiro steps in the current voltage characteristics (IVC) at voltages $V = nhf/2e$, where n is an integer and f is the frequency. This has made it possible, among other things, to develop a *Josephson volt standard* based on a series connection of more than 10,000 nonhysteretic junctions. The present definition of volt (V) in the SI system of units is based on this Josephson standard.

B. The dc SQUID

In a dc SQUID (Fig. 2) the two junctions interrupting the superconducting loop are shunted resistively as described above, this way eliminating the hysteresis of the current voltage curve. The IVC of the dc SQUID depends on the external flux Φ_{ext} applied to the loop, as shown schematically for two values of Φ_{ext} in Fig. 2B.

The critical current of the dc SQUID, I_m depends periodically on the flux threading it as $I_m = 2I_c|\cos(\pi\Phi/\Phi_0)|$. Here we assume identical junctions with critical current of I_c for both. If the SQUID is biased at a current I_b above $2I_c$, which is the maximum supercurrent through the SQUID, the voltage will therefore oscillate as a function of Φ with period Φ_0 (Fig. 2C). When operated in the steep part of the $V - \Phi_{ext}$ curve, the SQUID produces an output voltage in response to a small flux $\delta\Phi_{ext} \ll \Phi_0$. It is thus essentially a flux-to-voltage converter.

C. The rf SQUID

The one junction loop, the rf SQUID, can also be employed as a sensitive detector of flux. Assume a loop with inductance L, and a junction in it with critical current I_c again. The flux threading the loop, Φ, is then, by adding up the total flux due to external flux and the flux due to the circulating current, given by

$$\Phi = \Phi_{ext} - LI. \tag{2.5}$$

Here I is the current circulating in the loop. The Faraday induction law, $V = -d\Phi/dt$ for the almost closed SQUID loop, yields a relation between the magnetic flux and the Josephson phase difference

$$\phi = 2\pi\Phi/\Phi_0 \tag{2.6}$$

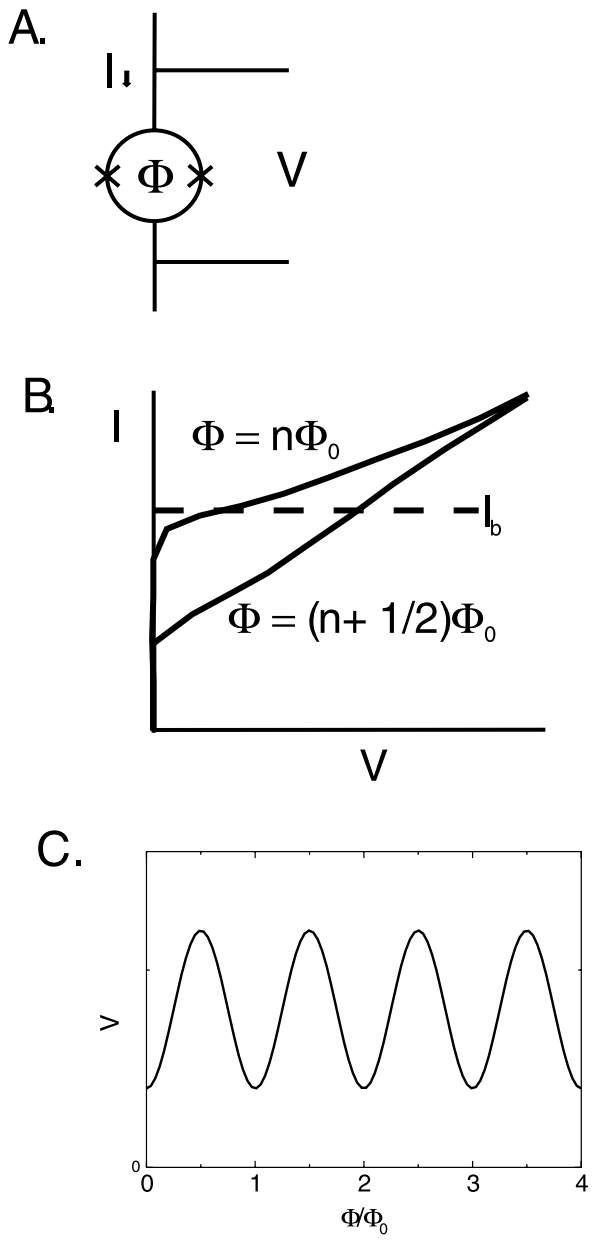


FIGURE 2 The dc SQUID. The SQUID loop is shown in (A); the junctions are shunted resistively as described above, but this has not been shown in the figure. Schematic current (I)-voltage (V) characteristics at two different values of the applied external magnetic flux Φ_{ext} in (B). V at a constant bias current $I > 2I_c$ in (C).

by integration of the ac Josephson relation [Eq. (2.2)]. Using the dc Josephson relation [Eq. (2.1)] and combining it with Eqs. (2.5) and (2.6) we finally obtain

$$\Phi + LI_c \sin\left(2\pi\frac{\Phi}{\Phi_0}\right) = \Phi_{ext}. \tag{2.7}$$

Equation (2.7) presents a nonlinear dependence of the flux threading the loop as a function of the external flux. In particular, for parameter values $LI_c \geq \Phi_0/2\pi$ the $\Phi_{ext} - \Phi$

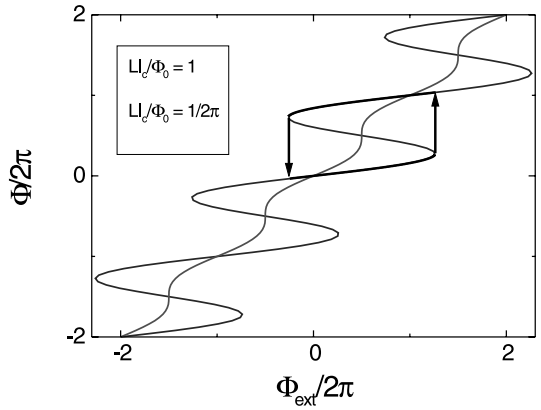


FIGURE 3 The flux Φ through a SQUID loop against the applied magnetic flux Φ_{ext} at two different values of the loop and junction parameters, $LI_c/\Phi_0 = 1/2\pi$ and $LI_c/\Phi_0 = 1$. In the latter case the backbending of the curve induces hysteresis. A typical hysteresis loop on excursion of the external flux is indicated by arrows.

characteristics show re-entrant behavior, and the dynamics are hysteretic on changing Φ_{ext} (see Fig. 3).

The sections with positive slope are stable and the ones with negative slope are unstable. With $LI_c \simeq \Phi_0$ the energy dissipated upon traversing the simplest hysteresis loop indicated in Fig. 3 is about $I_c\Phi_0$.

The radio frequency (rf) operation of the SQUID means connecting an RLC tank circuit with inductive coupling to the SQUID loop (see Fig. 4).

The detection of flux is possible, since the (amplitude of the) current in the inductor of the tank circuit at which the hysteresis loop first gets traversed depends on the bias point, i.e., the quasistatic value of Φ_{ext}. When the flux becomes hysteretic it can be detected as energy loss, i.e., change of rf voltage on the tank circuit.

D. Noise in SQUIDs

It is typical to express the noise in SQUIDs in terms of flux noise spectral density $S_\Phi(f)$, where f is frequency. The most useful quantity indicating the flux noise, $S_\Phi(f)^{1/2}$, is usually expressed in units $\Phi_0/\sqrt{\text{Hz}}$. A convenient way of characterizing the flux noise is the noise energy per unit bandwidth, $\varepsilon(f) = S_\Phi(f)/2L$, which, in turn, can be expressed in units $\hbar$/Hz.

Present day dc SQUIDs operating at 4 K typically have flux noise of order $S_\Phi(f)^{1/2} \simeq 10^{-6}\Phi_0/\sqrt{\text{Hz}}$, and $\varepsilon(f) \sim 10^2\hbar$. By cooling the dc SQUID and the shunt resistors down to well below 1 K, one can achieve $\varepsilon(f)$ of less than $10\,\hbar$. One limitation of dc SQUIDs is the $1/f$ noise, which depends on the quality and material of the Josephson junctions. $1/f$ noise is an issue especially in biomagnetic measurements where good resolution at frequencies down to below 0.1 Hz is required.

In contrast to the dc SQUID, noise in the rf SQUID depends also on the pre-amplifier and the tank circuit. Today rf SQUIDs can, when operated at >1 GHz frequencies, achieve flux noise of the order of few times $10^{-6}\Phi_0/\sqrt{\text{Hz}}$. The corresponding noise energy is typically $\varepsilon(f) \sim 10^3\hbar$, which can be improved, however, by using a cooled pre-amplifier.

E. SQUID Instruments

The SQUIDs can be used in a large variety of ultrasensitive measurements where the signal to be detected can be converted into magnetic flux. The basic measurement by a SQUID is to measure magnetic field. A schematic illustration of a dc SQUID as a *magnetometer* is shown in Fig. 5A, where a flux transformer circuit with a pick-up loop is connected to the input coil to make a superconducting flux transformer. The inductances of the pick-up coil

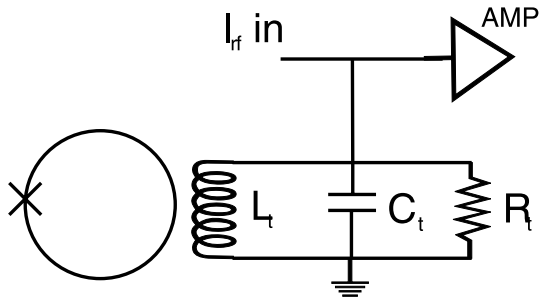


FIGURE 4 An LRC resonant tank circuit coupled via the inductance L_t to an rf SQUID.

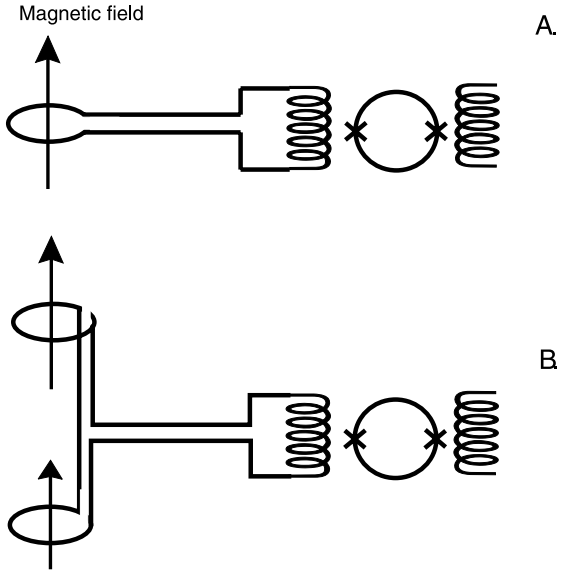


FIGURE 5 SQUID in measuring magnetic fields. Flux transformer as a magnetometer in (A) and as a first-derivative gradiometer in (B).

and the input coil determine the flux into the SQUID loop corresponding to the change in magnetic field at the measurement position (pick-up loop). The third coil is used to set the optimum bias point of the device, and to operate in the flux-locked mode. SQUID-magnetometers are used in geophysics and in various research purposes. The flux transformers have typically involved loops made of Nb. A typical noise level of a superconducting magnetometer is of the order of 10^{-11} mT/$\sqrt{\text{Hz}}$.

The next step is to measure spatially varying magnetic fields. A *gradiometer* measures basically the gradient of magnetic field along the axis of a pick-up coil with two oppositely wound loops (Fig. 5B). With a uniform axial field the gradiometric measurement yields a zero signal if the coils are equal. The input coil and the SQUID operate like a magnetometer. Highly symmetric gradiometer coils can be made by thin film techniques. An important application of gradiometry is biomagnetism, where spatially varying magnetic fields from local sources (e.g., human brain or heart) can be discriminated against virtually uniform fields from distant sources.

A gradiometer set-up of Fig. 5 can be used as a *susceptometer* by inserting the sample whose susceptibility χ is to be determined inside one of the halves of the pick-up coil. The signal of the susceptometer is proportional to χ in the presence of a static uniform magnetic field. SQUID-susceptometers are available commercially.

SQUIDs can be easily configured as very sensitive *ammeters* or *voltmeters* also. For example, in measuring the current through a voltage-biased superconducting transition edge desensor, TES, as will be described in Sections III.C and D, the current is allowed to pass through a series connection of several (in some cases 100) input coils of individual dc SQUIDs; this way the current to be determined is transformed into flux through the SQUID loop.

A variation of susceptometry is the measurement of *displacement*. Position of a magnetic object in the vicinity of one-half of the pick-up coil can be detected with high accuracy due to the change of the "filling factor" of this object with nonzero χ. This way, besides conventional magnetic objects, position of a superconductor with perfect diamagnetism ($\chi = -1$) or of a nonmagnetic object attached to a magnetic one can be measured.

F. RSFQ, Principle and Devices

Binary "0" and "1" are often realized by latching circuits in semiconductor logics. This was how researchers attempted to mimic the realizations of superconducting electronics in the 1970s and 1980s. With superconductors, in particular in Josephson junctions, this unnatural choice of representing the logic states corresponds to the two voltage states of the junction when current through it is below the critical value I_c. The "0" could be presented by the supercurrent branch, where current flows with zero voltage, whereas "1" corresponds to the dissipative branch with voltage $V \simeq 2\Delta/e$. Because of the complicated dynamics of the junction, especially upon reducing the current, the speed of switching is limited to a few GHz in such devices. Latching circuits also lead to severe cross-talk problems when integrated. Therefore, two extensive programs to develop and promote superconducting electronics, one by IBM in the United States (1969–1983) and the second one, MITI, in Japan (1981–1990) were terminated without commercial continuation, mainly not because of unsuitability of the technology, but because of low operation frequency due to the principally hopeless attempt to operate Josephson junctions in an analogous mode as semiconductor-based digital electronics. Comparable performance of superconductor-based electronics could not justify their use against conventional semiconductor technology: cooling of superconductors to helium temperature (4.2 K) is a handicap anyway.

The development of RSFQs, or alternatively dynamic single flux quantum (SFQ) digital technology presents a new attempt to develop superconductor-based fast digital electronics. The speed of the Josephson junctions can be fully benefited in the SFQ devices. Besides the speed of the active device, one should realize that superconductors also provide the fastest passive transmission lines across the chip essentially without cross talk. This is made possible by the fact that superconductors provide very little AC-attenuation up to the frequency $f = 2\Delta/h$, which is of the order of 700 GHz for Nb, the most common superconductor in planar technology. (Above this frequency attenuation increases due to Cooper pair breaking.) The low cross talk, in turn, is guaranteed by the very thin insulator gap between the conductor and the ground plane allowed by the short penetration depth in a superconductor, the *London depth* which is $\sim 0.1\ \mu$m for Nb.

Single flux quantum logic is based again on the fundamental property of superconductors to quantize magnetic flux through a loop in multiples of the flux quantum. The most elementary example of an SFQ circuit is provided by the rf SQUID with external current input I_{ext}, see Fig. 6. The external flux $\Phi_{\text{ext}} = MI_{\text{ext}}$, induced by the inductance M between the current path and the SQUID loop, affects the total flux Φ through the loop according to Eq. (2.5). If we now return to Fig. 3, with $LI_c = 1$, we see that with a flux bias of $\Phi_{\text{ext}} \simeq \Phi_0/2$, we have two flux states of equal stability, separated from each other by $\delta\Phi \simeq \Phi_0$.

The "quantization" of the flux difference above is accurate only for very inductive SQUID loops, $LI_c \gg 1$, but it holds approximately also for $LI_c = 1$, in which case only two flux states are simultaneously stable. The RSFQ logics are fundamentally based on this switching between

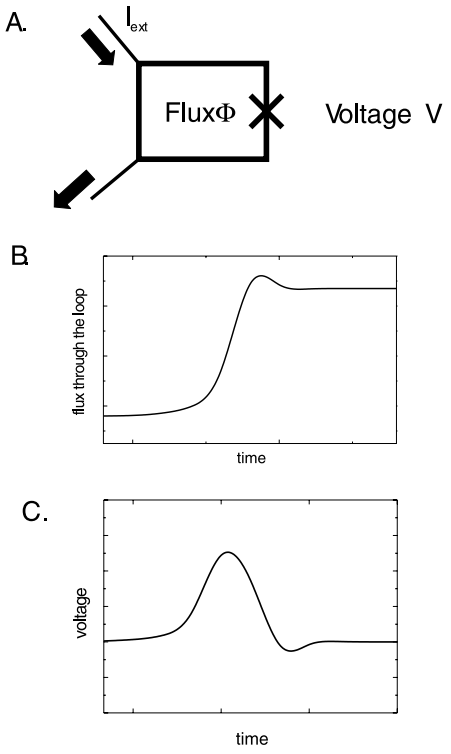


FIGURE 6 RSFQ. The simplest SFQ circuit, an rf SQUID, with an external current input I_{ext} to couple flux into the loop in (A). Schematic presentation of the time evolution of flux through the loop (B) and voltage across the junction (C) when the external current reaches the threshold value corresponding to different flux branches of Fig. 3.

the two neighboring flux states in the SQUID loop. During this switching, induced, for example, by a current pulse taking the Φ_{ext} beyond the stability limit in Fig. 3, a short voltage $V(t)$ pulse, typically a few picoseconds long, is formed across the junction. By Faraday induction law we obtain that the time integral of the pulse is

$$\int V(t)\,dt \simeq \Phi_0 \simeq 2\ \text{mV} \cdot \text{ps}. \tag{2.8}$$

Digital bits are thus coded in statics by the single quanta of magnetic flux, whereas in dynamics the data are transferred as picosecond SFQ voltage pulses with quantized area. It is expected that RSFQ-based VLSI circuits may operate at clock frequencies well above 100 GHz, which is at least an order of magnitude higher than those of the semiconductor-based devices fabricated with similar design rules.

The high speed of the RSFQ circuits makes them attractive for special applications, despite the psychological barrier posed by the necessity of operating them at helium temperatures. Such recent practical devices include analog-to-digital and digital-to-analog converters in radar technology, digital SQUIDs in supersensitive magnetometry, and digital autocorrelators in radio astronomy. It should be emphasized that no true commercial RSFQ system exists at the moment. Within a decade goals to implement RSFQ techniques include digital signal processing and petaflop-scale computing.

III. THERMAL DETECTORS

A. General

Thermal sensing of radiation in both X-ray and submillimeter wave ranges is a technique that is being actively developed presently. Essential elements of a thermal detector are the absorber of radiation and the ultrasensitive temperature sensor. Superconductors and semiconductors are generically well suited as base materials for temperature measuring, because they exhibit strong, sometimes even abrupt, changes in their electrical conductance properties against temperature. In this article we focus on two different types of superconductor-based thermometers: TES and normal metal-insulator-superconductor (NIS) junctions. Recent advances in microlithography, micromachining, and sub-kelvin cooling techniques have made it possible to make integrated detectors with ultrahigh resolution.

B. TES for Microcalorimetry and Microbolometry

A natural choice for a very sensitive thermometer in a narrow temperature interval is that based on a phase transition in a solid (TES). To avoid hysteresis in the transition one should restrict to second order transitions with vanishing latent heat. Normal to superconductor transition of a metal film is a good candidate. The resistance at this transition has been used to measure small temperature excursions over many years. In a textbook the transition to the superconducting state is infinitely sharp at the transition temperature $T = T_c$, but in practical samples it is spread over a small temperature interval due to material inhomogeneities, stress, and dimensional effects. The width of the transition is characterized by the parameter α, defined by

$$\alpha \equiv d \ln R / d \ln T. \tag{3.1}$$

Here R is the temperature dependent resistance of the film. By this definition α is a dimensionless parameter which gives the inverse of the relative temperature width of the transition. Thin film transitions have typically α's in the range from several tens up to several thousands.

Besides α, or the temperature width of the transition, ΔT, another parameter characterizing the TES is the mean transition temperature T_c, defined usually as the temperature where R has dropped down to one-half of its value in

the normal state just above the transition. With this definition we can write $\Delta T \simeq T_c/\alpha$. In addition to high α and practically negligible hysteresis, another advantage of a superconducting TES is that there are a number of ways either to select or tune T_c of the detector. This is necessary because the thermometer should work at a designed temperature interval and the superconducting TES has no sensitivity ($dR/dT \equiv 0$) below the transition and almost no sensitivity (dR/dT positive but very small) in the normal state above the transition.

A traditional way of tuning T_c is to apply an external magnetic field to the metal sample. This is, in many cases inconvenient, e.g., because it involves extra apparatus for producing this field. A recently developed method to set the transition temperature by fabrication is the implantation of magnetic ions to otherwise clean metal films. Magnetic impurities lower the T_c of the virgin metal film. By this method T_c can be tuned accurately, for example, with $^{56}Fe^+$ ion implantation in W films transition temperature can be tuned highly reproducibly in the range of 40–140 mK.

The most widely used method of tuning T_c is to make use of the superconducting proximity effect in metallic thin film bilayers. In the proximity effect a normal metal in good metallic contact lowers the transition temperature of the composite film from that of the bare superconductor component. This method also yields relatively predictable and reproducible T_c's but the method is not free from material problems. The key issue is the quality of the interface between the two metals. An involved theory of the proximity effect has been developed. The results are readily comparable to the measured T_c's without uncertain fit parameters in the case of an absolutely clean and abrupt interface between the metals. In practice, however, the metals tend to interdiffuse forming an alloy layer, or the interface between them cannot be made into a low-resistance metallic contact from the outset: in the worst case it can be a Schottky or a tunnel barrier. For example, Al/Cu bilayer results in a predictable T_c initially right after its fabrication, but these thin films form an alloy in just a few weeks' time in the ambient air. Yet stable material pairs are provided by Ti/Au, Mo/Cu, and Al/Ag with T_c's in the 0.1 K range.

C. Thermal Properties and Operation of TES in a Calorimeter

The thermal properties of a TES detector are largely characterized by two quantities. The first one is the thermal conductance G between the detector at temperature T_s and the heat bath at temperature T_0. In steady state G can be defined as $G \equiv P/(T_s - T_0)$, where P is the (small) constant input power to the detector. The second one is the heat capacity C of the absorbing element, which is directly integrated to the detector itself. The natural thermal time constant τ of the detector is $\tau = C/G$. The parameters of the detector are chosen based on the needs of the calorimetric detection. For example, in single X-ray quantum detection, C has to be set such that the initial rise during the temperature excursion, $\delta T = E/C$, after the hit of the X-ray quantum of energy E on the absorber does not bring the TES into saturation. This condition can be written as $\delta T \ll \Delta T$. Another design consideration of a calorimeter comes from its speed: τ has to be much shorter than the inverse of the count rate of X-ray pulses.

The measurement of a superconducting TES is best done in the constant voltage bias mode. This is because of the positive temperature coefficient (PTC) of the detector at the transition, $dR/dT > 0$. If, instead, the TES with a PTC is set into constant current bias, the sensor can experience thermal runaway and get damaged eventually. In other words, a sudden temperature rise leads to increased voltage and power for a detector with PTC and current bias. With constant voltage bias, increase of temperature leads to decrease in current and consequently the power dissipation diminishes and temperature tends to drop again: the system is thermally stable. Furthermore, in voltage bias the detector becomes faster than what the natural thermal time constant C/G might expect. This is due to the so-called *electrothermal feedback*.

D. TES X-Ray Calorimeter

Transition edge detectors can be used in soft X-ray calorimeters, e.g., in astronomy or elemental microanalysis. A typical TES X-ray detector is shown schematically in Fig. 7.

The TES is grown on a thin silicon nitride membrane window with bulk silicon etched away in this area. This window, often further sliced into a mesh, provides low G between the TES film and the heat bath of the bulk silicon outside the window. The TES film is electrically connected to its bias circuitry via thin film leads made of a higher T_c superconductor, like Nb, to avoid electrical resistance and to eliminate thermal conduction along them. The *quantum efficiency*, i.e., the relative number of detected X-ray quanta is very low if the bilayer TES film also acts as the absorber: the film is thin and often made of not very heavy metals. To increase the quantum efficiency from the level of only a few per cent close to unity, one typically covers the TES film by a relatively thick, 1–3 μm, Bi film. The advantages of using Bi are that it does not strongly influence the proximity effect of the underlying bilayer because Bi is a semimetal, it does not increase the heat capacity too much, and this way the detector will not be too slow.

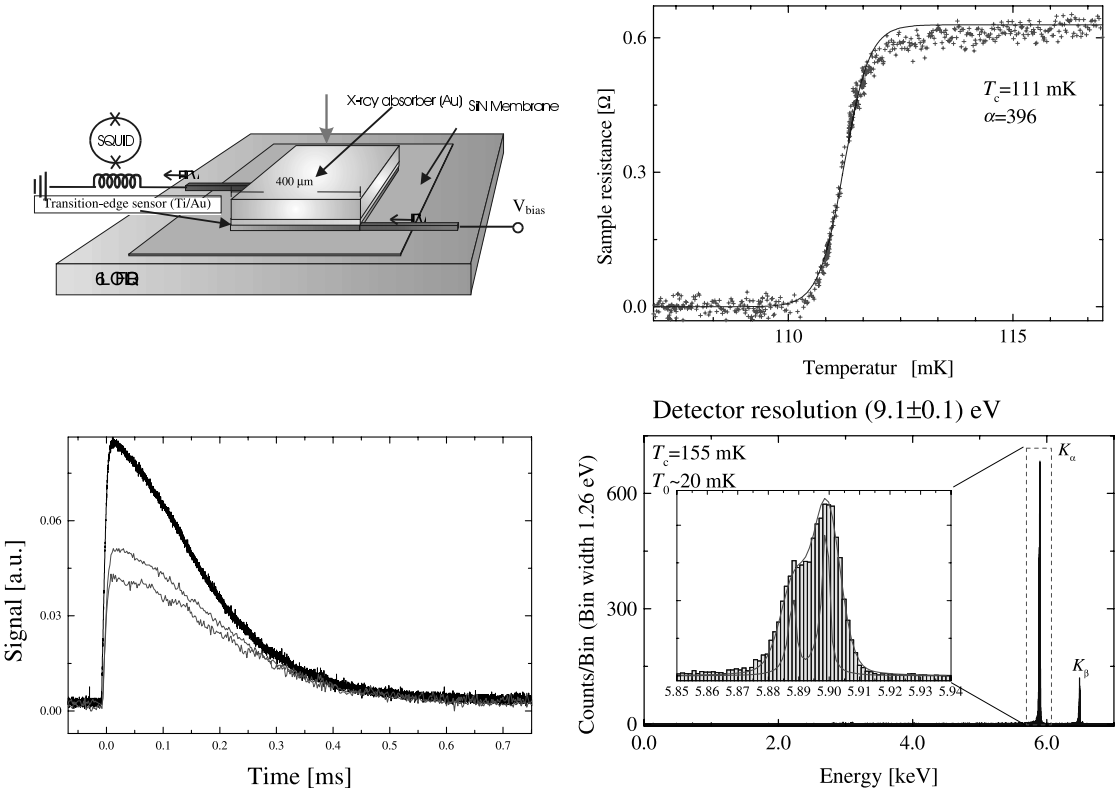


FIGURE 7 Characteristics of a TES-based X-ray microcalorimeter. Voltage-biased bimetal detector, measured by a SQUID ammeter on the top left. A typical superconductivity transition of a Ti/Au film on the top right. Measured SQUID current on X-ray hits on the detector on the bottom left, and the corresponding X-ray spectrum on the bottom right.

E. Antenna Coupled TES Infrared Bolometer

Besides soft X-rays, one can use TES detection in infrared and sub-millimeter wave micro-bolometry, primarily in astronomy. Figure 8 shows an antenna-coupled bolometer. It has been fabricated on silicon substrate. The spiral antenna is used to absorb the THz radiation from a weak source and the resulting ac current heats the TES microbridge in the center. It can be seen from the expanded view in Fig. 8 that the submicron wide TES strip is "airborne," i.e., it is etched from the substrate to improve thermal isolation and thus increase the sensitivity. Such devices, and many variations of them, are currently under intense investigation.

IV. NIS TUNNELING DEVICES

A. General

An NIS tunnel junction can be readily made between a number of material combinations. The simplest example with nearly ideal performance is provided by aluminum-aluminum oxide-copper, where aluminum is the superconductor at sub-kelvin temperatures and copper is one of the possible normal metals. (Also gold and silver are good examples of a normal metal.) A tunnel barrier of aluminum oxide forms easily on pure aluminum under an oxygen atmosphere at room temperature. Besides normal metal, also heavily doped semiconductors, e.g., Si, can be used in combination with aluminum, in which case the tunnel junction forms as a Schottky barrier even without oxidizing aluminum. Other superconductors can be used as well, for example, Nb with again higher T_c.

B. Transport Through a NIS Junction

The useful properties of NIS junctions stem from the nonconstant density of quasiparticle states n_S in a superconductor. There is the gap with width of twice the energy gap, 2Δ, with forbidden single particle states around the Fermi level E_F, at which the Cooper pairs sit. The density of levels diverges at $\pm\Delta$. Explicitly we can write

$$n_S(E) = n(0)|E|/\sqrt{E^2-\Delta^2}, \qquad |E| \geq \Delta \qquad (4.1)$$

$$= 0, \qquad |E| \leq \Delta. \qquad (4.2)$$

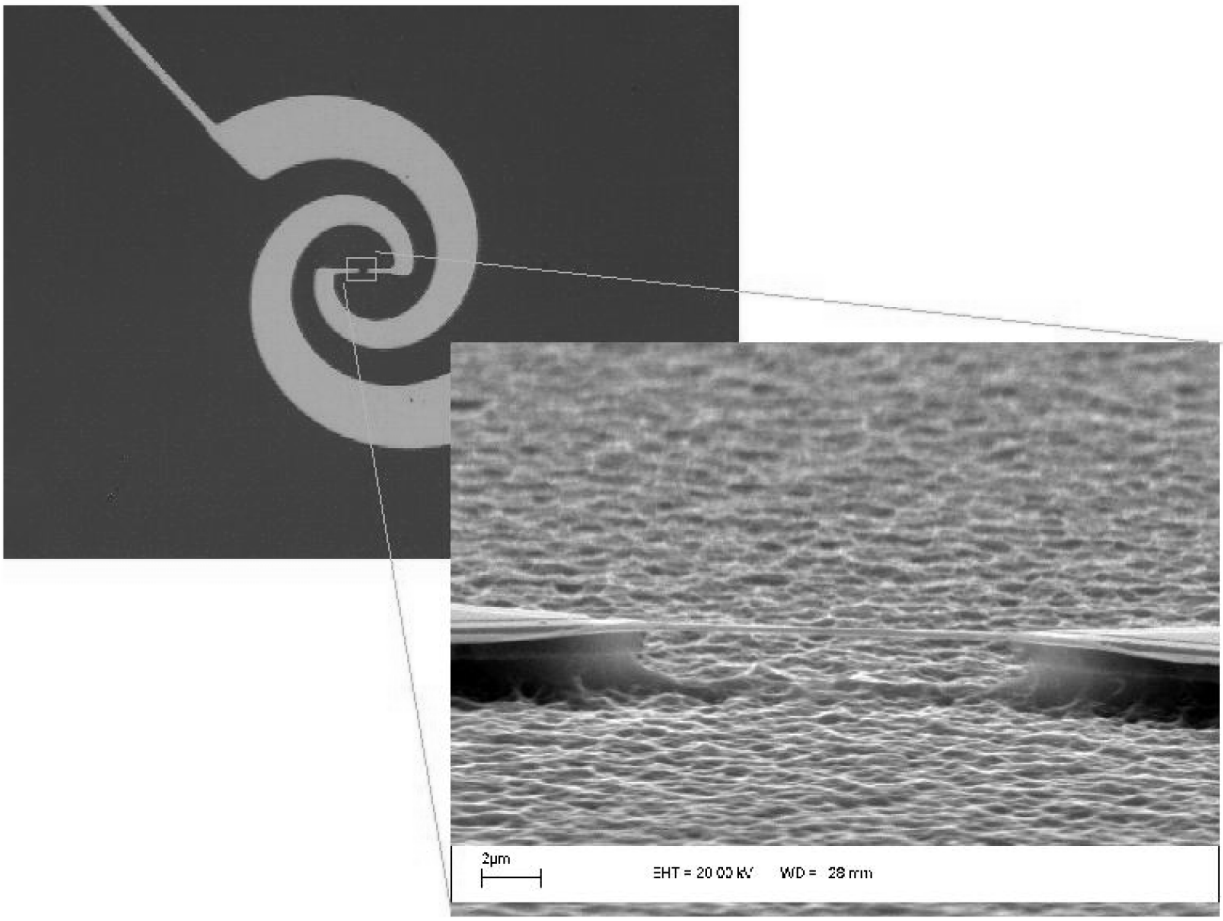


FIGURE 8 An antenna coupled infrared microbolometer. The spiral antenna picks radiation which induces ac current in the superconducting TES nanobridge lifted from the substrate (expanded view). The detector senses very weak IR radiation as temperature rise of the poorly thermalized TES bridge.

Here $n(0)$ is the corresponding (almost) constant density of electron levels in the normal metal state near the Fermi level, which is assigned to be the zero of energy. E is the energy of electrons counted from the Fermi level. The occupation of the levels is given by the Fermi distribution, $f(E) = 1/(1 + \exp(E/k_B T))$, whereby at zero temperature only states below the gap are occupied, and at nonzero temperatures there are thermally occupied states above the energy gap.

Figure 9 is the "semiconductor" energy diagram of an NIS tunnel junction, with occupied states shaded at a nonzero temperature. The bias voltage V lifts the Fermi level of the superconductor with respect to the normal electrode by the amount eV. If we make the typically very well justified approximation of low environmental impedance of the conductors, tunneling represents a horizontal crossing between S and N through the insulator I from an occupied state into an empty one, but not into the forbidden gap region. Therefore, at low voltages, $|\mathrm{eV}| < \Delta$, and at low temperatures, $k_B T \ll \Delta$ or $T \ll T_c$, virtually no current flows through the junction. Above the threshold voltage of $|\mathrm{eV}| = \Delta$ the current increases rapidly and soon resumes the ohmic value determined by the normal state tunnel resistance of the junction, R_T. If the temperature is not very low as compared to T_c, current "leaks" at sub-gap voltages due to the smeared Fermi distribution in the normal electrode with tail extending up to energies above the top of the energy gap in the superconductor.

At low temperatures the full current-voltage curve can be calculated approximately by

$$I(V) = (eR_T)^{-1} \int_{\Delta}^{+\infty} [f(E - \mathrm{eV}) - f(E + \mathrm{eV})] n_S(E)\, dE. \tag{4.3}$$

Figure 10A displays a few IVCs of NIS junctions calculated at different temperatures T/T_c. The measured ones closely follow this simple theoretical prediction.

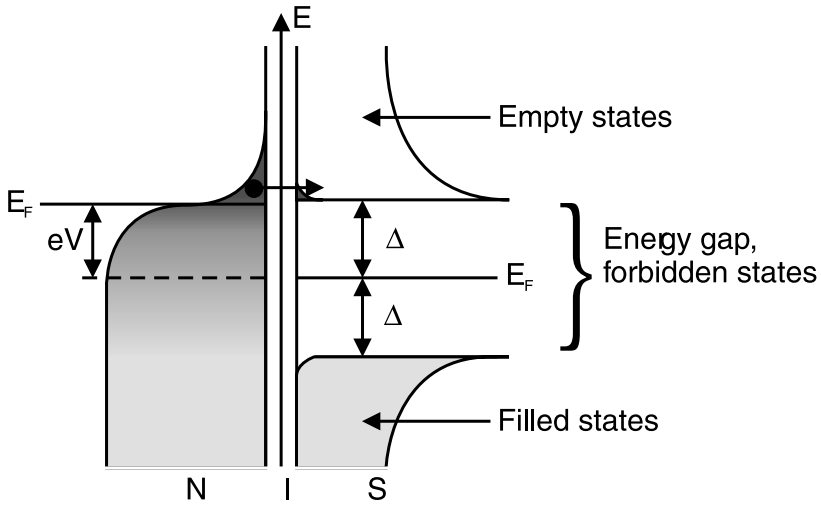


FIGURE 9 The energy diagram of a biased (at voltage V) NIS tunnel junction. The shaded area in the normal metal (N) gives the electron occupation according to the Fermi distribution. In the superconductor (S) there is the gap of width 2Δ with forbidden levels.

C. NIS Junction as a Sensitive Thermometer

Based on Fig. 10 it is obvious that an NIS junction can be used as a sensitive thermometer by biasing the junction at a constant current I or at a constant voltage. The first method is usually preferable because of the very strong temperature dependence of voltage at constant current within the gap region. In this bias mode the slope in the relatively wide temperature range with linear $V-T$ dependence is

$$dV/dT \simeq (k_B/e)\ln(I/I_0), \tag{4.4}$$

where the scaling current is $I_0 \equiv (eR_T)^{-1}\sqrt{\pi k_B T \Delta^{3/2}/2}$. The NIS thermometer is typically characterized by a large impedance: especially the dynamic resistance in the most sensitive gap region is very high, whereby measurement using a SQUID-based amplifier is not especially advantageous. Therefore measurement by a high input impedance voltage amplifier is usually preferred.

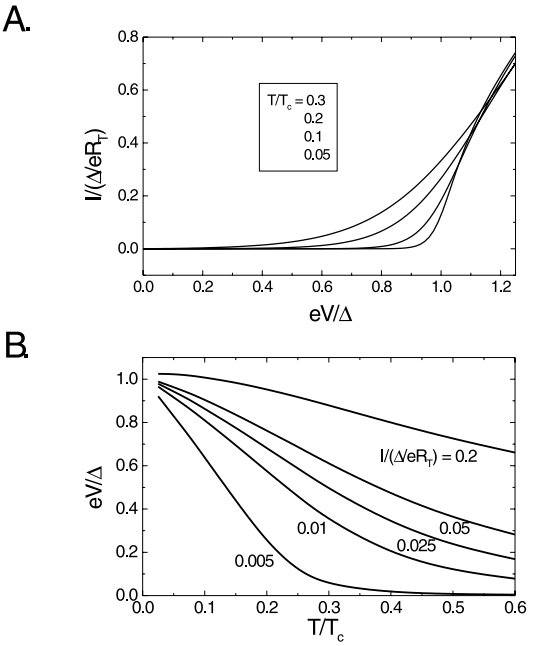


FIGURE 10 Current through an NIS junction. IVC of an NIS junction at different temperatures in (A). Voltage at a few fixed values of current providing a thermometer in (B).

V. NIS MICROCOOLING

A. Thermal Considerations

Besides the fact that current is preferably transported only at voltages exceeding the gap voltage, there is also a nonvanishing net heat current through the junction at nonzero bias voltages. This makes NIS junctions a very attractive candidate for on-chip cooling. When V has a value close to Δ/e, it is obvious that only the hot electrons in the tail of the Fermi distribution can be transported across the barrier. This way the average energy and thus the temperature of the normal metal electrode drops. The cooling power P_{NIS} of one NIS junction can be calculated analogously to what we did to determine the IVC [Eq. (4.3)] as

$$P_{NIS}(V) = \left(e^2 R_T\right)^{-1}\int_{\Delta}^{+\infty}(E-eV)[f(E-eV) - f(E+eV)]n_S(E)\,dE. \tag{5.1}$$

Figure 11 shows P_{NIS} calculated at a few values of temperature T/T_c. P_{NIS} reaches its maximum value (optimum

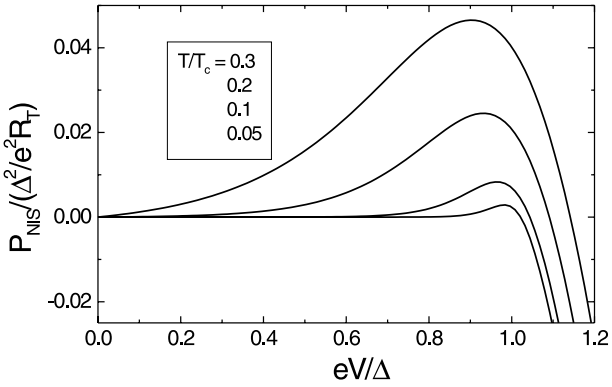


FIGURE 11 Cooling power of an NIS junction against bias voltage calculated at a few values of temperature. Maximum cooling power is reached at V slightly below Δ/e. At large values of V the curve approaches that of Joule heating with resistance R_T.

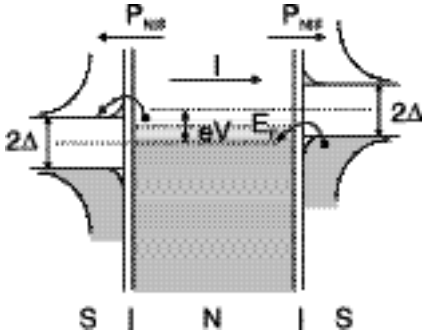


FIGURE 12 The energy diagram of a symmetric SINIS cooler at a bias voltage $V \simeq 2\Delta/e$ across the two junctions. Symmetry of the Fermi distribution leads to equal cooling power in the two junctions: the current is flowing unidirectionally through the SINIS pair (top), but heat is flowing out through both the two junctions.

bias point) just slightly below Δ/e. The value of this maximum cooling power is approximately

$$P_{\mathrm{NIS}} \simeq 0.6\frac{\Delta^2}{e^2 R_{\mathrm{T}}}\left(\frac{k_{\mathrm{B}}T}{\Delta}\right)^{3/2} \tag{5.2}$$

at temperatures $T \ll T_{\mathrm{c}}$.

If the normal electrode is not connected to a heat bath by a metallic contact, the electrons in it can really cool down, because the electron system decouples efficiently thermally from the underlying lattice due to the very weak electron-phonon (e-p) coupling at low temperatures. Heat current $\dot{Q}$ between the electrons and lattice is given by $\dot{Q} = \Sigma\Omega(T_{\mathrm{e}}^5 - T_{\mathrm{p}}^5)$, where Σ is a material dependent e-p coupling constant, Ω is the volume of the electrode, and T_{e} and T_{p} are the electron and lattice temperatures, respectively.

There are at least two ways of isolating the normal metal electrode from its environment in order to allow it to cool down. The N electrode can be made with an NS "Andreev" contact at its one end and the refrigerating NIS junction at the opposite one. A metallic NS contact does not allow heat to be transported electrically through. Yet this configuration can be replaced by a more efficient, stable, and easy-to-fabricate symmetric SINIS cooler. In SINIS the N electrode is symmetrically terminated at both its ends by an NIS junction. Figure 12 shows the energy level diagram of the SINIS cooler with a bias of approximately $2\Delta/e$ connected across.

In a relatively symmetric SINIS pair the voltage drop across each of the two junctions is close to Δ/e, thus allowing optimum cooling in the NIS junction. In the opposite SIN junction the cooling power is equal. The simple explanation of this is that while the hot electrons are skimmed away through the NIS junction, as explained above, they are replaced by cold ones entering into energy levels below E_{F} in the N electrode through the SIN junction. Since the Fermi distribution is symmetric around E_{F}, cooling power of the NIS and SIN junctions are the same, and the resulting total cooling power is doubled with respect to the simple NIS cooler.

One of the most important considerations in designing an NIS cooler is how to get rid of the heat dissipated in the secondary side of the cooler: the superconductor receives heat and as a very poor thermal conductor it tends to warm up whereby the cooling power of the NIS cooler diminishes. The coefficient of performance, COP, of the NIS junction, i.e., the ratio of the heat extracted from the normal electrode as compared to the total heat dissipated at the optimum bias point can be approximated by

$$\mathrm{COP} = \frac{P_{\mathrm{NIS}}}{IV} \simeq \frac{0.6\frac{\Delta^2}{e^2R_{\mathrm{T}}}\left(\frac{k_{\mathrm{B}}T}{\Delta}\right)^{3/2}}{(\Delta/e)I_0} \simeq 0.3T/T_{\mathrm{c}}. \tag{5.3}$$

For an aluminium-based cooler with (thin film) T_{c} of about 1.3 K, we have $\mathrm{COP} \simeq 0.02$ at $T = 0.1$ K, i.e., approximately 50 times more heat is dissipated in the superconductor than what is removed from the N electrode. This can lead to degraded cooling performance since the cooling power with a "hot" superconductor is reduced from that of Eq. (5.2) to

$$P_{\mathrm{NIS}} \simeq 0.6\frac{\Delta^2}{e^2R_{\mathrm{T}}}\left(\frac{k_{\mathrm{B}}T_{\mathrm{n}}}{\Delta}\right)^{3/2} - \frac{\Delta^2}{e^2R_{\mathrm{T}}}\sqrt{\frac{2\pi k_{\mathrm{B}}T_{\mathrm{s}}}{\Delta}} \times \exp\left(-\frac{\Delta}{k_{\mathrm{B}}T_{\mathrm{s}}}\right), \tag{5.4}$$

where we have now indicated by the subscripts n and s if the temperature T is that of the N or S electrode, respectively.

To avoid or at least to reduce heating of the superconductor, i.e., to reduce excess population of hot quasiparticles, so-called *quasiparticle traps* can be used. These are

typically normal metal films in contact with the superconductor near the junction, so that the excess heat in the superconductor can be conducted into this normal metal which has superior heat transport capability. The trap can be made of copper, and it can be either in direct metal-to-metal contact with the superconductor, or it can contact through a thin insulator layer. The most efficient geometrical arrangement may turn out to be a sandwich cooler, where both the NIS junction and the trap are all on top of each other. Such a trap geometry has been illustrated in Fig. 13. At the time of writing this volume, the best achieved performance using quasiparticle traps is a cooling power of about 15 pW per NIS junction (junction area $\sim 20\ (\mu\mathrm{m})^2$) at $T_\mathrm{n} = 0.1$ K and $T_\mathrm{s} = 0.3$ K.

To use the NIS cooler as a conventional refrigerator one needs to provide an intermediate cooled heat bath that is only weakly coupled to the heat bath of the precooling stage. The object to be cooled, e.g., a microcalorimeter or a bolometer, then sits at this intermediate heat bath, but not electrically connected to the N electrode

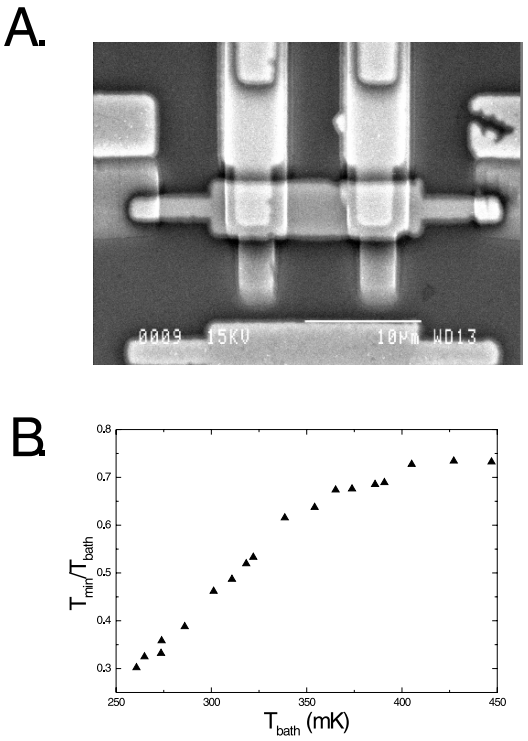


FIGURE 13 A SINIS cooler with quasiparticle traps. In (A) the NIS junctions of the SINIS pair are seen in the central area. The superconductors extend up vertically in the figure, and the quasiparticle copper traps are sandwiched down to the junction area. The horizontal bar is the N electrode cooled, with SINIS thermometers at the two ends. The relative temperature reduction of N as a function of the starting temperature is shown in (B). The cooling power is about 15 nW per junction with N electrode at 100 mK.

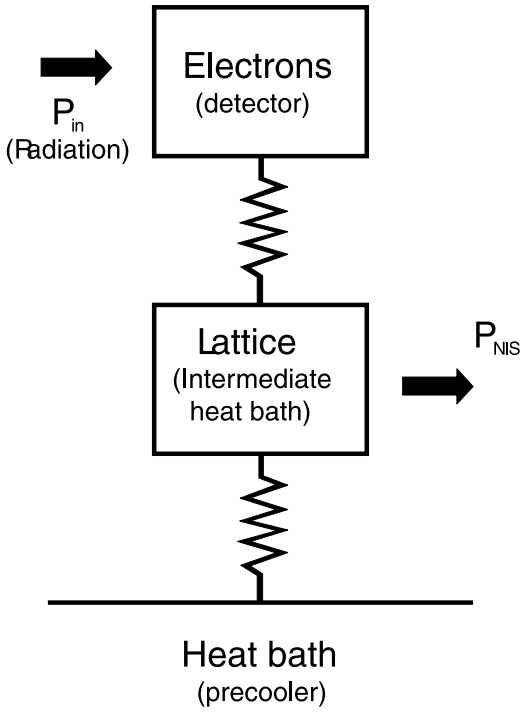


FIGURE 14 Thermal diagram of the indirect cooling method, where SINIS is used to lower the lattice temperature of an intermediate heat bath, formed by, e.g., a silicon nitride membrane with a cooled copper cold finger.

of the cooler. Figure 14 illustrates this *indirect* cooling method.

In some cases it might be tempting to cool down only the electrons directly, for example, if the N electrode could act as a detector as such. This *direct* cooling is, however, not usually a good solution, because the noise properties of the detector to be cooled depend not only on its electron temperature but usually equally as much on the temperature of the lattice. For example, the fundamental limit of NEP, the noise equivalent power, i.e., the minimum resolvable power of a radiation bolometer assumes a form

$$\mathrm{NEP} = \sqrt{5\Sigma\Omega\left(T_\mathrm{e}^6 + T_\mathrm{p}^6\right)}. \tag{5.5}$$

Thus, by cooling down the electrons only, one reduces the NEP by not more than about 30%.

Figure 15 shows one possible solution of an on-chip SINIS refrigerator which is capable of cooling down a micromachined silicon nitride membrane on which the object to be cooled can be processed also. The cooling junctions are patterned on bulk silicon, and the heat is flowing out from the membrane toward the cooling junctions via a copper cold finger.

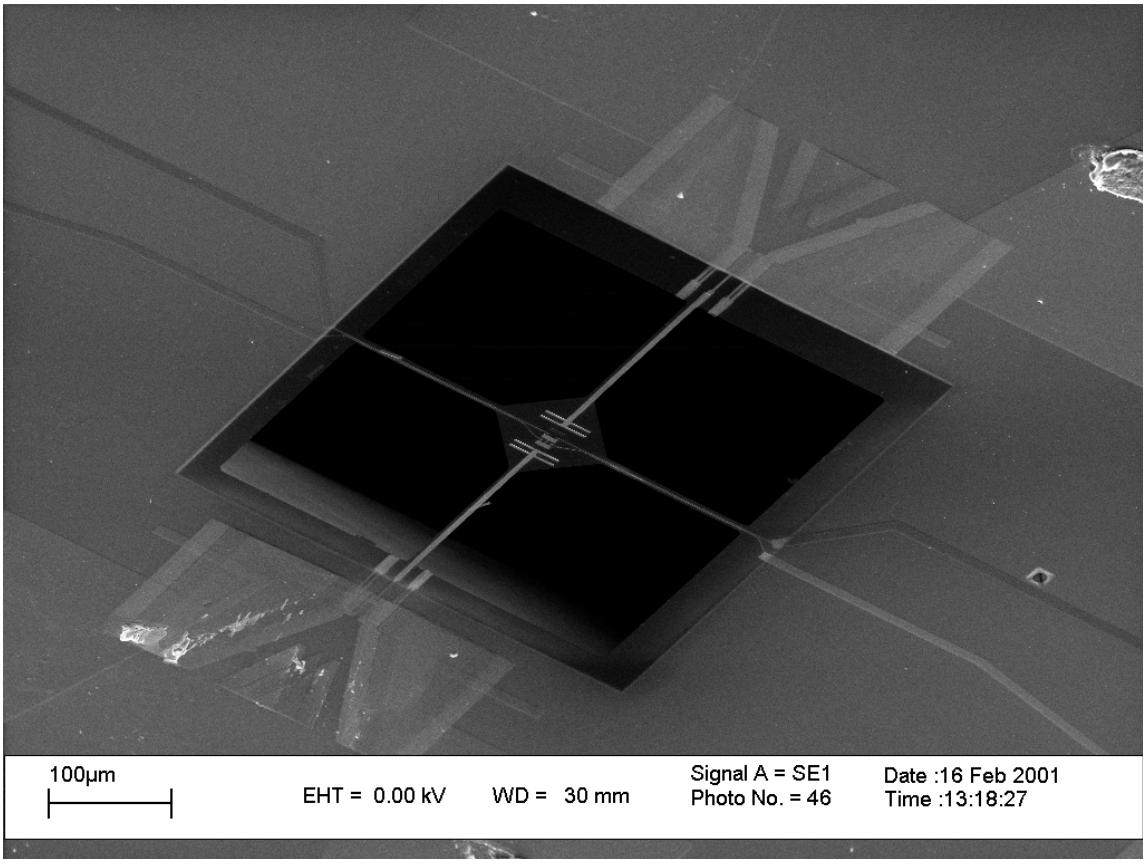


FIGURE 15 An example of an SINIS cooler in an indirect cooling configuration. The cooling junctions are at the top and bottom of the figure. The central area is the thermally semi-isolated silicon nitride membrane, which forms the intermediate heat bath together with the copper cold fingers extending from the cooling junctions to the membrane. Separate sensors, in this case NIS thermometers, can be placed on the cooled membrane in the central area of the membrane.

VI. SUPERCONDUCTING SINGLE CHARGE DEVICES

A. Basics of Single Electron and Single Cooper Pair Tunneling

A family of new generation devices, including single-charge transistors and logic, metrological current and charge standards, a Coulomb blockade absolute thermometer, and eventually a fully solid-state quantum computer, is expected to emerge along the technical development after the feature size in patterning of components has dropped down to deep below the 1-μm range. What governs charge transport then, either electron or Cooper pair current, is the charging energy by single electrons or Cooper pairs. These effects emerge, in particular, when temperature is low such that thermal energy $k_B T$ is much smaller than the charging energy needed to add or remove one elementary charge carrier to or from the circuit. So far, in respect of superconducting devices, particular attention has been paid to aluminium-based, single-charge devices. These are typically structures where two lithographically patterned thin film conductors are separated from each other by an oxide barrier forming a tunnel junction.

B. Fabrication Issues

The exclusive technique to produce superconducting single-charge devices is electron beam lithography combined with shadow mask evaporation. The aim is to obtain superconducting thin film conductors with 100 nm order line widths connected to each other by a high quality oxide barrier. The standard technique to do this is illustrated in Fig. 16.

An electron-beam-sensitive, double-layer resist has been spun on a flat, preferably semiconducting substrate. The bottom resist layer is more sensitive to the electron shower than the top one, and by this action the developed pattern after electron beam lithography produces an undercut profile as demonstrated in Fig. 16. This way the

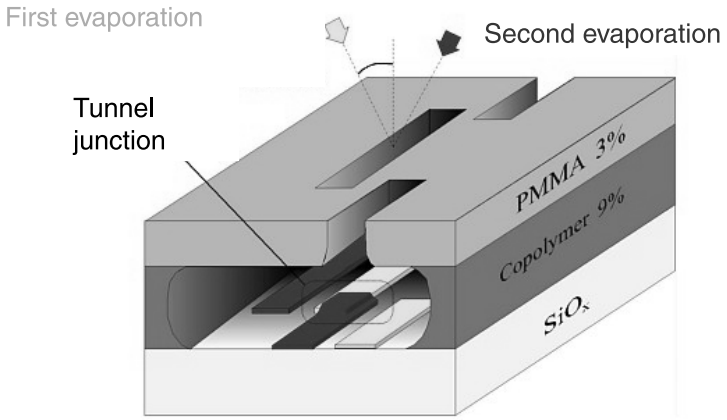


FIGURE 16 Schematic presentation of a resist mask in electron beam lithography to allow multiple angle deposition in fabricating tiny tunnel junction structures for single electron and single Cooper pair devices. Typical feature sizes are of the order of 100 nm.

patterned resist forms a suspended mask which can be used in multi-angle deposition of the superconducting metal. When the neighboring conductors are near enough the undercut in these areas connects the pattern in the lower resist layer. Now, by depositing superconducting metal in the two angles indicated, and by oxidizing the first aluminium layer inbetween the metal depositions, an ultrasmall superconductor-insulator-superconductor (SIS) tunnel junction can be formed. The computer-aided pattern generation allows easy design of complicated multijunction circuits as well: new designs are easy to realize in electron beam lithography but the patterning itself is slow for industrial purposes. Figure 17 illustrates an example of a simple double junction structure fabricated by the technique described above.

C. Single Cooper Pair Box, Transistor, and Multijunction Circuits

The basic device concept in single electronics is a single electron box (SEB) and its superconducting counterpart single Cooper pair box (SCB), schematically presented in Fig. 18A. A small superconducting island between the gate capacitor C_g and the "semitransparent" tunnel junction capacitor C can house an integer number, N of Cooper pairs.

Electrostatically, the energy of the system, E_{ch}, corresponding to each N and gate voltage V_g can be written as

$$E_{ch} = (2eN - C_g V_g)^2 / [2(C + C_g)]. \qquad (6.1)$$

For each N, Eq. (6.1) presents a parabola against V_g, centered at $C_g V_g = 2eN$ (Fig. 19). Each state $|N\rangle$ gives the minimum energy in the interval

$$2e(N - 1/2) < C_g V_g < 2e(N + 1/2). \qquad (6.2)$$

When temperature is low, $k_B T \ll (2e)^2/[2(C + C_g)] \equiv E_c$, N is determined to within one integer value by the setting of the gate voltage within this interval, and N changes by one upon crossing the limits of this range. This change is allowed by Cooper pair tunneling through the junction.

Each Josephson junction has, in addition to the classical charging energy above, which is important for small junctions, the more familiar Josephson coupling energy,

$$E_J = -E_{J0} \cos \varphi, \qquad (6.3)$$

where φ is the phase difference across the junction and the strength of the Josephson coupling is given by $E_{J0} = \hbar I_c / 2e$. The total Hamiltonian of a Cooper pair box, in the absence of quasiparticle (single electron) excitations and in the environment with vanishing impedance, is the sum of E_{ch} and E_J.

The net effect of the two energy contributions is the formation of a *band structure*, where, instead of sharp crossing of the parabolas of E_{ch}, a gap of magnitude E_{J0} opens near the degeneracy points of succeeding states $|N\rangle$. This means, in turn, that upon cranking V_g across one of the degeneracy points, the eigenstate evolves smoothly as a superposition of the two "nearest" charge states from one pure charge state into another in the two-state

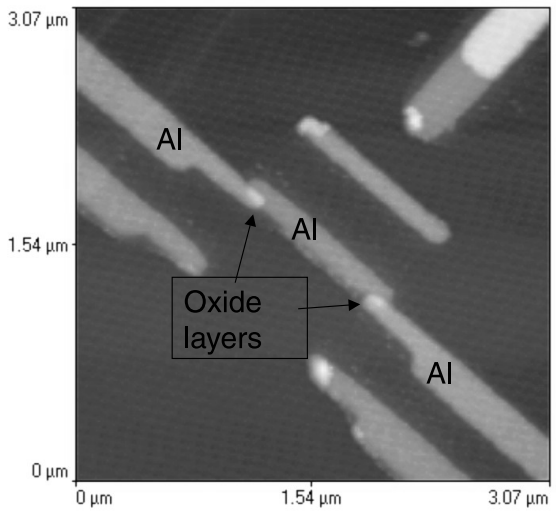


FIGURE 17 A single electron transistor fabricated through a resist stencil mask of the type shown in Fig. 16. The shadows are of aluminum.

approximation. In particular, at the exact degeneracy value of the gate voltage, the two-charge states exist at an equal probability (1/2). Figure 20 demonstrates the influence of E_J on both the eigenstates and the probabilities, i.e., occupations p of different charge states in these eigenstates. The lowest band corresponds to the ground state, the higher ones are naturally excited states, which can be reached, e.g., by so-called Landau-Zener crossing upon fast enough (nonadiabatic) sweep of V_g.

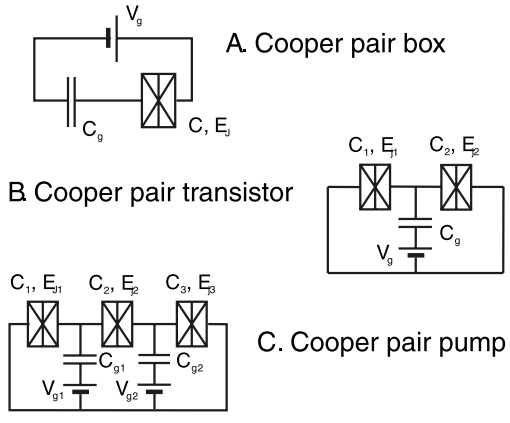


FIGURE 18 (A) A single Cooper pair box connected to a gate bias V_g through gate capacitance C_g. The crossed square depicts a small Josephson junction with capacitance C. (B) A single Cooper pair transistor consisting of two Josephson junctions and a gate. (C) A three-junction linear Josephson junction array with gates for the two islands.

The interplay between the charging energy and the Josephson coupling provides the possibility of using the SCB with a *small Josephson junction as a quantum bit*, possibly in future quantum computation.

More complex circuits with small Josephson junctions can be described in a similar manner. The charging Hamiltonian is technically more complicated and the Josephson energy is the sum of the contributions of the type given by Eq. (6.3), but no conceptual complications arise. Perhaps the best known "device" is the single Cooper pair transistor (SCT) consisting of two series Josephson junctions and of a gate electrode (see Fig. 18B). Josephson current

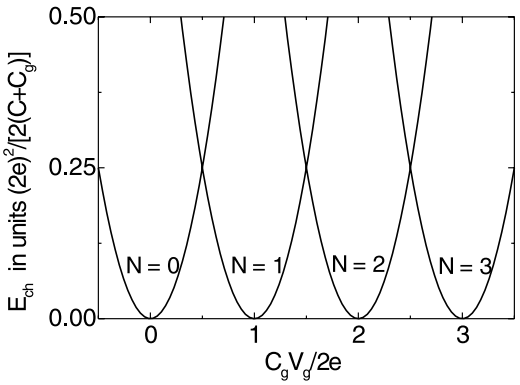


FIGURE 19 Energy diagram including just the charging energy of a Cooper pair box against the gate voltage for states with different number of Cooper pairs on the island.

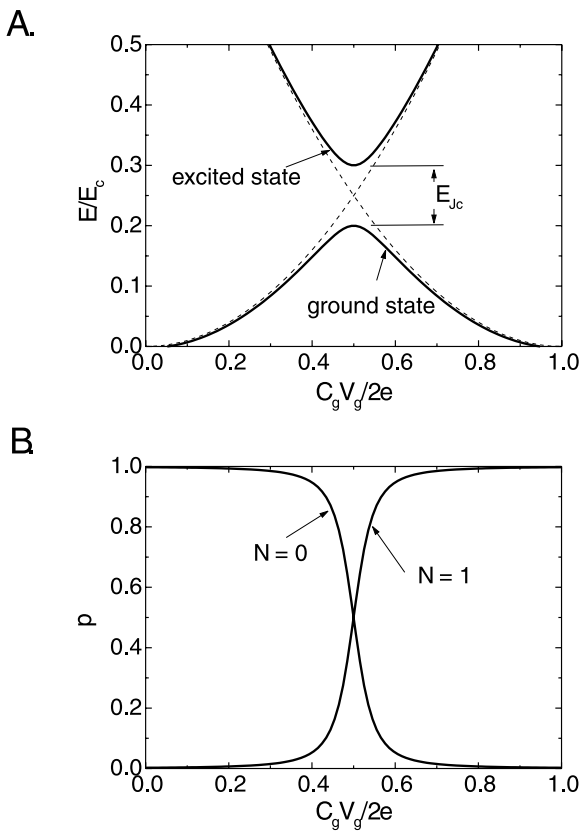


FIGURE 20 Cooper pair box in presence of Josephson coupling. (A) Eigenstates as a function of gate voltage form a band structure with a gap of magnitude $E_{Jc} = E_{jo}$ opening at the degeneracy point. $E_c \equiv (2e)^2/[2(C + C_g)]$. (B) Occupation p of different charge states against gate voltage. On crossing the degeneracy point the state evolves from an almost pure charge state into another. As a controllable two-state system the Cooper pair box forms a candidate of a solid-state quantum bit.

through the SCT can be modulated by the gate voltage periodically. Multijunction arrays of small junctions with gates connected to the intermediate islands (Fig. 18C) can be used for quantized charge transport by Josephson tunneling.

SEE ALSO THE FOLLOWNG ARTICLES

SUPERCONDUCTING CABLES • SUPERCONDUCTIVITY • SUPERCONDUCTIVITY MECHANISMS • SUPERCONDUCTORS, HIGH TEMPERATURE

BIBLIOGRAPHY

Averin, D. V., and Likharev, K. K. (1991). Single electronics: correlated transfer of single electrons and Cooper pairs in systems of small tunnel junctions. *In* "Mesoscopic Phenomena in Solids," B. L. Altshuler, P. A. Lee, and R. A. Webb eds., Elsevier, New York.

Bunyk, P., Likharev, K., and Zinoviev, D. (2001, to appear). "RSFQ Technology: Physics and Device," *Int. J. High Speed Electron. Systems*, special issue, M. Rodwell (ed.)

Clarke, J. (1996). SQUID fundamentals. *In* "SQUID sensors: Fundamentals, Fabrication and Applications," H. Weinstock, ed., Kluwer, The Netherlands.

de Korte, P., and Peacock, T., eds. (1999). "Proceedings of the 8th International Workshop on Low Temperature Detectors, LTD-8," Elsevier, The Netherlands.

Leivo, M. (1999). On-chip Cooling by Quasiparticle Tunnelling below 1 Kelvin, Ph.D. thesis, University of Jyväskylä, Finland, unpublished.

Tinkham, M. (1996). "Introduction to Superconductivity," 2nd ed., McGraw-Hill, New York.

Van Duzer, T., and Turner, C. (1998). "Principles of Superconductive Devices and Circuits," 2nd ed., Prentice Hall, New Jersey.



Thin-Film Transistors

Yue Kuo
Texas A&M University

I. Introduction—Status of Thin-Film Transistors
II. Thin-Film Transistor Structures
III. Material Issues in Thin–Film Transistors
IV. Device Physics and Requirements in Thin-Film Transistors
V. Process Issues in Thin-Film Transistor Fabrication
VI. Summary and Technology Trends

GLOSSARY

Active matrix liquid crystal display Liquid crystal display with each pixel driven by an individual solid-state device, such as a thin-film transistor or diode.
Amorphous silicon Silicon with no long-range order structure, which has broad tail states and large deep states.
Dangling bond Valence electron that can form a bond but actually is not part of a bond, thus can be in a more stable state by forming a bond with elements such as hydrogen.
Deep states Gap states near the center of the energy gap, which are caused by defects.
Microcrystalline crystal silicon Silicon with very small, nanometer-sized grains imbibed in an amorphous matrix.
Off current Minimum current flowing from source to drain at a specified drain voltage.
On current Maximum current flowing from source to drain at specified gate and drain voltages.
Polycrystalline silicon Silicon-containing crystals of different orientations joined at grain boundaries.
Tail states Gap states near conduction and valence band edges.
Threshold voltage Minimum gate voltage required before an appreciable amount of drain current starts flowing through a thin-film transistor.

A THIN-FILM TRANSISTOR (TFT) is a solid-state field-effect transistor (FET), which contains three electrodes, i.e., source, drain, and gate. The operation principle of a TFT is similar to that of a metal–oxide semiconductor (MOS) transistor in a very large-scale integrated circuit (VLSIC), i.e., the magnitude of the source-to-drain current is manipulated by the voltage applied to the gate. However, due to the difference in structures, materials, and fabrication processes, the device characteristics of a TFT are very different from those of a MOS transistor. These factors widely separate their applications. For example, MOS transistors are used mainly in high-speed,

very small-geometry integrated circuits, while TFTs are commonly used in large-area, nonwafer displays.

I. INTRODUCTION—STATUS OF THIN-FILM TRANSISTORS

A. TFT Applications

Currently, the main application of TFTs is in the liquid crystal display (LCD). The TFT LCD is a kind of active-matrix LCD (AMLCD) of which each pixel is addressed by an individual device. Each pixel can be turned on and off separately without the influence of electrical signals of adjacent lines or pixels. In addition, since the TFT is a triode device, it has gray-scale capability, which is difficult to achieve with a diode device. Therefore, the TFT LCD has superior display qualities, such as an absence of ghost-image effects, a high switching speed, and large viewing angles, compared to other types of LCDs.

Since the beginning of their mass production in the early 1980s, TFT LCD products have been widely applied in commercial, industry, transportation, entertainment, education, and military markets. In the early stages, the TFT LCD was used mainly in high-price products such as laptop computers, projectors, instruments, and airplanes. As the production technology has matured and the yield has improved, TFT LCD products have quickly spread to other markets such as desktop computers, televisions, and games. So far, TFTs have been used in direct-view or projection displays whose size varies from less than 1 to more than 30 in. diagonally.

TFT applications are not restricted to displays. Table I lists some of the known TFT products that either are commercially available or have been fabricated into prototypes. In these applications, the TFT functions as a switch that drives a specific device, reads information from a designated local structure, or engages in simple circuit operations. In one case, i.e., SRAMs, the TFT replaces a high-resistance resistor to control the resistivity better. Since knowledge on TFTs is available nowadays, many new applications are emerging continually.

TABLE I Applications of TFTs

TFT function	Examples of applications
Driving devices	Liquid crystal displays: direct view and projection
	Printers and facsimiles
	Inorganic and organic light-emitting devilces
Readout devices	Two-dimensional medical imagers
	Sensors
	Radiation detectors
Integrated circuits	SRAM loading cells
	EEPROM
	High-voltage circuits
Others	Artificial retina
	Photo imagers

Compared with other semiconductor transistors, such as the VLSI MOS, the major advantage of TFT technology is that the fabrication process is independent of the substrate size and materials. Currently, TFT arrays can be fabricated on a low-temperature glass substrate with a size of 1×1 m.

B. Requirements for LCD Applications

The basic requirements for a TFT are the same as those for other semiconductor transistors, i.e., a high field-effect mobility (μ_{eff}), a high on current (I_{on}), a low leakage or off current (I_{off}), and a low threshold voltage (V_{th}). Depending on the application, some of the above characteristics may be more important than others. For example, for LCD pixel driving, a low I_{off} is more desirable than a high μ_{eff}. For the circuit operation, a high μ_{eff} is most important. For low power consumption, the V_{th} has to be kept very low. There are certain unique issues regarding TFTs, such as photosensitivity and reliability, which need to be addressed. They are discussed in subsequent sections of this chapter.

II. THIN-FILM TRANSISTOR STRUCTURES

A TFT is composed of several layers of thin-film materials deposited by various methods. The substrate serves only as a supporting material for the TFT. It indirectly influences the transistor's structure, for example, through the stress mismatch or impurity distribution. Due to the flexibility of the deposition processes, TFTs can be fabricated in many different structures. In principle, as long as the semiconductor, dielectric, and conductor films can be arranged into a structure similar to that of a field-effect transistor, a TFT can be manufactured. However, even with the same structure, the TFT performance can vary drastically according to other factors such as the thin-film material properties and process conditions. These issues are discussed in subsequent sections.

Figures 1a–d show four of the most common TFT structures. The first three structures are used in the amorphous silicon (a-Si:H) TFT. The last structure is used exclusively in the polycrystalline silicon (poly-Si) TFT. Recently, the structure in Fig. 1a has also been used to prepare poly-Si TFTs.

For the a-Si:H TFT, the inverted, staggered structures, i.e., Figs. 1a and b, are more popular than the staggered structure, i.e., Fig. 1c, because the former give better

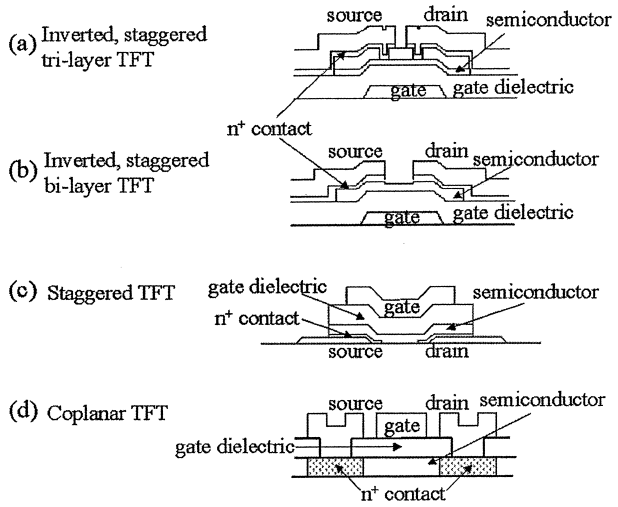


FIGURE 1 Common TFT structures.

device characteristics, such as a higher μ_{eff} and lower V_{th}, than the latter. The trilayer structure, i.e., Fig. 1a, has characteristics superior to those of the bilayer structure, i.e., Fig. 1b, because the former allows the use of a very thin a-Si:H layer and has a back-channel protection film. However, the structure in Fig. 1a needs one more mask to define the back-channel passivation region than does the structure in Fig. 1b. For the mass production environment, the additional masking step can affect the throughput. Therefore, despite its inferior device performance, the bilayer structure is used in some production processes. The physical and chemical properties at the semiconductor/dielectric interfaces are critical factors that can be used to explain why certain structures give better TFT characteristics than other structures. These interface properties are also closely related to the process conditions. Detailed discussions of these factors are included in Sections III and IV.

Based on the structures in Fig. 1, many advanced TFTs have been fabricated. For example, a light-blocking layer is required to protect the a-Si:H channel area from light irradiation. If this blocking layer is made of a metal and is connected to the gate, a light-insensitive, dual-gate TFT is formed. The I_{off} of the dual-gate TFT is high because of the lack of a back-channel interface. By inserting a dielectric film into the a-Si:H layer, a vertically redundant TFT can be formed. This new TFT has a low I_{off} and an enhanced I_{on}. In another case, by splitting the gate into a U-shape, a horizontally redundant TFT can be formed. This new transistor occupies about the same space as the non-split-gate TFT but has a higher I_{on}. In addition, the channel length of a TFT can be drastically reduced to close to the a-Si:H layer thickness by forming a vertical structure where the source and drain are formed beneath and above the a-Si:H film, separately. In another design, the a-Si:H TFT can be used as a high-voltage switching device by purposely offsetting the drain region. In summary, the purpose of developing a new TFT structure is to achieve one or more of the following goals: to improve the device performance, to solve some intrinsic device problems, to obtain special functions, to reduce the area occupancy of the transistor, to form a redundant architecture, and to simplify the manufacturing process.

III. MATERIAL ISSUES IN THIN–FILM TRANSISTORS

A complete TFT involves three types of thin-film materials: semiconductors, dielectrics, and conductors. Table II lists common materials used in different device regions and some comments. Many of these materials are similar to those used in VLSI MOS transistors but with very different material properties.

TABLE II Thin-Film Materials for TFTs

Device region	Material(s)	Comments
Semiconductor layer	Hydrogenated amorphous silicon (a-Si:H)	Dangling bonds, traps, photosensitivity, etc.
	Polycrystalline silicon (poly-Si)	Grain boundaries, size, quality, etc.
	Cadmium selenide (CdSe)	Stoichiometry, donors, etc.
	Other semiconductors: Ge, GeSi, Te, Sb, organic, etc.	In the experimental stage
Gate dielectric	SiN_x (for a-Si:H TFTs); Ta_2O_5, Al_2O_3, SiN_xO_y, SiO_2 (for poly-Si TFTs)	One layer or multilayers with a critical interface layer
Passivation layer	SiN_x, SiO_2	Band bending, etc.
Interconnection layer	Al, Al alloys, Cu, refractory metals (Mo, Ti, Ta, W, MoW, etc.)	Conductivity, hillocks formation, etc.
Ohmic contact	Heavily doped n^+ region (a-Si:H, μc-Si, poly-Si)	Dopant concentration, efficiency, surface cleanness
Pixel electrode	Indium tin oxide (ITO), tin oxide (SnO_2), etc.	Electric conductivity, light transmittance, etc.
Substrate	Glass, plastics, etc.	Low mobile ion content, temperature and chemical resistant, transparency, etc.

Since there is no method for depositing a single crystalline film on the surface of an amorphous material, it is natural that the deposited semiconductor film has an amorphous or polycrystalline morphology. Currently amorphous silicon (a-Si:H) is the most commonly used semiconductor material because the deposition process is simple [e.g., by plasma-enhanced chemical vapor deposition (PECVD)], low-temperature (e.g., less than 350°C), and easy to scale up (e.g., to a large-area substrate). A-Si:H has the following characteristics: (1) an absence of long-range order, (2) a large number of defects within the band gap, (3) an n-type intrinsic film, and (4) photoconductivity. The high density of states (DOS) comes from defects near conduction and valence band edges (i.e., tail states) and deep within the band gap (i.e., traps). The tail states come from the lack of long-range order of the film. The trap states come from the unsaturated dangling bonds that can be partially passivated by hydrogen atoms. The band gap of a-Si:H is higher that of single-crystal silicon, i.e., 1.7 vs 1.1 eV. All a-Si:H TFTs are n-channel transistors because the as-deposited intrinsic a-Si:H film is n-type. The heavily phosphorus-doped a-Si:H (n^+) film can be deposited by introducing a large amount of phosphine (PH_3) into the deposition chamber. The n^+ film is used as the ohmic contact layer in the source and drain areas. Although a-Si:H can be doped into a p-type semiconductor, e.g., by introducing diborane (B_2H_6) gas into the feed stream, the p-type TFT is difficult to fabricate because of the high defect density near the valence band. The photoconductivity of an a-Si:H film comes from the splitting of electron-hole pairs under the exposure of light, which is also known as the Staebler–Wronski effect. The film's conductivity is increased several orders of magnitude by this effect. The a-Si:H TFT becomes highly leaky when it is illuminated by light. Several methods have been suggested to reduce the photosensitivity of the a-Si:H TFT, such as reducing the a-Si:H thickness, reducing the back-channel interface states, introducing defect centers into the film, and applying a light-blockage layer on the transistor. All of them, except for the addition of defect centers, have been implemented in commercial products.

A microcrystalline silicon (μc-Si) film is composed of small silicon crystals, e.g., smaller than 10 nm, imbibed in a matrix of amorphous silicon. The defect density of μc-Si is not much lower than that of a-Si:H even though small grains exist in the film. In addition, due to the harsh deposition conditions, there is a high interface DOS at the μc-Si/dielectric interface. Therefore, μc-Si is not a good candidate to replace a-Si:H as the semiconductor layer of a TFT. However, the n^+ μc-Si film can have a conductivity more than two orders of magnitude higher than that of the n^+ a-Si:H film, which makes it a desirable ohmic contact layer in the source and drain regions.

Poly-Si contains crystalline silicon grains in various orientations and has an energy gap similar to that of single-crystal silicon, but a large number of defects. These defects are located in two areas: at grain boundaries and within grains. Grain boundaries trap charges and form potential barriers between adjacent grains. They also serve as short paths for current. Most grain boundary defects are due to the existence of unpassivated dangling bonds. The quality of the grain is dependent on the deposition or crystallization condition of the film. The grain's bulk quality is more important than its size. For example, low-quality grains usually contain many dislocations and microtwins, which serve as defect centers. A high-performance poly-Si TFT usually contains large-size, low-defect density silicon grains. In addition to high defect densities, grain boundaries have a high impurity diffusion rate and a high oxidation rate. Therefore, grain boundary quality control is an important issue in poly-Si TFTs.

The electrical requirements of a TFT's gate dielectric are the same as those of a MOS FET's gate dielectric, i.e., low interface states and bulk film charge trapping capability. However, the material properties of these two dielectrics are very different. For example, thermally grown silicon oxide (SiO_2) is used exclusively in the MOS FET. The most popular gate dielectric for an a-Si:H TFT is the slightly nitrogen-rich silicon nitride (SiN_x), which contains a large amount of hydrogen, e.g., more than 25%. These hydrogen atoms passivate dangling silicon and nitrogen bonds in the film as well as at the interface with the a-Si:H film. For large-area a-Si:H TFT array fabrication, a dual-gate dielectric structure is often used. Although various types of dielectrics, such as SiO_x, SiO_xN_y, TaO_x, and AlO_x, have been used as the bulk film, the nonstoichiometric SiN_x always serves as the interface layer because it gives low interface states. As for the poly-Si TFT, the deposited SiO_2 is commonly used as the gate dielectric material. In certain cases, SiN_x has also been used. The material properties of these dielectric films, such as the stoichiometry, density, and hydrogen content, are changed during the subsequent thermal processing steps. For the inverted, staggered, trilayer TFT, material properties of the channel passivation layer can influence the TFT properties, such as the leakage current, through varying charge densities and interface states. The high hydrogen-content SiN_x film can be used as a hydrogen source to passivate grain boundaries of a poly-Si film.

The main material requirement for the interconnect line is the conductivity. Although copper has a very low resistivity, it is rarely used in TFT products because of material and process difficulties. For example, copper has a low adhesion force on dielectric surface, it is easy to corrode by chemicals, and the etch process is difficult to

control. Recently, aluminum is getting popular as the gate interconnect material for large-area TFT arrays. However, aluminum has the hillocks formation problem, which can be solved by adding an extra element, such as zirconium, to the film or by using a capping layer. Currently, refractory metals are the most popular interconnection material. A redundant interconnect structure is commonly used for the large-area TFT array for the yield improvement. As the array size is increased and the line width is shrunk, aluminum or copper has to be used. More effective methods of solving the above material problems have to be discovered.

Indium tin oxide (ITO) is exclusively used as the pixel electrode in TFT LCD because it has a high light transmittance in the visible wavelength and a high electrical conductivity. The highly conductive ITO has an In/Sn ratio of 90/10 and is in the polycrystalline form.

Alkaline earth boroaluminosilicate glass is widely used as the substrate for large-area TFT arrays. It has a strain point around 550 to 650°C, which greatly limits the TFT process temperature. Since the a-Si:H TFT is fabricated at a much lower temperature, the substrate has little influence on transistor characteristics. However, for poly-Si TFT fabrication, glass is an important factor to consider. Even for low-temperature processes, the glass can be exposed to a high temperature for a short period of time. Glass properties, such as compaction, stress, impurity release, and surface smoothness, can be changed by the short heat pulses. Recently, there are many efforts on fabricating TFTs on plastic substrates that typically can stand a temperature lower than 150°C. The main material properties to consider for a plastic substrate are light transmittance, thermal expansion coefficient, stress, and process compatibility.

IV. DEVICE PHYSICS AND REQUIREMENTS IN THIN-FILM TRANSISTORS

The energy gap structure of a-Si:H is composed of a continuum of states near the conduction and valence edges, i.e., tail states, as well as near the midgap, i.e., deep states. Figure 2 shows diagrams of the density of states (DOS) of a-Si:H and poly-Si band gaps. For a-Si:H, the tail states are contributed by the restructured silicon bonds. The deep states come from the dangling unsaturated silicon bonds. For poly-Si, the situation is more complicated since it has a heterogeneous structure, i.e., grains and grain boundaries. The grain boundary contains high concentrations of defects including restructured and dangling bonds. Depending on the film formation process, the defect states within a grain may be low or high. Therefore, its DOS varies with

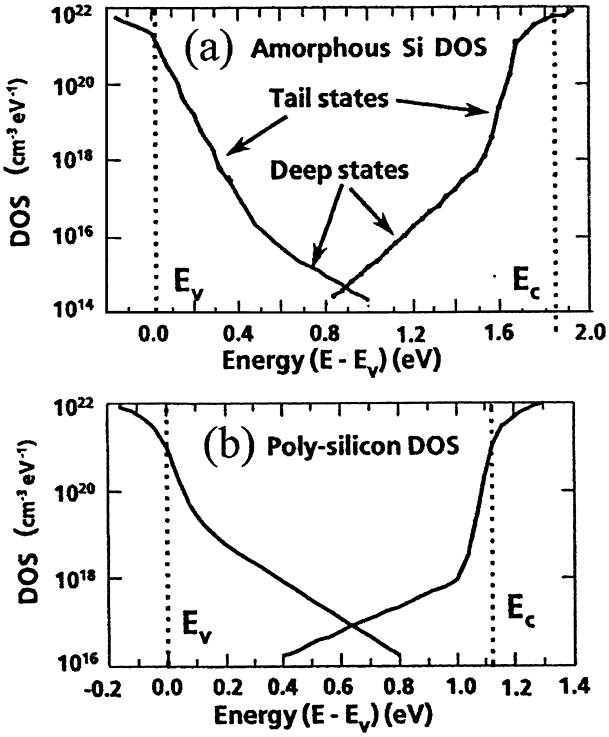


FIGURE 2 Density of states (DOS) within (a) a-Si:H and (b) Poly-Si band gaps. [From Hack, M. (1992). "Proceedings of the First Symposium on Thin Film Transistor Technologies," p. 53, Electrochemical Society, Pennington, NJ. Reproduced by permission of The Electrochemical Society, Inc.]

the location. In general, an effective DOS, which spatially averages all defects at grain boundaries and within grains, is used for poly-Si film. The poly-Si band tail profile is sharper than that of a-Si:H.

Figure 3 shows the band-bending diagram of a TFT when a positive gate voltage is applied and the drain voltage is zero. After traps in the semiconductor layer are filled, an accumulation layer is induced adjacent to the

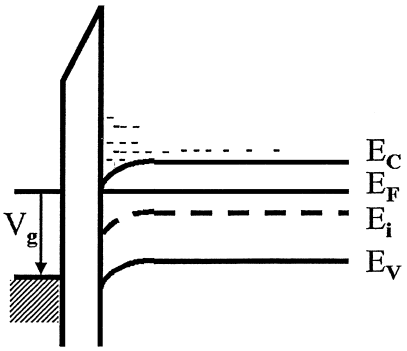


FIGURE 3 Band bending and accumulation layer formation of an n-type TFT at $V_g > 0$ and $V_{ds} = 0$ V.

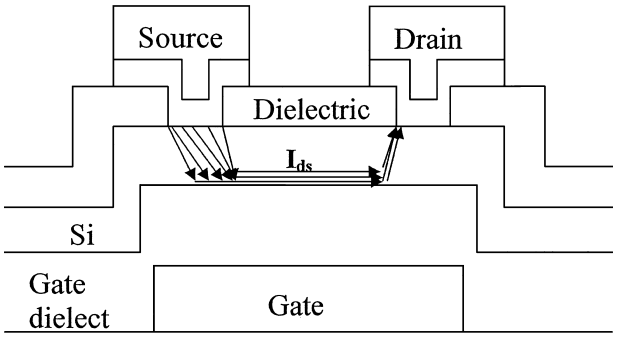


FIGURE 4 Current path in an inverted, staggered trilayer TFT in the on state.

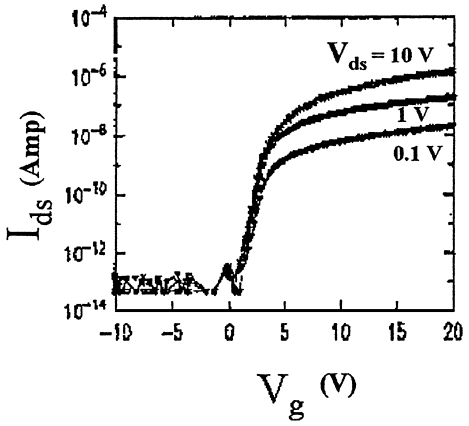


FIGURE 5 Transfer characteristic curves of an inverted, staggered trilayer a-Si:H TFT.

interface. Figure 4 shows the current path in an inverted, staggered TFT in the on state. The current flows from the source contact area to the drain contact area through a very thin region, e.g., less than 200 Å thick, adjacent to the interface of the gate dielectric layer.

The total resistance R_{total} along the current path can be expressed as follows:

$$R_{\text{total}} = R_{\text{source contact}} + R_{\text{Si layer}} + R_{\text{channel interface}} + R_{\text{Si layer}} + R_{\text{drain contact}}, \quad (1)$$

where $R_{\text{source contact}}$ and R_{drain} contact are the contact resistances at the source and drain areas, R_{si}'s are the resistances of the silicon layer, and $R_{\text{channel interface}}$ is the resistance of the thin accumulation layer near the interface.

The total resistance R_{total} can be dominated by any one of the above resistances. Each resistance is a function of the thin-film material properties (e.g., Si/N ratio in SiN_x, dangling bonds in a-Si:H, interface DOS, and roughness), manufacturing processes (e.g., temperature, feed gas composition, PECVD power density), and transistor structure (e.g., bottom or top gate, bi- or trilayer). Therefore, when device properties of two TFTs are compared, these factors cannot be neglected.

For practical applications, there are five parameters that are commonly used to judge the TFT's performance: on current (I_{on}), off current (I_{off}), field-effect mobility (μ_{eff}), threshold voltage (V_{th}), and subthreshold slope (S). Figure 5 shows the transfer characteristic curves of an inverted, staggered, inverted trilayer a-Si:H TFT. Three of the above characteristics, i.e., I_{on}, I_{off}, and S, can be directly read from this curve.

I_{on} is the maximum drain current (I_{ds}) that the transistor can carry at a specified gate voltage (V_{g}) and drain voltage (V_{ds}). It is a function of the density of deep states, i.e., traps within the band gap. The a-Si:H TFT has an I_{on} several orders of magnitude lower than that of a poly-Si TFT because of the large number of trap states in the a-Si:H layer. The I_{on} value can be improved by changing the design of the TFT, such as decreasing the channel length or increasing the channel width. The I_{on} of an a-Si:H TFT increases with the increase in temperature, while the I_{on} of a poly-Si TFT is almost independent of temperature.

I_{off} is the transistor's leakage current, which is commonly taken as the lowest I_{ds} in the transfer characteristics curve. For an a-Si:H TFT, I_{off} is a function of the back-channel interface characteristics, such as the fixed positive charge density and the interface. Its value is affected by the deposition condition and film composition of the SiN_x passivation film. A trilayer TFT usually has a lower I_{off} than the bilayer TFT because of the lower back-channel interface states. For a poly-Si TFT, the leakage current is from the Poole–Frenkel mechanism, i.e., holes tunneling out of traps near the high-voltage drain region. The grain boundary quality and grain size distribution are important factors affecting the leakage current. Conventionally, the leakage current can be reduced by passivating traps with hydrogen. Several methods, e.g., offsetting the drain region, lightly doping the drain (LDD) contact structure, and using a multigate structure, have been proved effective in further reducing the I_{off} of a poly-Si TFT. However, these methods complicate the TFT manufacturing process and may even deteriorate certain transistor characteristics.

S is a subthreshold characteristic corresponding to the amount of V_{g} increase required for a one-order of magnitude increase in I_{ds} under the condition $V_{\text{g}} < V_{\text{th}}$. The value of S is dependent on the deep states density. A small S value is desirable for the fast buildup of the accumulation layer.

Figure 6 shows the I_{ds}-vs-V_{ds} curves of an a-Si:H TFT. Both μ_{eff} and V_{th} can be derived from data in this figure using the following equations:

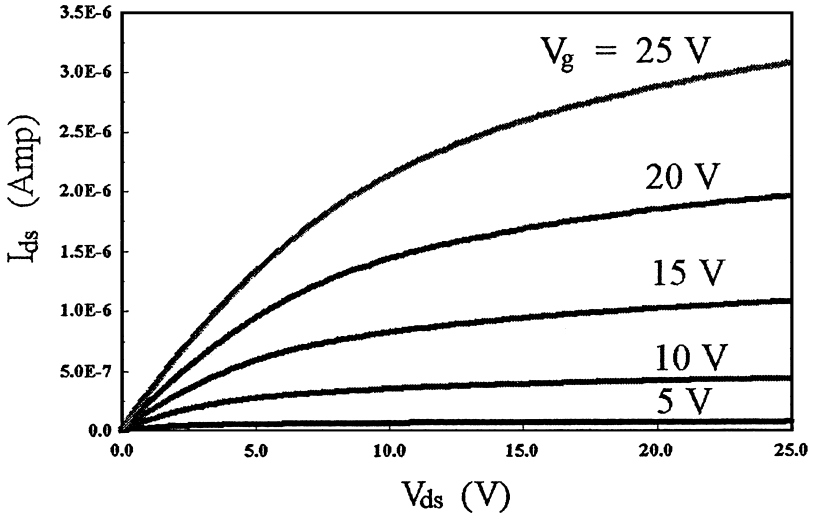


FIGURE 6 I_{ds}-vs-V_{ds} curves of an inverted, staggered trilayer a-Si:H TFT.

for linear-region parameters

$$I_{ds} = \mu_{eff}\frac{W}{L}C_{ox}\left(V_g - V_{th} - \frac{V_{ds}}{2}\right)V_{ds}; \qquad (2)$$

for saturation-region parameters,

$$I_{ds} = \frac{1}{2}\mu_{eff}\left(\frac{W}{L}\right)C_{ox}(V_g - V_{th})^2. \qquad (3)$$

μ_{eff} is controlled by the tail states and the gate voltage. Depending on the measurement region, μ_{eff} can be the saturation mobility, i.e., measured in the saturation region, or the linear mobility, i.e., measured in the linear region. For the LCD application, the linear mobility is critical because the pixel liquid crystal is operated in this region. However, most mobility values reported in the literature are measured in the saturation region because they are easy to measure (e.g., using a high V_{ds} and a high V_g) and the results are repeatable. V_{th} is the minimum V_g required before I_{ds} starts responding to the change in V_g apparently. It is dependent on the deep states of the film. The V_{th} value should be as low as possible to reduce the power consumption. A poly-Si TFT has a mobility one or two orders of magnitude higher than that of an a-Si:H TFT, e.g., >100 Vs vs. < 1.0 cm^2/Vs, because of its lower tail states. The former also has a lower V_{th} than the latter.

The poly-Si TFT often shows the kink effect in its I_{ds} vs. V_{ds} curves, which is a sudden increase in I_{ds} under the high-V_{ds} condition. The same phenomenon has been observed in the silicon-on-insulator (SOI) MOS transistor. It is caused by the impact ionization in the drain region under a high electric field.

Another important characteristic that can be detected from the I_{ds} vs. V_{ds} curves is the ohmic contact formation. If the source or drain contact is nonohmic, e.g., a native oxide layer exists between n^+ and a-Si:H, these curves will be crowded in the low-V_{ds} range. The contact resistance reduces the effectiveness of the gate voltage. A simple method to estimate the contact resistance is to draw a linear (V_{ds}/I_{ds}) vs. L (channel length) curve. The interception of the line to the y-axis ($L = 0$) is the contact resistance contributed by both source and drain.

Device degradation is another important issue in TFT technology. When a gate voltage is applied to the TFT for a period of time, the transistor characteristics deteriorate. For the a-Si:H TFT, the transfer characteristics curve shifts parallel to the positive direction, with a positive V_g bias. The change in V_{th} (ΔV_{th}) follows the power law of the bias time with a power parameter smaller than 0.5. The magnitude of ΔV_{th} is independent of the value of V_{ds}. The degradation phenomenon is composed of two mechanisms. Under the low-bias voltage condition, it is due to the state creation, e.g., bond breaking, in the a-Si:H layer. Under the high-bias voltage condition, it is due to the charge trapping in the gate dielectric layer. For the poly-Si TFT, the degradation mechanism is different. The magnitude of ΔV_{th} is dependent on the values of both V_g and V_{ds}. The change in the subthreshold slope S is dependent on the drain field strength, i.e., under the low-field condition, S is constant; under the high-field condition, S deteriorates appreciably. This is due to the defect creation in different locations within the band gap. Defect creation in TFT is directly related to the life expectancy of the device. Therefore, for any product applications, the degradation of the TFT, such as the maximum tolerable ΔV_{th}, has to be specified.

TABLE III Common Processes for TFT Fabrication

Process(es)	Function(s)	Comments
PECVD	a-Si:H TFT: gate dielectric, a-Si:H, passivation, n^+ ohmic contacts, etc. Poly-Si TFT: a-Si:H and SiO_x	Low temperature, large area, high throughput, one pump-down
Sputtering	Deposition of metals and ITO	High throughput
Wet etching	Various thin-film layers including ITO	Simple setup, low equipment cost, possible residue formation
Plasma etching	Etching of critical layers (n^+ layer, etc.), slope control (a-Si:H TFT gate, contact vias, etc.)	Flexible and versatile chemistry
Lithography: stepper or scanner	Alignment and exposure	Critical large-area throughput issue
Doping	a-Si:H TFT: PECVD n^+ a-Si:H or μc-Si	High throughput
	Poly-Si TFT: ion implantation	Critical large-area throughput issue
Thermal annealing	a-Si:H TFT: repairing of plasma etch damages	Reduction of density of states in films and at interfaces
	Poly-Si TFT: activation of dopant, removal of hydrogen, Si crystallization	
Laser annealing	Poly-Si TFT: Si crystallization, dopant activation	Low temperature, high quality
Plasma hydrogenation	a-Si:H TFT: modification of a-Si:H/gate dielectric interface, precleaning before n^+ deposition, etc.	Low temperature, *in situ* process
	Poly-Si TFT: passivaiton of grain boundaries	Low temperature

V. PROCESS ISSUES IN THIN-FILM TRANSISTOR FABRICATION

Basically, TFT manufacturing processes are similar to some VLSI manufacturing processes. Table III lists common processes that are used to fabricate TFTs.

Plasma-enhanced chemical vapor deposition (PECVD) is commonly used to deposit semiconductor and dielectric films. The advantages of the PECVD process are (1) the low deposition temperature (e.g., less than 350°C), (2) the variable film properties (e.g., nonstoichiometric, nitrogen-rich SiN_x, hydrogenated a-Si:H), (3) the low interface contamination (e.g., deposition of several layers with one pump-down), (4) the large-area capability (e.g., up to a 1×1-m glass substrate), and (5) the high throughput (e.g., high film deposition rate). All the above characteristics can be prepared with the conventional parallel-plate electrode reactor design. Since the material requirements for TFTs are different from those of MOS transistors, their process windows, which involve feed gases, temperature, pressure, power, etc., are very different.

A μc-Si can easily be prepared by PECVD under conditions similar to those for a-Si:H except for a high hydrogen concentration and a high plasma power density. Both conditions damage the surface of the underneath layer, i.e., creating a high interface DOS for the inverted, staggered TFT structure. Therefore, the μc-Si TFT usually has inferior device characteristics, such as a high V_{th} and low μ_{eff}, compared to the a-Si:H TFT. However, the harsh deposition conditions of μc-Si are effective in removing the thin native oxide on a-Si:H surface. Therefore, the ohmic contact is often formed using an n^+ μc-Si film. A μc-Si film has a dual-layer structure. The bottom thin layer, e.g., less than 200 Å thick, is amorphous and the top layer is composed of microcrystals imbedded in an amorphous matrix. The crystal size and volume fraction increase with the film thickness. Although there are reports that poly-Si film can be prepared by PECVD, the deposition rate is very slow and the film quality is not good enough for TFT application.

There are other applications of the PECVD chamber. When the feed gas is hydrogen, the plasma can remove the native oxide on the source/drain vias before the deposition of an n^+ layer. The SiN_x/a-Si:H interface of an a-Si:H TFT can be modified with hydrogen, ammonia, or oxygen plasma to alter the transistor characteristics. Hydrogen plasma is also commonly used to passivate the grain boundaries of a poly-Si film, which lowers the TFT's leakage current. The PECVD reactor can be modified, e.g., by heavily biasing the substrate electrode independent of the cathode power source, to form a doping apparatus, i.e., non-mass-separation ion implanter. Although the doping efficiency is low, the process is complicated, and the power consumption is large, this equipment is effective for large area applications.

The magnetron sputtering technique is widely used to prepare all metal and ITO films for the large-area TFT array. Due to the layered structure of a TFT, the step coverage directly affects the yield. Conformal step coverage is necessary for many deposition processes. ITO is commonly deposited by reactive sputtering, i.e., the feed gas includes argon and oxygen. A high-quality ITO film is in the polycrystalline form, which can be prepared by one of the following two methods: deposited in the amorphous form and subsequently annealed to the polycrystalline

form or directly deposited into the polycrystalline form. The former method is tedious but the process is easy to control. The latter method is straightforward but the process is complicated and often involves substrate heating. A polycrystalline ITO film sometimes has a two-layer structure that forms a reverse-sloped profile after a wet etching process.

Both wet and dry etching methods are used in large-area TFT array fabrication. Although the wet etching process is simple and the equipment is cheap, dry etching is necessary in certain steps. For example, a wet etching process for the n^+ layer often leaves a conductive residue that shortens the source and drain areas. A dry etching process can avoid this problem. In certain designs, several layers of different materials need to be etched as one material with a desired wall profile. Dry etching can match this requirement. However, there are some unique issues in the dry etching of a TFT. For example, since films used in TFTs have properties, e.g., stoichiometry, composition, and bond structures, different from those used in MOS transistors, their dry etching process results, such as the etch rate and selectivity, can be very different. The plasma-etched TFT can have a high V_{th} and a high leakage current due to the radiation damage mechanism. This is because traps are easily generated in PECVD a-Si:H and SiN_x films by the short-wavelength light source emitted from the plasma. This effect is not obvious in VLSI transistors, in which the films are deposited at higher temperatures by the LPCVD or thermal growth method. Fortunately, these traps can be removed with a thermal annealing step.

In some cases, a film has to be etched with a wet process because of the lack of a dry etching capability. For example, ITO is commonly etched with an aqua regia-type solution. This method creates a large undercut and the solution attacks all kinds of metals, which affects the TFT design rules and the production yield. A dry etching method can avoid these problems. However, it is very difficult to etch ITO with a plasma process because the indium and tin compounds are not volatile. There is a great deal ongoing research on this subject around the world.

A stepper is used as the lithography tool for TFT fabrication. Since the substrate size is large, the aligner has to have a large exposure field for the high throughput and a high resolution to reduce the TFT's occupancy of the pixel territory. A typical stepper has an exposure field of several square inches and a resolution higher than 3 μm. Just as in the VLSI case, the aligner's throughput is a bottleneck for the whole production process. For small-size TFT arrays, such as those used in projection displays or viewfinders, a VLSI stepper can be used since the substrate is small and the resolution needs to be high.

In poly-Si TFT fabrication, the deposition or formation of polycrystalline silicon is a critical step. Due to the restriction of the substrate material, the process temperature has to be low, e.g., below 600°C. It is difficult to deposit poly-Si film directly at such a low temperature. For example, the conventional LPCVD poly-Si deposition process requires a temperature higher than 650°C. The grain size and quality increase with the temperature. Recrystallization, in which the film is deposited in an amorphous form at a low temperature and is subsequently transformed into a polycrystalline form, is the most commonly used method. The amorphous film can be deposited by PECVD or LPCVD. A dehydrogenation step is required before the recrystallization step.

Recrystallization can be carried out with thermal or photon energy. Thermal crystallization is a solid-to-solid transformation process. The low-temperature furnace annealing method may take several days to complete the crystallization process. The crystal quality is generally poor because of the inclusion of large numbers of defects within the grain. Recently, the metal-induced crystallization method has been proved effective in reducing the crystallization time. However, it still requires more than 10 hr to complete the process, which is too slow for mass production. The rapid thermal annealing method has been proved effective in completing the crystallization process within a few minutes. However, the high-temperature duration is too long for a low-temperature glass substrate.

The excimer laser annealing (ELA) method, e.g., using XeCl, XeF, or KrF gas, produces the best-quality poly-Si for TFTs. For example, a high μ_{eff}, e.g., >200 cm^2/Vs, and a low V_{th}, e.g., <2 V, can routinely be achieved with this method. The principle of ELA is to transform the photon energy into heat that is high enough to melt silicon. Poly-Si silicon is formed after the cooling step. Since the wavelength of the laser beam is short, e.g., in the deep-UV range, all the incident energy is absorbed within the top 100 Å of the film. The pulse time is so short that the substrate is not exposed to the high heat during the process. Therefore, a low-temperature glass can be used. The ELA crystallized silicon has large, low-defect grains that are responsible for the good TFT characteristics. However, the laser crystallization process is complicated, e.g., multipath exposure, beam-to-beam overlap, substrate heating, and energy density and duration controls. It is difficult to apply ELA to a large-area substrate with a high throughput. It has been proposed that a-Si:H and poly-Si TFTs can be integrated into the same display, for example, where only the drivers are fabricated with ELA crystallized poly-Si and the pixel area is driven by a-Si:H technology. This would reduce the impact of the low-poly-Si throughput issue. However, other thin-film material properties can become an important issue in this approach because the best gate dielectric for the a-Si:H TFT is often not the best gate dielectric for the poly-Si

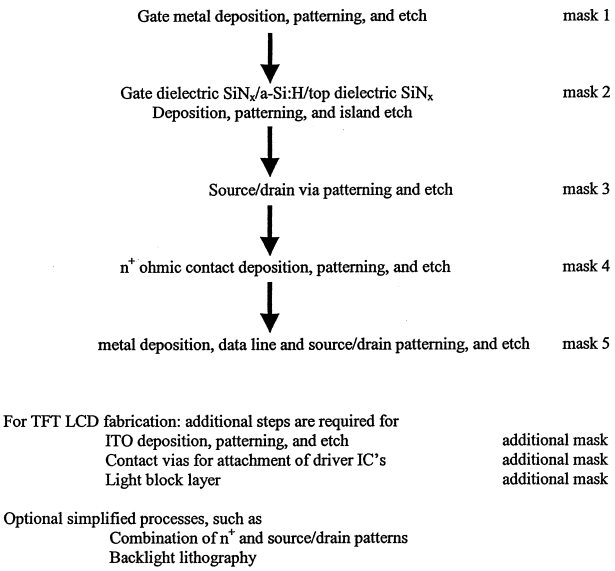


FIGURE 7 Fabrication process of an inverted, staggered trilayer a-Si:H TFT.

TFT. In general, despite many research efforts spent on poly-Si TFT, it is still an open area requiring further investigation.

Figure 7 shows the flowchart of a simplified fabrication process for an inverted, staggered trilayer a-Si:H TFT. Five masks are required to complete the TFT. There are several key steps that affect the TFT performance, reliability, and yield: gate line slope control, deposition of the trilayer, source/drain alignment, contact resistance, and n^+ dry etching.

The step coverage of films deposited on the gate line is directly influenced by the profile of the gate line. A dry etch process can supply a gate line with a proper slope angle. Properties of the two interfaces of the trilayer film influence some major TFT characteristics, such as μ_{eff} and I_{off}. These interface characteristics can be manipulated by the plasma conditions of the PECVD chamber at various stages of the process. The accuracy of the source/drain-to-gate alignment is critical to the function of the TFT. A backlight lithography that uses the gate pattern as the mask guarantees a self-aligned source/drain structure. The overlap of the source/drain with the gate is small, e.g., 0.5 μm. Therefore, it is a simple and reliable process for source/drain definition and has been adapted for many production processes. The contact resistance is dependent on the cleanness of the source/drain via. Ohmic contacts, which requires the removal of the native oxide in the via region, can be obtained with a wet dip or a dry hydrogen method before n^+ deposition. n^+ residing between the source and the drain, which is often left after the wet etch step, can cause transistor failure. A dry etching process can avoid this problem. However, a damage-repair step has to be carried out.

For the mass production of TFT arrays, a high throughput is most desirable. The simplest way to increase the throughput is to decrease the number of masks. Conventionally, the inverted, staggered, trilayer TFT requires one more mask to fabricate than the bilayer TFT. However, there are many new designs that can reduce the number of masking steps of the former to equal that of the latter. A complete TFT array for LCD application should include ITO pixels and contact pads, which increases the mask number. So far, the minimum number required to prepare a complete inverted, staggered, trilayer a-Si:H TFT array including ITO and contact patterns is four.

Figure 8 shows a simplified poly-Si TFT fabrication flowchart. The process is similar to that of a MOS transistor except for the silicon crystallization and the dopant activation steps. For small-size TFT array fabrication, quartz is used as the substrate. A high-temperature method can be used. For large-size TFT array fabrication, a low-temperature process, such as laser annealing, has to be used. The hydrogenation step can be done with a conventional or a high-density plasma reactor. It can also be carried out by annealing a high-hydrogen content PECVD SiN_x film deposited on top of the TFT. Currently, there are very few large-area poly-Si TFT manufacturing processes.

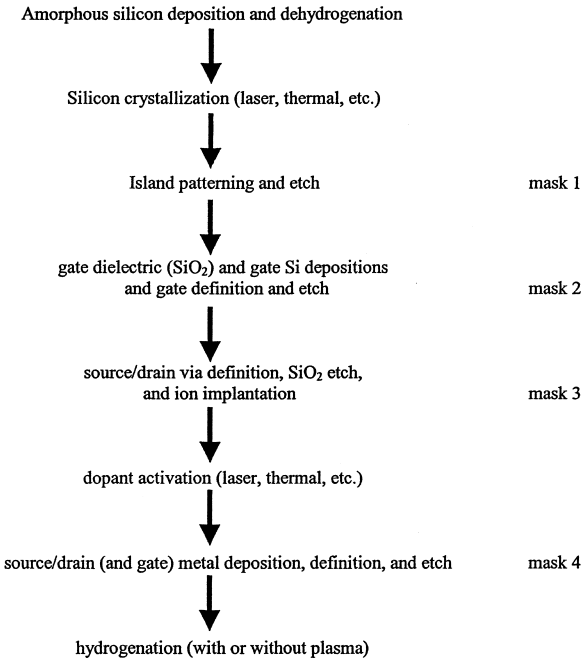


FIGURE 8 Fabrication process of a coplanar poly-Si TFT.

The main limitation is the low-temperature crystallization step.

VI. SUMMARY AND TECHNOLOGY TRENDS

Since its invention more than 40 years, TFT has played a minor or negligible role in industries until the successful fabrication of the a-Si:H TFT. Today, TFT LCD alone is a major industry with a predicted high growth rate for the next few decades. TFT applications have been spread out from laptop computers to all major consumer and industry products.

Due to the tremendous amount of research and development activities on a-Si:H and poly-Si TFTs, the basic physics and chemistry of the thin-film materials and the device operation principles are well understood. The advantages and disadvantages of TFTs over other semiconductor devices, such as VLSI MOS transistors, have been thoroughly studied. In the long run, TFT will be a technology that complements rather than competes with VLSIC in applications requiring large-area and low-temperature substrates. Most of these applications cannot be fulfilled with non-TFT technologies.

Although the a-Si:H TFT has the disadvantage of a low field-effect mobility, it is suitable for applications that require a local switching capability but no logic functions. The poly-Si TFT has a major advantage over the a-Si:H TFT in mobility. However, a bottleneck in poly-Si TFT technology is the mass production process, especially in the low-temperature poly-Si formation step. A method for producing a high-quality poly-Si film with a high-throughput capability using a low-temperature substrate is most desirable.

SEE ALSO THE FOLLOWING ARTICLES

BIPOLAR TRANSISTORS • DIAMOND FILMS, ELECTRICAL PROPERTIES • FIELD-EFFECT TRANSISTORS • LIQUID CRYSTAL DEVICES • SUPERLATTICES

BIBLIOGRAPHY

Hack, M., Shaw, J. G., LeComber, P. G., and Willums, M. (1990). "Numerical simulations of amorphous and polycrystalline silicon thin-film transistors," *Jpn. J. Appl. Phys.* **29,** L2360–L2362.

Kuo, Y. (1995). "Plasma etching and deposition for a-Si:H thin film transistors," *J. Electrochem. Soc.* **142,** 2486–2506.

Kuo, Y. (ed.) (1992, 1994, 1996, 1998). "Proceedings of Thin Film Transistor Technologies I, II, III, and IV," Vols. 92-24, 94-35, 96-23, and 98-22, Electrochemical Society, Pennington, NJ.

Migliorato, P. (1992). Thin-film transistors. *In* "Encyclopedia of Physical Science and Technology," 3rd ed., Vol. 16, pp. 749–761, Academic Press, San Diego, CA.

Migliorato, P., and Meakin, D. B. (1987). "Materials properties and characteristics of thin-film silicon transistors for large area electronics," *Appl. Surf. Sci.* **30,** 353–371.

Powell, M. J., van Berkel C., and French, I. D. (1987). "The resolution of a-Si:H thin film transistor instability mechanism," *J. Non-Crystal. Solids* **97/98,** 321–324.

Shur, M. S., Jacunski, M. D., Slade, H. C., Owusu, A. A., Ytterdal, T., and Hack, M. (1996). SPICE models for amorphous silicon and polysilicon thin film transistors. *In* "Proceedings of Thin Film Transistor Technologies III," Vol. 96-23, pp. 242–259, Electrochemical Society, Pennington, NJ.

Street, R. A. (1991). "Hydrogenated Amorphous Silicon," Cambridge University Press, Cambridge.



Ultra Large-Scale Integration Design

Saburo Muroga
University of Illinois, Urbana-Champaign

I. Outline of Design and Fabrication of a Chip
II. Logic Networks
III. Memory
IV. Comparison of Design of Logic Networks and Memory
V. Limitations of Integration Size
VI. Different Design Approaches
VII. Full-Custom Design Approach
VIII. Semi-Custom Design Approaches
IX. Comparison of Custom Design Approaches
X. Reconfigurable Computers with Field-Programmable Gate Arrays
XI. Design Tools and Future Problems

GLOSSARY

Chip Rectangular thin sheet of semiconductor up to about 2 cm on a side and about 0.5 mm thick. Integrated circuits are realized on it. Often called a die (plural: dies).

Computer-aided design (CAD) Design of computers, electronic circuits, or other miscellaneous items, aided by a program run on a computer.

EEPROM A type of nonvolatile semiconductor memory. EEPROM stands for electrical erasable programmable memory.

Flash memory Nonvolatile semiconductor memory of large capacity, where a large number of EEPROMs are compactly laid out on an IC chip.

FPGA A type of programmable hardware, where blocks, each of which consists of few logic gates and memory, are laid out along with interconnections in a matrix form on a chip. Hardware for desired logic functions can be realized as easily as software. FPGA stands for field-programmable gate array.

Full-custom design approach Most time-consuming design approach of a chip, in which every stage, in

particular, layout is most deliberately designed to attain the smallest chip size and the best electronic performance.

Integrated circuit (IC) Electronic circuits entirely realized on a chip, in contrast to old realization where components (e.g., transistors and diodes) are separately manufactured and then assembled into an electronic circuit.

Integration size (or integration scale) Number of devices (i.e., components of electronic circuits, such as transistors, diodes, and resistors) realized on a single chip. When packing many devices into a single chip was difficult, the number of devices in a single chip was meaningful because of difference in cost and performance. Chips of different integration sizes were classified by SSI (small-scale integration), MSI (medium-scale integration), LSI (large-scale integration), VLSI (very large-scale integration), and ULSI (ultra large-scale integration). For example, LSI means integration size of a few hundred to a few thousand devices on a chip and ULSI far more, though the number of devices for each size has not been formally defined. In recent years, such classification becomes less meaningful and LSI has been used to mean any large number of devices, including VLSI and ULSI.

Layout Arrangement pattern of electronic circuits which is to be realized onto a chip as an integrated circuit.

Logic family A type of electronic circuit to realize a logic gate which can be realized with MOSFETs, bipolar transistors, or other devices.

Logic gate A basic element for a logic operation (such as NOT-AND, i.e., NAND operation) in a computer. A network of logic gates is called a logic network; a computer contains a large number of logic networks.

Mask A glass plate, covered with photographic film, that has the image of the layout of electronic circuits to be realized on an IC chip.

Microprocessor A processor, with some amount of memory, realized on a single IC chip. A computer, whether it is a desktop, palmtop, server, mainframe, or supercomputer, consists of microprocessors, memory, input/output control circuits, and other auxiliary logic networks, which can be realized in a single or many chips.

PLD A type of programmable hardware on an IC chip, where interconnections are pre-laid in straight lines in a matrix form along with transistors at fixed places. Desired logic functions can be realized only by connecting appropriate interconnections, as easily as software, without design and layout of transistor circuits for each design.

Semi-custom design approaches Design approaches of a chip, in which some design stages are simplified to reduce design time and expenses, sacrificing the compactness of chip size or electronic performance. They are used when the full-custom design approach, which every design stage is deliberately done, is too time-consuming and expensive.

Wafer A round, thin disk of semiconductor, which is 8 or 12 in. in diameter and about 0.5 mm thick. On this, a few hundred to thousands of chips are fabricated.

Yield Ratio of the number of faultless products to the number of all manufactured products, where product means chip, wafer, or possibly others.

THE DESIGN of electronic circuits on a chip in very large-scale integration (VLSI) or ultra large-scale integration (ULSI) means a few thousand or more components (devices such as transistors, diodes, and resistors) of electronic circuits, on a single chip of semiconductor. An IC chip, as shown on a finger in Fig. 1, is at most about 2 cm on a side and about 0.5 mm thick. Because of utmost compaction of a large number of components, the design of such electronic circuits is very complex and requires very strenuous, time-consuming engineering efforts. Compaction of so many electronic components in a tiny space is being improved over decades by the progress of microminiaturization technology. Now millions of devices can be packed on a chip and more will be packed by further miniaturization. Such miniaturization has many important advantages: reduction of cost of electronic circuits, electrical power consumption, space, and weight, and also improvement of speed and reliability. Recent technological marvels in computers, consumer electronics, and others are due to the progress of IC technology. Computers, which used to be room size, are replaced by much smaller personal computers with far greater performance. Many new electronic products are coming into being with IC chips. Recent advent of new types of semiconductor memory, such as nonvolatile memory tremendously broadens the applications of IC chips, such as digital cameras, mobile computers, and reconfigurable computers. A complex integrated circuit to be realized on an IC chip with the smallest size and highest speed can be designed by full-custom approach (the most deliberate design approach) but the full-custom approach is too time-consuming and expensive for many products. The full-custom design of such a very complex IC chip can easily take several dozens of engineers about four years to complete. Thus many IC chips are designed with semi-custom approaches that reduce design time and expenses, while sacrificing the compactness of the chip and its speed.

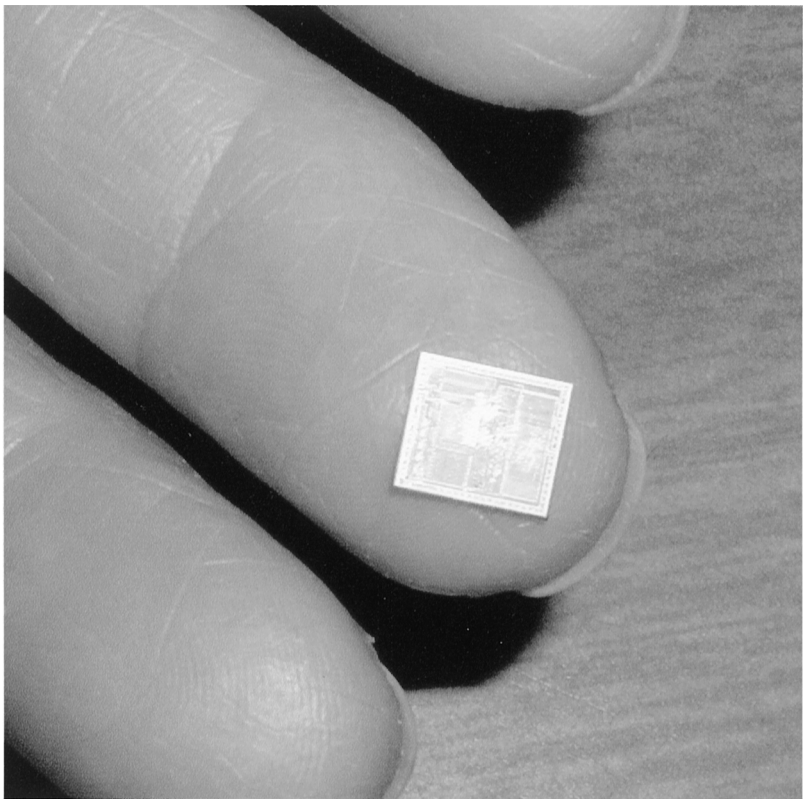


FIGURE 1 An IC chip on a finger.

I. OUTLINE OF DESIGN AND FABRICATION OF A CHIP

A. Design

A chip of any integration size is made by a long, complex sequence of steps of design and fabrication, as illustrated in Fig. 2. First, the architecture of a computer to be realized on the chip is defined and designed in detail as in Fig. 2a. Then further details, in other words, design of logic networks (logic design) and electronic circuits is done in (b). Then in (c), the layout, i.e., location pattern of components of the electronic circuits to be realized on a chip, is made. For a VLSI or ULSI chip, the layout, if to be most compacted, is the most time-consuming stage among all the design stages. Design will be discussed in detail later again.

B. Fabrication

Electronic circuits are fabricated with several different materials on a chip. These materials are typically aluminum, n-type silicon (pure silicon mixed with a trace of the donor-type impurity such as phosphorus or arsenic), n^+-type silicon (pure silicon mixed with a greater amount of the donor-type impurity than the n-type silicon), p-type silicon (pure silicon mixed with a trace of the acceptor-type impurity such as boron or gallium), p^+-type silicon (pure silicon mixed with a greater amount of the acceptor-type impurity than the p-type silicon), and polysilicon. For example, an n-channel MOSFET (a metal-oxide-semiconductor field effect transistor), shown in Fig. 3a, which is often abbreviated as nMOS, consists of two regions of n^+-type silicon in the wafer of p-type silicon, an insulation layer of silicon dioxide (which is usually called silicon oxide for simplicity), and three separate metal strips. The two regions of n^+-type silicon are called the source and drain. The electrode located between them is called the gate (this term is different from "logic gate") and is often realized by polysilicon, replacing aluminum, but recently metal such as tungsten. Then n-channel MOSFET is denoted by the symbol shown in Fig. 3b. Connections among n-channel MOSFETs are made of aluminum, copper, polysilicon, or n^+-type silicon. Thus, the layout of electronic circuits actually consists of many layouts, corresponding to layers of these different materials which are to be made by different processing methods. The layout of electronic circuits is drawn by draftpeople in Fig. 2c, being converted into digital information. The layout for each layer is transferred by electron beam onto a reticle (often also called a mask) which is a photographic glass plate, in Fig. 2d. In

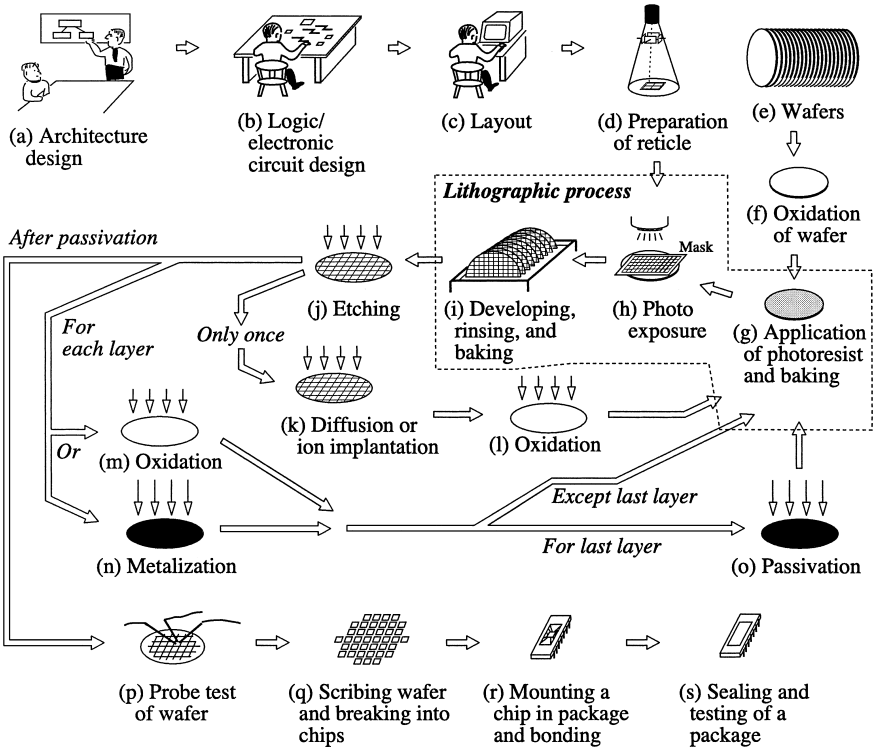


FIGURE 2 The entire sequence of design and fabrication of an IC package.

Fig. 2h, the layout pattern on the reticle is repeated many times on a wafer by an optical system called a stepper during the lithographic process of the fabrication of the IC chip. About two dozens reticles (or masks) are prepared for the entire chip fabrication, letting each reticle contain the layout pattern for a different layer of material for the electronic circuits.

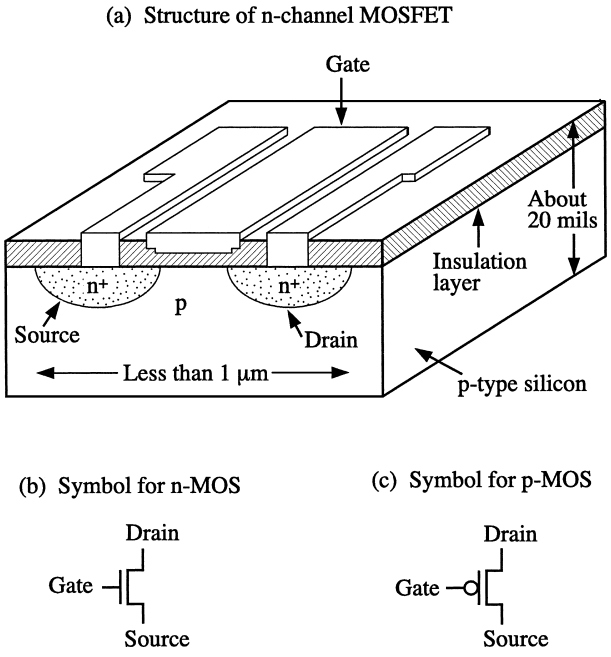


FIGURE 3 MOSFET.

Fabrication of an integrated circuit chip is the realization of electronic circuits on a wafer which is sliced from a large sausage-shaped ingot of a single crystal of semiconductor, as illustrated in Fig. 2e. Fabrication starts with oxidation of a surface of wafer in Fig. 2f. Fabrication consists of "processing" for physical and chemical treatments of the wafer, and "packaging" for assembling a processed chip and a container into an integrated circuit package. The processing consists of lithographic process shown in (g)-(i) in Fig. 2 and other processes, such as ion implantation, diffusion, metallization, or etching, shown in (j)-(n). The processing is repeated for each layer. Finally in Fig. 2o, a protective layer from scratches, moisture or chemicals is formed.

As an illustration, the fabrication of the n-channel MOSFET shown in Fig. 3a is illustrated in detail in Fig. 4 with a simple example, along with Fig. 5 which illustrates the lithographic process. The oxidized wafer of p-type silicon shown in Fig. 4a is processed by the lithographic process shown in Fig. 5 to form windows in the insulation layer (the oxidized silicon), as shown in Fig. 4b. In the lithographic process, the oxidized surface of the wafer in Fig. 5a is coated with a photoresist, which is a lacquer-like material, as shown in Fig. 5b. The photographic image on

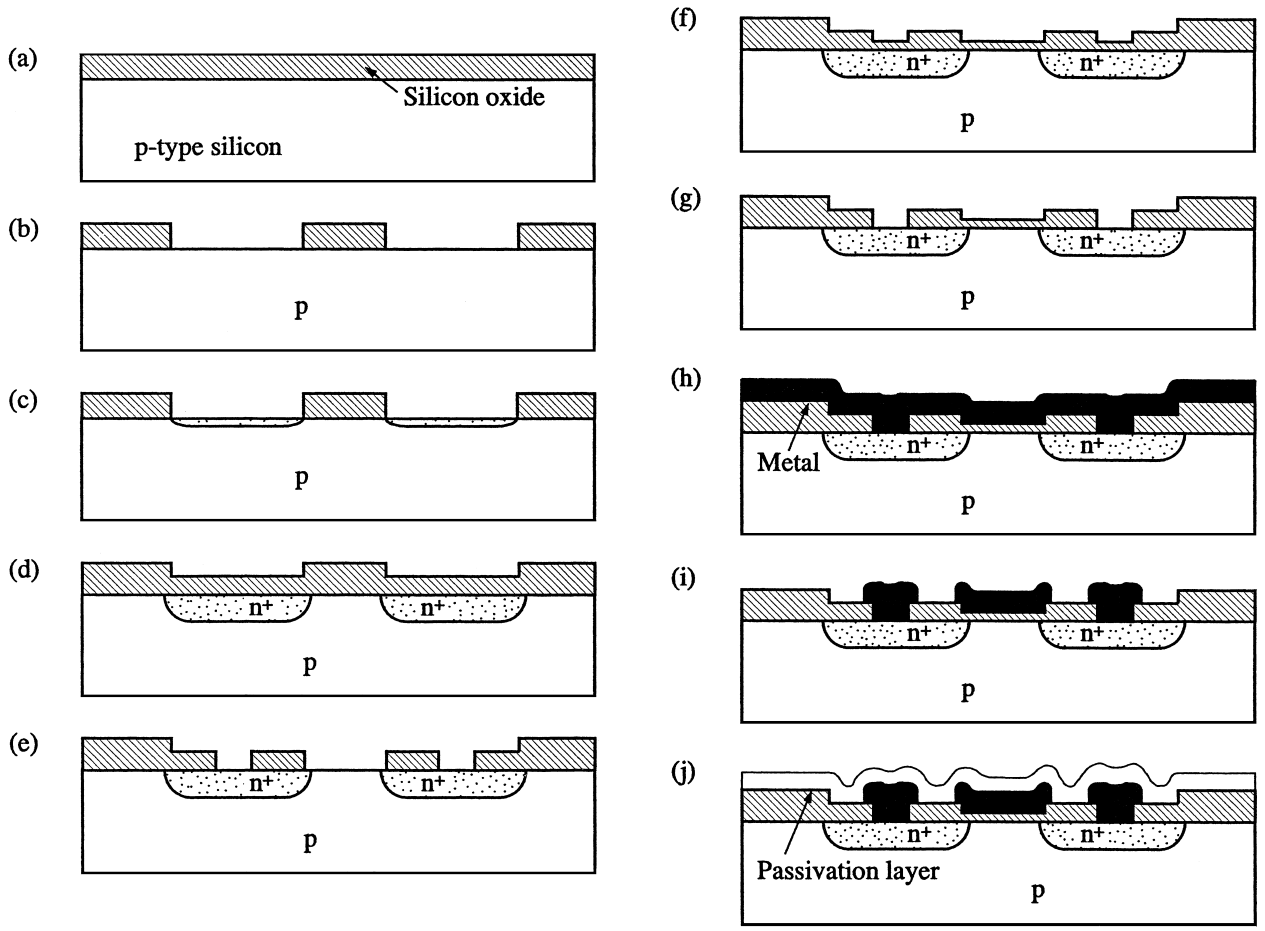


FIGURE 4 Fabrication of n-cahmmel MOSFET (not scaled).

a mask (or a reticle) is projected on it in Fig. 5c with ultraviolet light. The areas of the photoresist corresponding to the transparent areas of the mask change their property because of the ultraviolet light. After removing the mask, these areas of the photoresist are removed by a solvent in Fig. 5d. The other areas remain. Then the wafer is baked to harden the photoresist and then soaked in a chemical solution to etch away the insulation layer where it is not protected by the photoresist, as shown in Fig. 5e. The photoresist is then removed by soaking the wafer in acids, as shown in Fig. 5f, i.e., Fig. 4b. A thin surface layer of phosphorus concentration is formed, resulting in the n^+regions shown in Fig. 4c. Then oxidation in a furnace grows a thin protective insulation layer of silicon oxide (more precisely, silicon dioxide) over the diffused regions, as shown in Fig. 4d and at the same time heat for this oxidation further diffuses phosphorus (n^+-type silicon) into the silicon wafer. A window for gate is opened, as shown in Fig. 4e, by using the lithographic process of Fig. 5 again. A thin oxide layer is grown in the window, as shown in Fig. 4f. Then the insulation layer over the diffused regions is removed to form contact windows (each contact between diffusion region and a metal electrode) by the third lithographic process of Fig. 5, as shown in Fig. 4g. Metal is deposited over the entire surface as shown in Fig. 4h. Then metal is removed everywhere except at the contacts and the gate electrode region by the fourth lithographic process of Fig. 5, as shown in Fig. 4i. Also at this stage, connections among MOSFETs and connections for the power supply and ground are also made by metal. For second and further layers of connections, the formation of an insulation layer (Fig. 2m) and a connection layer (Fig. 2n) along with contact windows is repeated. Finally, a layer to protect against chemical contamination or scratch is formed as shown in Fig. 4j, i.e., Fig. 2o, and this is called passivation.

The processing which is actually used in industry is far more complex than that illustrated in Figs. 4 and 5. The processing consists of more than a hundred steps, taking at least a few weeks. Bipolar transistors along with diodes and resistors can be similarly fabricated. The number of masks to fabricate MOSFET circuits, or bipolar transistor circuits, is about two dozen.

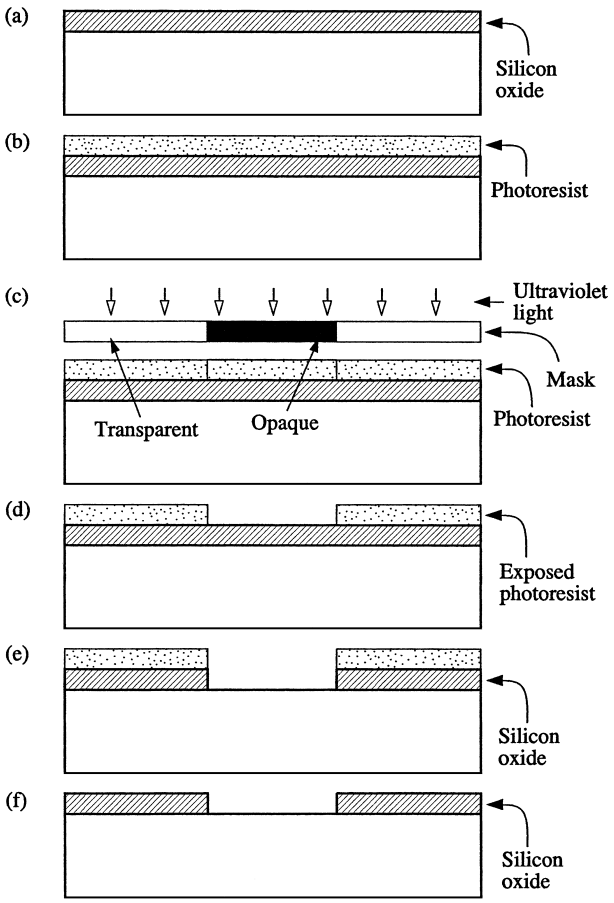


FIGURE 5 Lithographic process (not scaled).

Each chip on the wafer is tested in Fig. 2p and then we have chips by scribing and breaking the wafer as illustrated in Fig. 2q. Then faulty chips are rejected. Then a chip is placed inside a package (plastic or ceramic), bonding the pads on the periphery of the chip to the metal pins of the package by metal wires, as illustrated in Fig. 2r. Then the package is sealed with a lid and is tested, as illustrated in Fig. 2s. Thus we have an IC package.

The entire sequence of design and fabrication is very long and complex, and mistakes or faults can sneak in at any time and any place, making a completed integrated circuit package faulty. At every stage throughout the entire sequence, elaborate, complex, and time-consuming tests and verifications are absolutely required.

After fabrication, IC packages are mounted and interconnected on printed circuit (PC) boards. Printed circuit boards are assembled into cabinets, or a desk top computer case, to complete the computer.

II. LOGIC NETWORKS

At the system design stage illustrated in Fig. 2a, an architecture of a computer to be realized on a chip is designed with an appropriate connection configuration of logic networks and memory. Being tailored to the architecture, computational tasks to be realized are converted into logic operations, each of which is to be realized by a logic network. At the logic design stage in Fig. 2b, logic networks are designed using logic gates (e.g., AND gates, OR gates, NOT gates, NAND gates, and NOR gates) which are unique to the logic family adopted. On a VLSI or ULSI chip, we need to design logic networks with millions of logic gates.

Then, at the electronic design stage which is usually done with logic design at the same time, these logic networks are converted into electronic circuits, considering speed, voltages, currents, and other electronic parameters. Logic gates can be realized with different types of electronic circuits, using nMOS, bipolar transistors, or other devices. These different types of electronic circuits are called logic families. Because each logic family has unique features in speed, power consumption, logic operations, and others, designers have to choose the most appropriate logic families, in designing a chip, by considering design motivations and requirements of the chip. Logic families based on bipolar transistors have higher speed, delivering large output power, but larger power consumption than those based on nMOSFETs. Emitter-coupled logic (ECL) has the fastest speed, but it also has the largest power consumption. Because of the use of resistors and bipolar transistors, ECL occupies a large area on a chip and accordingly is expensive. A typical logic gate realized in ECL is shown in Fig. 6. Emitter-coupled logic has been used when high speed along with delivery of large output power is required regardless of the cost. Logic families based on bipolar transistors, including ECL, have not been often used, as CMOS, a logic family based on MOSFETs, is improved. Logic families based on MOSFETs have generally much lower power consumption, and smaller areas than those based on bipolar transistors. Among these MOSFET families, the static nMOS

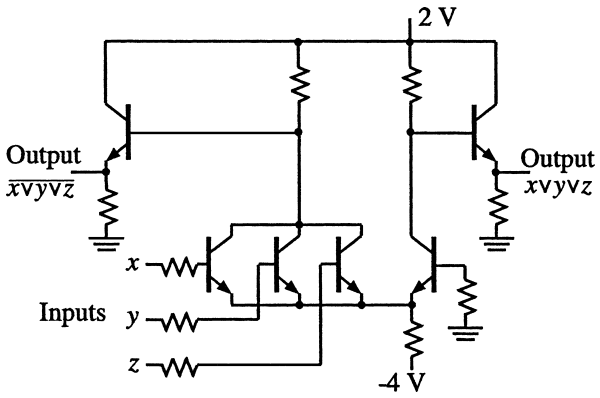


FIGURE 6 Emitter-coupled logic.

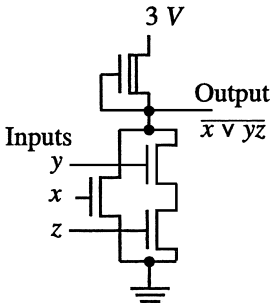


FIGURE 7 Static nMOS.

which consists of n-channel MOSFETs is the fastest and has the largest power consumption. A typical logic gate in static nMOS is shown in Fig. 7. Static nMOS is the family with the smallest area on a chip. (If the speed can be sacrificed, static nMOS can be made smaller with a lower power consumption.) CMOS, which stands for complementary MOS, consists of n-channel MOSFETs and p-channel MOSFETs, where a p-channel MOSFET has p-type silicon and n-type silicon regions exchanged in an n-channel MOSFET and its symbol is shown in Fig. 3c. CMOS consumes the least power among all logic families, but if we increase the rate at which signals change, CMOS can consume power comparable to that of ECL. CMOS has lower speed and greater area than static nMOS. A typical logic gate in CMOS is shown in Fig. 8. For VLSI or ULSI design, CMOS is the most important logic family because a far greater number of transistors can be packed on a chip without being limited by the maximum power consumption than nMOS or ECL. In addition to the smallest power consumption, CMOS has the unique advantage of being reliable for changes in temperature and power supply voltage and also for noise. Thus, CMOS has been used widely in all kinds of computers, including those inside wristwatches and cameras. Domino-CMOS is a modification of the above CMOS combined with clock pulses. The modification gives domino-CMOS a higher speed than ordinary (static) CMOS. There are other logic families such as cascode voltage switch logic (CVSL).

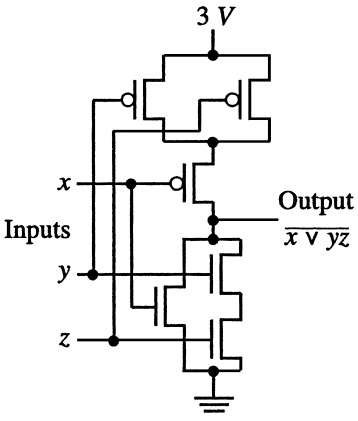


FIGURE 8 Complementary MOS.

III. MEMORY

Memory is to store data or programs and its technology has made remarkable progress in recent years, stimulating the introduction of new types of products. In contrast to magnetic discs, memory realized with semiconductor is very fast but expensive. A memory of any type (RAMs, ROMs, or others described in the following) consists of memory cells, an address decoder, a memory register, and a read-write control, as shown in Fig. 9. Memory cells (shown with many small squares), each of which stores information 0 or 1, are arranged in a matrix form. When a memory address code, $(A_1 A_2 \ldots A_n)$, is given to a decoder, one row of memory cells (a memory word) that corresponds to the address code is selected by an output of the decoder. Then the information, 0 or 1, stored in these cells is read out as outputs $f_1, f_2, \ldots f_t$ as a word into the memory register by the read-write control, or the information stored in the memory register is written into the memory cells in the selected memory word.

A. RAM

Random access memory (RAM), is the most important type of memory in semiconductors because of high-speed access to a memory word at any address. Information can be written in and read out of a memory at any of its address locations with the same access time, unlike floppy disks. Depending on how memory cells are realized by an electronic circuit, RAMs are further classified into two types—static RAM (SRAM) and dynamic RAM (DRAM). Each

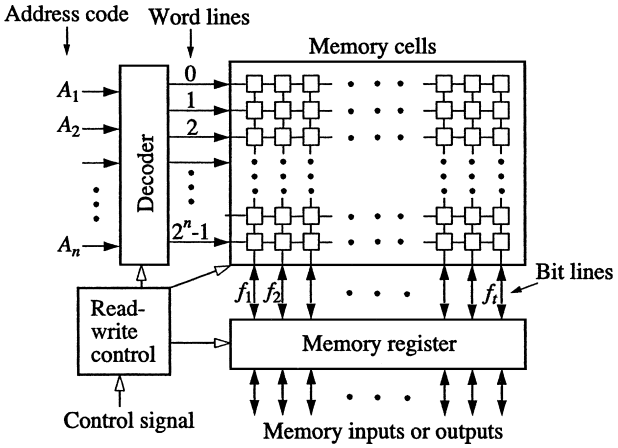


FIGURE 9 A memory structure.

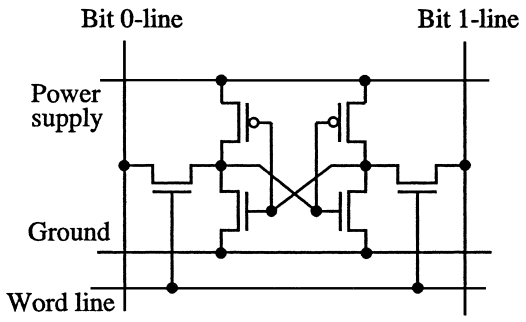


FIGURE 10 CMOS memory cell for static RAM.

cell of a static RAM consists of a transistor circuit (called flip-flop) realized in CMOS, as shown in Fig. 10. Thus as long as a power supply is connected, stored information is maintained. Each memory cell of a dynamic RAM keeps information by storing an electric charge on a very small capacitance, which is usually called storage capacitance. Information 0 and 1 is represented by electric charge and no charge, respectively, on the storage capacitance in a memory cell. Figure 11 shows an example of a dynamic RAM, where only one MOSFET is connected to the ground through a storage capacitance C. Since an electric charge stored on capacitance C gradually leaks (mostly from the diffusion region of the MOSFET to the ground), the same information must be written again before the charge completely leaks out. In other words, the memory cell must be refreshed. Because a single MOSFET is used instead of many MOSFETs used for SRAM in Fig. 10, a dynamic RAM occupies a much smaller area. Thus in the same chip area of a static RAM, a dynamic RAM can be packed about 4–10 times more, depending on the technology, but SRAM is usually much faster than DRAM in reading and writing.

Every four or four and a half years, RAM with memory capacity four times greater than the previous one has been introduced with a new processing technology. In other words, the number of transistors on a chip doubles every year and half and this empirical observation is called Moore's law. When the new RAM is introduced, the price of a memory package is initially high, and goes down due to manufacturing improvements.

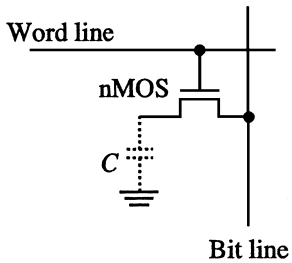


FIGURE 11 Memory cell for dynamic RAM.

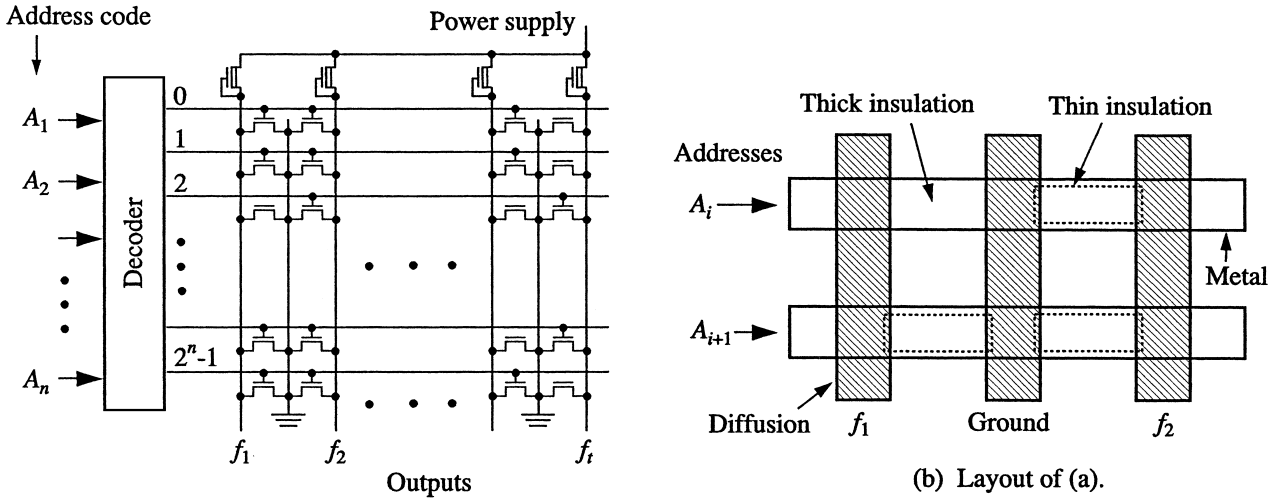
The performance of DRAMs became a bottleneck for faster operation of a computer. Conventional DRAMs have been replaced by new types of DRAM, such as extended data out (EDO) DRAM, double-data rate (DDR), synchronous DRAM (SDRAM), and Rambus DRAM.

B. ROM

ROM stands for read-only memory. Information can be read out like with RAM but cannot be written in as freely as in the case of RAMs. It is further classified into mask-programmable ROM and field-programmable ROM. The latter is further classified into PROM, EPROM, and EEROM.

1. Mask-Programmable ROM

In mask-programmable ROM, information is permanently stored by masks during fabrication by semiconductor manufacturers, and once set, it cannot be changed. The packing density of a mask-programmable ROM is higher (roughly two or three times) than that of a dynamic RAM whose packing density is higher than a static RAM. A ROM is much easier to manufacture than a dynamic RAM because of its simpler storage mechanism, as explained in Fig. 12a. For a chosen horizontal line in the memory structure shown in Fig. 9, only vertical lines that are not connected to that horizontal line through MOSFETs have high voltages. The connection of the gates of MOSFETs for this purpose can be realized as shown in Fig. 12b, where MOSFETs are formed if the insulation layer underneath the gates is thin enough, as shown by the dotted lines. If that portion of the layer is thick, a MOSFET is not formed. Figure 12b can be prepared by changing only the mask for making thin regions in the insulation layer. The custom preparation of only one mask, in addition to small size, makes ROM very economical. Thus ROM, in high-volume production, is the most economical among all memories, such as SRAM, DRAM, PROM, and others. When computer programs are completely debugged and need not be rewritten during computation, there is no point in using RAMs. Programs can be packed into a much smaller space of a ROM with lower cost and programs in this form are called firmware. Another advantage of using ROMs for software is the convenience that a ROM cartridge with software stored can simply be inserted into a calculator, game machine, or portable computer, with metal contacts, but without any mechanically moving equipment, whereas if software is to be stored in a floppy disk, we need mechanically moving equipment such as a disk drive, which is unreliable, expensive, and power-consuming.



(a) Structure of a ROM with MOSFETs.

(b) Layout of (a).

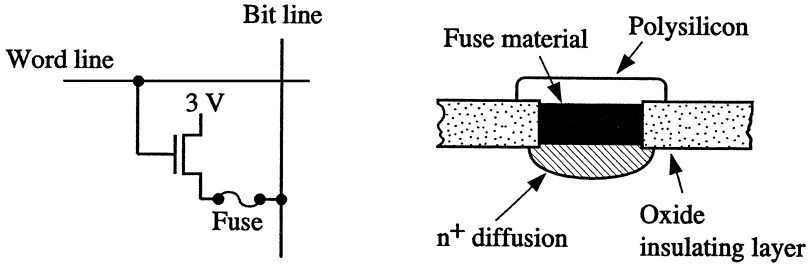
FIGURE 12 ROM with MOSFETs.

2. Field-Programmable ROM

Field-programmable ROMs (FPROM) are ROMs where users can store information by themselves and have the following types of ROMs: PROM (programmable ROM), EPROM (erasable programmable ROM) and EEPROM (electrically erasable programmable ROM). For small-volume production of ROM, custom preparation of even a single mask for connections may be too expensive, or delivery time from a semiconductor manufacturer (usually a few weeks) may be too long. In this case, FPROMs are useful.

3. PROM (Programmable Read-Only Memory)

FPROM, into which information can be written only once (permanently), is called programmable ROM (PROM). A commonly used structure of PROMs is illustrated in Fig. 13a. A MOSFET and a fuse, or antifuse, are connected in series and placed in each memory cell in Fig. 9. A cross section of the fuse is shown in Fig. 13b. Two lines, one in polysilicon and the other in diffusion, are initially connected. To store information, users blow the fuses in desired memory cells by feeding heavy currents, disconnecting the two lines. On a chip, electronic circuits are provided to supply these heavy currents, adding an extra large area on the chip. Antifuses connect two lines that are initially not connected and once connected, the two lines cannot be disconnected. An antifuse has its cross section which is similar to the fuse in Fig. 13b with different material. Use of fuses or antifuses has the following trade-off. If fuses are used, the majority of fuses must be usually blown up and this is very time-consuming. If antifuses are used, there is no need for connecting the majority of antifuses but connected two lines have rather high resistance and unconnected two lines have high parasitic capacitance between them.



(a) MOSFET with fuse

(b) Fuse

FIGURE 13 PROM.

4. Erasable ROM

In erasable ROM, information can be rewritten, unlike ROM and PROM. In this sense, erasable ROM is often called nonvolatile memory. Since writing new information usually requires high voltage and a long writing time, erasable ROM is mostly used for reading information, while information is occasionally rewritten. There are several different types of erasable ROM with slightly different working principles and structures, as described in the following. All of them are based on MOSFET with a floating gate but are made of different materials, as illustrated in Fig. 14a, where the insulation layer beneath the floating gate is thin. A MOSFET with floating gate is denoted with a MOSFET-like symbol with three bars, as shown in Fig. 14b. For writing information 0, a voltage that is higher than normal operating voltage is applied to the regular gate against the source and then electrons (shown with "-" inside circles) sneak into the floating gate, as illustrated in Fig. 14a. The floating gate is completely isolated from any other part of the MOSFET, and the electrons stay trapped in the floating gate, even after the voltage applied to the gate is turned off. For writing information 1, electrons are not stored in the floating gate. This writing takes a long time. (So, writing 0 in each floating gate one by one is often done in some type of erasable ROM after simultaneously erasing, i.e., emptying all floating gates.) For reading, a normal voltage is applied at the regular gate and also between the drain and the source. Then if the floating gate is full of trapped electrons, electrons do not flow through the channel of the MOSFET between the source and drain because the electric field caused by the voltage at the gate is not strong enough, being offset by the electric field caused by the electrons in the floating gate. If the floating gate has no electrons, electrons flow through the channel. Thus, information, 0 or 1, is read out, respectively.

The erasable programmable ROM (EPROM) is a ROM in which information stored in all memory cells is simultaneously erased by flood exposure of ultraviolet light, and then new information is electrically written into each cell. This flood exposure takes tens of seconds to tens of minutes.

In the case of electrically erasable programmable ROM (EEPROM), the previous information is not erased by the exposure of ultraviolet light but by applying appropriate voltages. This rewriting can be done in hundreds of microseconds to a few milliseconds per byte. This is much faster than EPROMs. Reading is somewhat slower than other ROMs (hundreds nanoseconds or less), but rewriting is much slower.

5. Flash Memory

Flash memory is specially-arranged EEPROMs, occupying a smaller area than EEPROM or DRAM arranged in other structures. Consequently, on a given chip area, flash memory with greater memory capacity than DRAM can be manufactured inexpensively. Also, hundreds to tens of kilobytes of data can be simultaneously erased, unlike other types of EEPROM. But unlike other types of EEPROM whose addresses can be selectively rewritten, all memory addresses in one block in flash memory are erased and then each memory address in the block can be written individually. This is inconvenient because even if we want to rewrite only one address, we have to rewrite the entire block.

A flash memory is currently structured in at least four different ways with different features. They are NOR type, NAND type, AND type, and DINOR type. The first two types have been widely used. A NOR-type flash memory is illustrated in Fig. 15. If one memory word consists of 8 bits, we have bit lines, $D_1, D_2, \ldots, D_8$. All these lines are connected to the drains of MOSFETs with floating gates. $A_1, A_2, \ldots, A_K$ are the outputs of a memory address decoder, which are connected to the gates of MOSFETs with floating gates and only one of which has a high voltage when the decoder is active. All sources of MOSFETs with floating gates are connected to S, which means a source. Applying appropriate voltages to S, D_i and A_j, writing, reading, or erasure is done. White circles denote contact windows between aluminum lines and drains. A

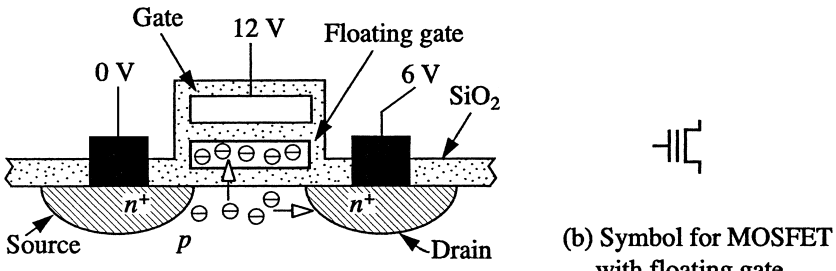


FIGURE 14 A MOSFET with floating gate.

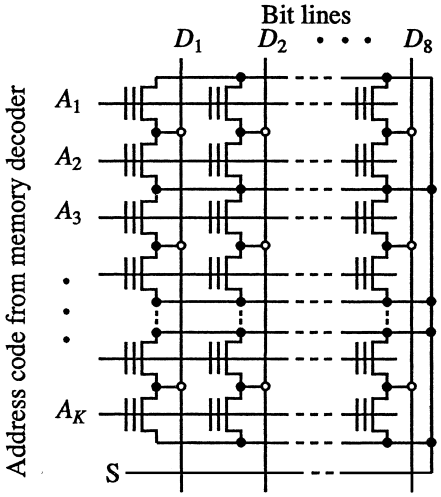


FIGURE 15 NOR-type flash memory.

NAND-type flash memory is illustrated in Fig. 16. Reading time in NOR-type flash memory is about hundreds of nanoseconds or less, being comparable to ROMs, but is faster than the NAND-type, though its area is larger (because of more contact windows which occupy large areas). Erasure and writing time in both types is a few microseconds, being far slower than RAMs, but is far faster than hard disks whose writing time is tens of milliseconds.

Flash memory has one serious shortcoming. If writing is repeated many times, flash memory gradually requires longer writing time and eventually stops working because of fatigue. Currently, we can repeat writing close to a million times and this is three orders of magnitude smaller than hard disks.

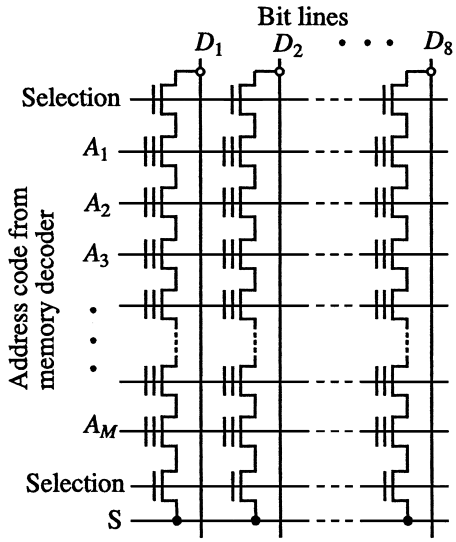


FIGURE 16 NAND-type flash memory.

Flash memory is cheaper with greater memory capacity than DRAM and yet faster, though more expensive per bit, than hard disks. Flash memory is immune to vibration, dust, and mechanical shocks. Also, flash memory works with far less power consumption, and accordingly if flash memory is installed in a notebook computer, replacing a hard disk, then batteries last far longer. A hard disk consumes a few amperes at the start and then consumes tens of milliamperes for rotation after that. In contrast, a flash memory in power-down mode consumes only a few microamperes. Flash memory has many unique applications. Flash memory is used in upgradable BIOS (Basic Input Output System), which is the hardware part of an operating system of some personal computers, and also in modems. Flash memory is widely used as credit-card-sized memory cards, which are called PCMCIA (personal computer memory card international association), for portable computers, as handy memory extensions. In digital cameras, photographic images are stored in flash memory. Flash memory is also used in cellular phones, where the identification of the owners and programs are stored and updated by wireless transmission. Flash memory is also used in pocketable recorders that record voice without microcassettes, i.e., without mechanical drives. Such pocketable recorders are far smaller than conventional pocketable recorders, and batteries last far longer. Flash memory is used also in FPGAs, as explained later.

C. Ferroelectric Memory

Ferroelectric memory has been recently used because of the following unique feature. It is nonvolatile and both reading and writing are fast, being almost comparable to those of DRAMs, whereas the other nonvolatile memories have a very long writing time. It is based on piezoelectric material instead of semiconductors. Each memory cell occupies a small area, but it fatigues for both repeated writing and reading. So it has been used in applications that need reading and writing only occasionally, such as smart cards.

D. Comparison of Semiconductor Memory

The performance and cost of semiconductor memory are compared in Table I, although precise comparison is not possible due to technological changes and product variations. ROMs are either mask- or field-programmable. In the case of mask-programmable ROMs, a semiconductor manufacturer charges a customer for each mask-programmable ROM and then a fixed fee for the entire set of masks to be custom-made which is much higher. But in the case of field-programmable ROMs, the price of

TABLE I Comparison of Semiconductor Memory in MOSFETs

	ROM			RAM	
	Mask-programmable	**EEPROM**	**Flash memory**	**Static**	**Dynamic**
Memory capacity per area	Highest	Medium	High	Low	Medium
Read time	100–200 nsec	200–500 nsec	90–150 nsec	10–70 nsec	60–100 nsec
Write time	—	1–10 msec	1–5 sec	10–70 nsec	60–100 nsec
Programming time	—	Very long	Very long	—	—
Cost for low volume	High	Low	Low	Medium	Lowest
Cost for high volume	Lowest				
Fatigue	None	For writing	For writing	None	None

each field-programmable ROM is higher than that of a mask-programmable ROM with the same memory capacity because of a larger chip size due to fuses or antifuses and electronic circuits for disconnecting them. But there are no mask charges.

IV. COMPARISON OF DESIGN OF LOGIC NETWORKS AND MEMORY

From the viewpoint of design, memory, and logic networks, which altogether constitute a computer, they each have almost completely different properties due to the difference in layout structure. If the layout of transistor circuits at the layout stage in Fig. 2c are to be the most compact, the designed chip will have the highest performance with the smallest size. For this, the compaction of layout is like a jigsaw puzzle because logic gates (or the corresponding transistor circuits) have different sizes and shapes along crooked interconnections. This is called random-structured layout and logic networks in such layout are called random-structured logic networks. Such layout is the most time-consuming. If the layout needs not be compact, interconnections can be straight lines and accordingly logic gates are sparsely laid out. This is called regular-structured layout and is far less time-consuming. In the case of random-structured logic networks for the highest performance or the lowest cost, designers have to lay out compactly on a chip a number of different complex logic networks with millions of logic gates and the corresponding transistor circuits that realize these logic networks. If designers want to change part of already laid-out logic networks, they often need to lay out from scratch the other part that does not need to change. Thus the compact layout of random-structured logic networks is the most time-consuming stage throughout the entire design sequence.

In the case of memory, a very simple transistor circuit of one transistor, or at most six transistors, must be designed as a memory cell and then this memory cell is repeated many times (millions of times), using straight lines in matrix form. So the design of the entire memory is short and simple, though in addition to one memory cell, peripheral electronic circuits for writing and reading must be designed. Whenever a new processing technology, such as new lithography technology, is developed, it is first applied to memory because of its short design time. Then after confirming the feasibility of the new technology with memory, the technology is applied to logic networks which require a much longer design time. Thus, most advanced processing technology is first demonstrated in memory. Because of the regular-structured layout, memory can pack a few more devices in the given area on a chip than logic networks in a random-structured layout (i.e., nonrepetitive layout using crooked lines), even when the same fabrication technology is used. Thus, in comparing the integration sizes of different chips, we have to know the proportions of memory and logic networks on each chip. The device densities of memory and logic networks cannot be directly compared because of the difference in layout structure, consequently in device packing density.

V. LIMITATIONS OF INTEGRATION SIZE

The cost of processing a wafer is fairly constant (hundreds of dollars per wafer), no matter how many chips the wafer has. Suppose the cost of processing a wafer is F, the number of chips in a wafer is N, and the yield is y (thus the number of faultless chips from the entire wafer is yN). The cost of a wafer before processing is usually much smaller than the processing cost F. Thus, the cost of a processed faultless chip is approximately F/yN. When the chip size S increases, keeping the same device density, the yield y decreases exponentially, as illustrated in Fig. 17. Also, the yield gets generally lower as the device density of the chip increases. The cost of a faultless chip can be rewritten as

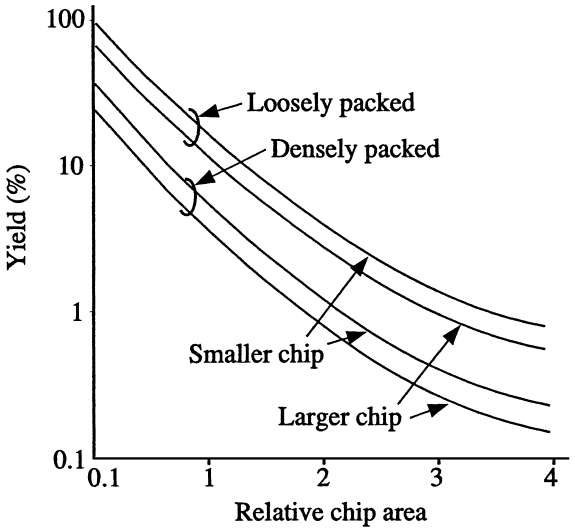


FIGURE 17 Yield of chip.

$F/yN = kS/y$ because the number of chips N in the wafer of a fixed size decreases for the increase of the chip size S, where k is a constant coefficient. The number of logic gates in a chip is linearly proportional to the chip size S. Thus, the cost per logic gate increases exponentially for the increase of chip size S, because the yield y decreases exponentially in Fig. 17.

As industry is improving the device density, the integration size is increasing, but the integration size is limited by either heat dissipation or yield. As the integration size increases, it is becoming difficult to dissipate the heat generated by transistor circuits. If the heat is not properly dissipated, the temperature inside the chip rises too high, damaging the chip. So far the maximum power consumption which is permitted to a chip is about 10 W by air cooling. Beyond this, a chip needs to be cooled by liquid. Logic families based on bipolar transistors, such as ECL, have large power consumption, and are limited in integration size by heat dissipation. The logic families based on MOSFETs have much lower power consumption than those based on bipolar transistors. The integration size is limited by the yield deterioration as shown in Fig. 17, if the MOSFET chips do not have high speed. Among them, CMOS has the least power consumption at low speed and is least likely to be limited by heat dissipation. For this reason, CMOS is the most important logic family for VLSI or ULSI design. But the power consumption of CMOS linearly increases as the switching frequency of inputs to CMOS logic gates increases (the speed of CMOS becomes faster) and the integration size is limited by heat.

In order to increase the integration size, transistors and other devices have been made smaller and also connections narrower. Line width is usually used as a measure of scale-down. With the technology for smaller line width, the channel length of a MOSFET (the distance between the diffused source and drain regions) gets shorter. The smaller the channel length, the faster the MOSFET works. So far, scale-down has continued by using photolithography with light source of shorter wavelength. This requires the exploration of new processing technologies.When the channel length approaches 0.1 μm, the improvement of photolithography gets difficult and complex physical problems are encountered, necessitating different types of technology progress.

VI. DIFFERENT DESIGN APPROACHES

The complexity, time, and expense of the design discussed so far increases very rapidly as the integration size of a chip increases. The design of a complex IC chip with the highest speed and the smallest size easily requires several dozen engineers working about four years, spending millions of dollars, if the most advanced technology is to be adopted (if not, that much manpower and expense are not required). If the design takes longer, the chip would be obsolete due to the progress of technology and change of market conditions. Roughly speaking, the expenses of VLSI or ULSI chip design are linearly proportional to the man-month, i.e., the product of the number of designers and the time required to complete a design. The major part of the expenses is the salaries of designers and indirect costs, such as those for computer terminals, workstations, and then if the new technology is to be tested, a pilot plant.

The most serious problem in the design is the elimination of mistakes. As the integration size increases, more mistakes enter into the design. This is further aggravated by an increasing number of designers involved because when a chip of a greater integration size must be designed within the time limit of about four years, more designers would be required. For this reason, the use of CAD programs has been intensified in order to shorten design time and minimize mistakes. The accelerating use of CAD programs, however, cannot drastically reduce the time and cost of the design, if the chips are to have the highest speed and the smallest size. But if we sacrifice the best result in chip size and speed, there are many ways to reduce the design time and cost. They are called semi-custom design (also called ASIC, i.e., application-specific integrated circuits) approaches, in contrast to the full-custom design approaches (the design approach discussed so far) in which the smallest chip size and the highest speed are aimed at deliberately at every design stage.

The most important factor in the competitiveness of an IC chip in the market is the total production cost

of the package with design cost considered, if the chip has the same performance as competitors' chips. All the full- and semi-custom design approaches center around this. Let the design cost be denoted by E, the number of IC packages (volume) to be produced throughout the market life of the package be denoted by V, and the manufacturing cost of a faultless IC package be denoted by M. Here M is the fabrication cost of a faultless chip, including the labor and material cost for assembling and packaging the chip into a package. Then, roughly speaking, the total production cost C of an IC package, with the design cost taken into account, is given by the following:

$$C = (E/V) + M$$

by dividing the design cost E by the production volume V and adding it to the manufacturing cost of each faultless IC package M. Using this formula which shows an approximate relationship among key factors E, V, and M for determining the total production cost C of an IC package, we can explain why we need many different semi-custom design approaches. The second term M in this formula is highly dependent on the chip size. If the integration size is very small, a large number of chips, say thousands of chips, can be made from a wafer with a high yield which is close to 100%. Thus the cost of a faultless chip would be in the order of pennies. Even with a plastic package, M would be a fraction of a dollar. In this case, the cost of the package is more expensive than the cost of the chip inside. As the integration size increases, fewer chips can be made from a wafer, and the yield is lower. In the case of ULSI chips, a few hundred chips can be made from a wafer, and the yield varies greatly, depending on many factors such as complexity of integrated circuits and processing technology. So the cost of a fabricated faultless chip and also M can be as high as hundreds of dollars. In this case, the labor and material cost for packaging is usually less than the cost of the chip inside the package.

The first term E/V in the preceding formula is highly dependent upon the production volume V. The relative size of the value of E/V to the second term M is important. When production volume V is very large, E/V does not add much to manufacturing cost M even if design cost E is fairly large. From the formula for the total production cost of an IC package, $C = (E/V) + M$, we can see the following: If production volume V is very large, the manufacturer can invest heavily in design without losing its competitive position and the full-custom design approach can be used for this IC package; if V is small, the manufacturer must reduce E and a semi-custom design approach must be used. Design cost E, roughly speaking, is linearly proportional to the man-month, as already mentioned. There are many ways to reduce the man-month required for design, as will be described for each semi-custom design approach in the following.

VII. FULL-CUSTOM DESIGN APPROACH

As discussed already, the full-custom design approach is justified only when large production volume is expected. So microprocessor chips which are extensively used in home appliances or personal computers have been designed by the full-custom design approach. At the system design stage in the sequence of design stages illustrated in Fig. 18 (also shown in Figs. 2a-c, "architecture design ⟶ logic design ⟶ electronic circuit design ⟶ layout"), an architecture of a computer to be realized on a chip is designed in detail, by breaking it into system components such as a processor, RAM, ROM, registers, etc. (If the computer is dedicated to special computational tasks, appropriate algorithms for the computational tasks must be devised before the architecture design.) At the logic design stage, logic networks for each system component are designed using logic gates. At the electronic circuit design stage, these logic networks are converted into electronic circuits, considering speed, voltages, currents, and other electronic parameters. At the layout stage, geometrical shapes and locations of these electronic circuits are determined on an chip by CAD or by hand by a number of draftpeople. The layout is the most time-consuming stage throughout the design sequence in the case of random-structured layout, though the layout is far less time-consuming in the case of regular-structured layout. The design sequence ends at the layout. Since each design stage is complex and time-consuming, design verification, simulation, and testing are done at every stage, using CAD programs. As the integration size increases, an increasing number of designers and draftpeople are involved, making the design more prone to mistakes. Thus, an increasingly longer time is spent on verification, simulation, and testing throughout the design sequence. If any mistakes are detected, the designers must go back to earlier stages to correct them and repeat the design sequence at least partly.

By using the full-custom design approach, the highest speed can be obtained by deliberate design at every stage of the design sequence. Minimizing the chip size by compacting as much as possible, we can obtain more chips from a wafer and consequently, the cost of a chip is minimized. At the same time, the speed is further improved because every connection is shorter than those on a larger chip.

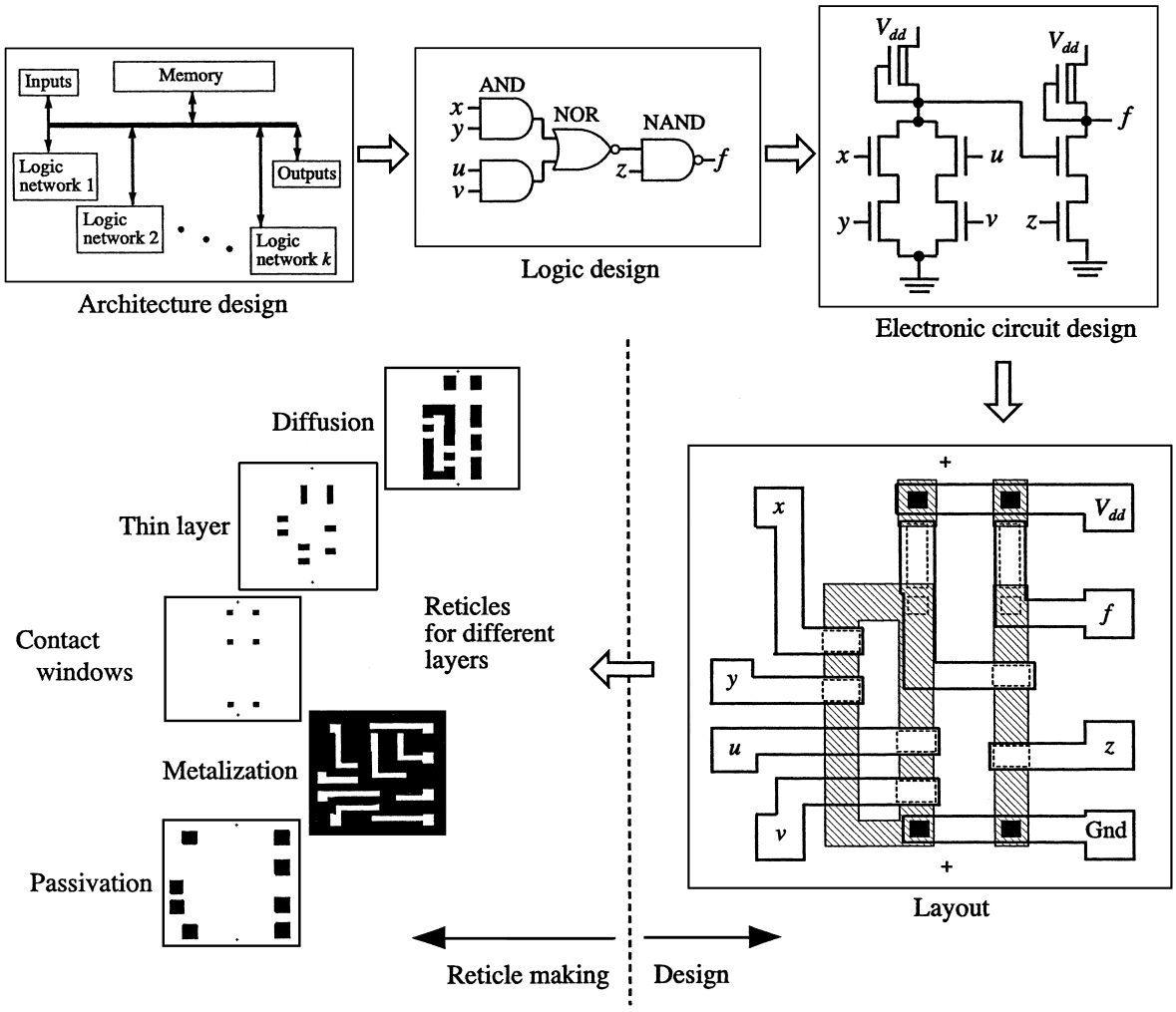


FIGURE 18 Design and reticles.

VIII. SEMI-CUSTOM DESIGN APPROACHES

There are many different semi-custom design approaches and designers of IC chips choose appropriate ones by considering trade-offs among design time, design cost, and the performance of the designed chips.

A. Programmable Logic Devices (PLDs)

Programmable logic devices are a type of semi-custom design approach where transistor circuits are directly realizable from given logic functions, skipping design and layout of transistor circuits.

1. Programmable Logic Arrays

Programmable logic arrays (PLAs) are a semi-custom design approach where transistor circuits can be realized immediately from a minimal sum for given logic functions, without designing logic networks and designing and laying out transistor circuits. A given set of logic functions generally has many different expressions, but a minimal sum, the simplest expression among them can be quickly derived by a CAD program. Straight lines are pre-laid out vertically and horizontally in matrix form on a chip along with MOSFETs at fixed points, as illustrated in Fig. 19. Then the horizontal and vertical lines are connected through MOSFETs (realized as illustrated in Fig. 12b) at some intersections, as specified by the derived minimal sum. Thus, by using PLAs, design of logic networks, electronic circuits, and layout can be completely skipped and running time of the CAD program is very short. So the design time is the shortest (in the order of minutes), though the chip size is the largest and the speed is the lowest among chips designed by other semi-custom approaches.

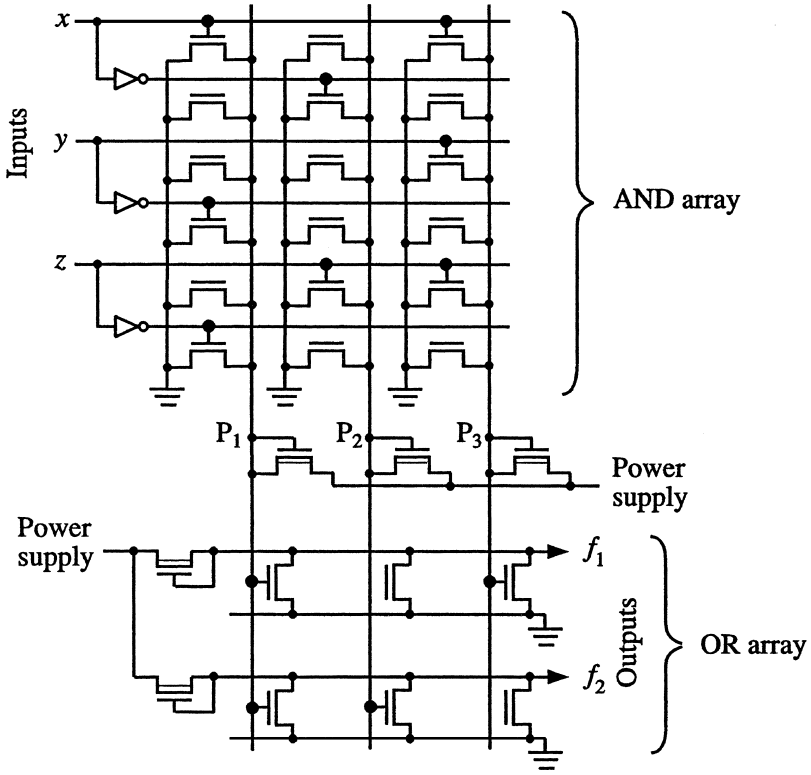


FIGURE 19 PLA without flip-flops.

In a PLA, the horizontal lines (metal line) and the vertical lines (diffused regions) are all straight lines, as shown in Figs. 12b and 19, unlike the layouts in other design approaches such as Fig. 18, where lines are not always straight.

Suppose logic functions

$$f_1 = x\bar{y}\bar{z} \vee xyz$$

$$f_2 = x\bar{y}\bar{z} \vee \bar{x}z$$

are to be realized in a PLA, where "$\vee$" denotes the OR operation. All distinct products, $x\bar{y}\bar{z}$, xyz, and $\bar{x}z$, are represented by the vertical lines, which are called product lines, in the upper part of the PLA in Fig. 19. Inputs x, y, z and then $\bar{x}, \bar{y}, \bar{z}$ through inverters have horizontal lines in the upper part of the PLA. The product $x\bar{y}\bar{z}$, for example, is represented by the product line (the first vertical line from the left in Fig. 19), which has dots at the intersections with the horizontal lines for x, $\bar{y}$, and $\bar{z}$. Although a certain product such as $x\bar{y}\bar{z}$ appears more than once in f_1 and f_2, we need to realize only one product line for each of them. The upper part of the PLA is called the AND array, because products such as $x\bar{y}\bar{z}$, the AND operation of x, $\bar{y}$, and $\bar{z}$, are realized. The lower part of the PLA is called the OR array, realizing the OR operation "$\vee$" for each function. The horizontal line for each function in the OR array has dots at the intersections with the product lines which represent the products contained in that function. For example, the horizontal line for function f_1 has a dot (a MOSFET is realized) at the intersection with the product line for $x\bar{y}\bar{z}$ because product $x\bar{y}\bar{z}$ is in the expression for function f_1. A minimal sum expression for the functions minimizes the size of a PLA because the minimal sum has the smallest number of distinct products (thus, the number of product lines) among all possible expressions for the given functions (note that the number of horizontal lines is fixed).

Programmable logic arrays are usually realized in dynamic-CMOS PLAs to reduce power consumption.

In addition to the short design time, PLAs have the advantages that only one mask must be custom-made and the dot pattern is unchanged even with new processing technology. PLAs are used frequently in places, such as control units in microprocessors, where design changes are expected in a chip which is designed by the full-custom or other semi-custom design approaches.

In the case of the PLAs thus far described, information has to be stored with masks by semiconductor

manufacturers, so, these PLAs are called mask-programmable PLAs. In contrast to this, field-programmable PLAs are PLAs where connections are cut off by disconnecting fuses or antifuses at desired places. These PLAs are larger than mask-programmable PLAs because of extra space for fuses (or antifuse) and electronic circuits for disconnecting.

PAL (programmable array logic) is a PLA
that its OR array is not programmabl

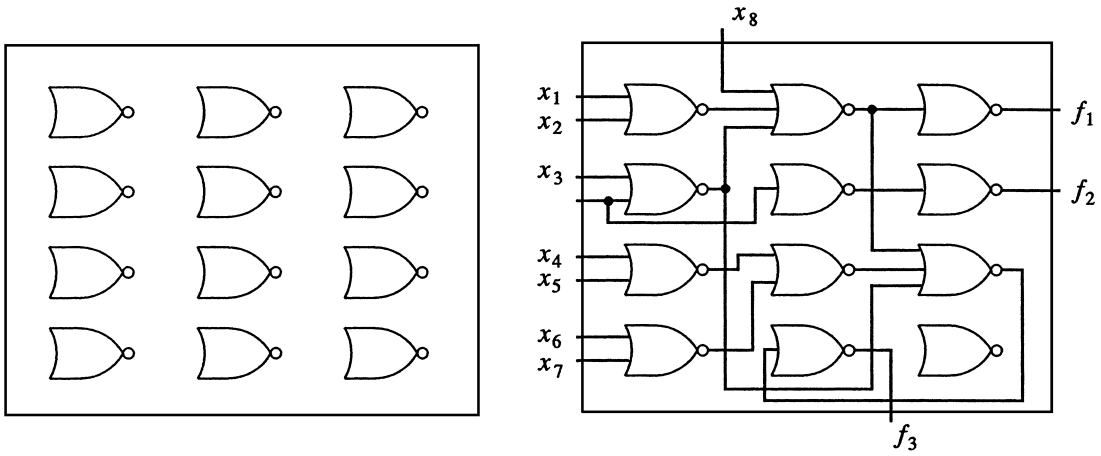
B. Gate Arrays

A gate array is logic gates that are pre-laid in matrix form on a chip, as illustrated in Fig. 20. Actually, unconnected components of logic gates, which are called cells (these cells should be differentiated from the cells in the cell library to be described later), are pre-laid instead of logic gates. By connecting the components in a cell appropriately, we can realize one of a few different types of logic gates. A cell of CMOS gate array, as an example, is shown in Fig. 21, where a pair of pMOSFETs and a pair of nMOSFETs are placed in the left and right, respectively, without connections among them. The NAND gate shown in Fig. 22a can be realized by connecting the components shown in Fig. 21 by two metal layers as shown in Fig. 22b. These two metal layers are formed by forming the first metal layer shown in Fig. 22c, the insulation layer (not shown), and then the second metal layer shown in (d).

After the designers finish the logic design, the logic networks are realized by appropriately connecting the components in a cell to form the logic gates and then appropriately connecting the logic gates. Thus, the layout of the entire chip has been obtained. In this case, the electronic circuit design and the layout, which are required in the full-custom design approach, are essentially skipped.

Gate arrays have the following advantages. Speed is improved over PLAs. Only 2 or 3 masks for connections and contact windows have to be custom-made, instead of all of two dozen masks required for full-custom design. Also only the connection layout, along with the allocation of gates, needs to be done effectively by CAD, greatly reducing the layout time. Thus the design with gate arrays is very inexpensive and quick, compared with the full-custom design. Gate arrays of CMOS have been extensively used. When new processing technology is to be adopted in gate array, we need to update the layout of unconnected components of only one cell, enabling us adopt the new technology quickly and inexpensively.

On the other hand, gate arrays have the following disadvantages, such as large chip size, many unused cells (limited by the available number of pads) and nonuniform delays on gates. The chip is larger than those designed by the cell-library design approach to be described next because each logic gate is not necessarily the most compact and all logic gates are not compactly laid out. Because ready-made components of the same size are used in all cells, each electronic circuit in a cell is not tuned to the need of each specific logic gate. For example, even if the gate has a long output connection, the size of layout for a transistor cannot be increased to offset delay increase. Thus the speed is sacrificed and logic gates at different sites have different delays. Also a CAD program may not provide good routing, despite the importance of routing by CAD for reducing design time and cost. When interconnections among gates are congested in some areas,



(a) Before making connections (b) After connections made

FIGURE 20 Gate array.

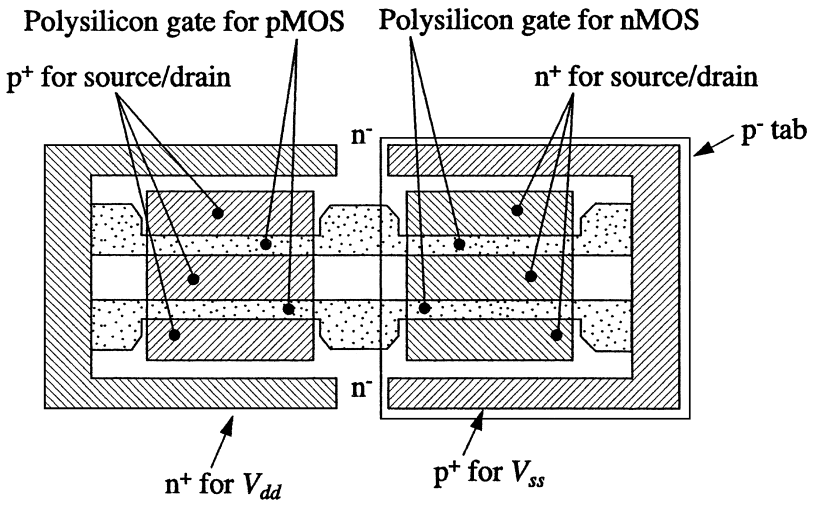


FIGURE 21 A cell of a CMOS gate array (Courtesy of Fujitsu Ltd.).

new interconnections have to be detoured by using long interconnections and thus increasing the delay.

In contrast to field-programmable gate arrays to be described later, the gate arrays here are called mask-programmable gate arrays.

C. Cell Library Approaches

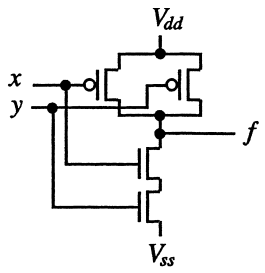
In the case of the cell library approach, the designs and layouts of electronic circuits for logic gates and simple subnetworks (such as flip-flops), which are to be used frequently, are prepared carefully in compact form by hand, in other words, by the full-custom design approach. These are called cells, or standard cells, and a library of cells is stored in a memory. When designers finish logic design for a chip to be realized, the cells for logic gates and simple subnetworks are retrieved from the cell library. Assembling these cells on a CRT (cathode ray tube) screen of a computer, the layout of the entire chip is completed.

The following two types of cell library approaches are popular. One is the standard cell library approach, where the layout of cells are prepared in rectangular shapes with the same height (their widths may be different), and when cells are assembled, all cells are arranged in rows with the same height and interconnections among cells are placed between the rows, as illustrated in Fig. 23. The area for interconnections between adjacent rows is called a routing channel. The second cell library approach is the hierarchical cell library approach, where each cell is prepared in a rectangle of different size appropriate for each specific logic gate or subnetwork (width and height can be different), and cells are arranged freely, not necessarily in rows or columns. Interconnections among cells are placed anywhere. Cells are first arranged, as closely as possible, into a block of rectangular shapes. Then several blocks are arranged into a larger block. We continue to build up blocks of smaller sizes into a larger one.

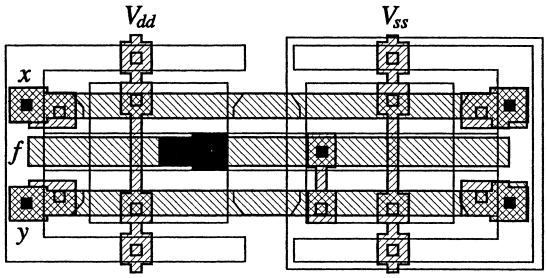
The advantage of the cell library design approach is the reduction of layout time. Once the cell library is prepared by hand to produce the most compact cells, layout by hand need not be repeated for any chips to be designed later and cells can be assembled by one designer on a CRT screen with a CAD program in short time, instead of dozens of draftpeople spending long time for the full-custom design approach. Thus the design cost is greatly reduced. The disadvantage is the increase of chip size and the decrease of the speed. The chip size increases because all cells are in rectangular shape, even if there are unused areas inside, and interconnections are routed in routing channels only (though some interconnections are allowed to cross rows of cells), widening routing channels whenever necessary, even if unused areas are created in the routing channels. Also the chip size increases, because the compaction of the entire chip is difficult, unlike the full-custom design approach where global compaction of the entire chip is always attempted. The speed decreases because interconnections tend to be longer due to the increase in chip size, and the electronic circuits and layout stored in the library are not tailored to each specific case of chip design (in the case of full-custom design, a transistor of larger size is used to prevent speed decrease whenever necessary).

The hierarchical library approach can produce faster and smaller chips than the standard cell library design approach. The performance and size of the designed chip is closer to the full-custom design approach, but this

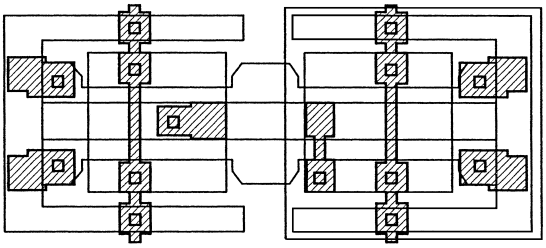
(a) NAND gate.



(b) The connections consist of two metal layers shown in (c) and (d).



(c) First metal connection layer.



(d) Second metal connection layer.

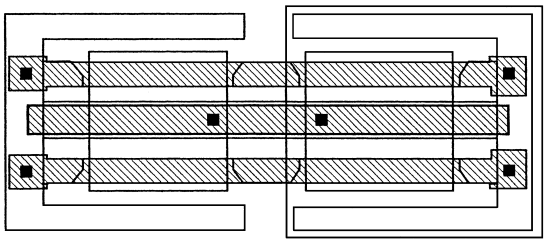


FIGURE 22 Connection layout example of the cell in Fig. 21 (Courtesy of Fujitsu Ltd.)

approach is more time-consuming and development of good CAD programs is more difficult.

IX. COMPARISON OF CUSTOM DESIGN APPROACHES

When the design cost E is reduced by a semi-custom design approach, the chip size S is generally sacrificed as illustrated in Fig. 24, where design time and speed are shown instead of E because both increase as E increases. Notice that both chip area and design time/speed are shown in logarithmic scale. Based on the formula, $C = (E/V) + M$, discussed in Section VI for the total production cost C per package with design cost taken into account, each of the full- and semi-custom design approaches is more cost effective than the others, only when the production volume V is within a certain range, as illustrated in Fig. 25. For example, the full-custom design approach can produce the least expensive package, when production volume V is very high. But when the production volume is lower, the total cost per package is higher than those made by any other design approaches.

On the same IC chip, more than one custom-design approach may be mixed. This is because different logic networks have different design problems, such as the need

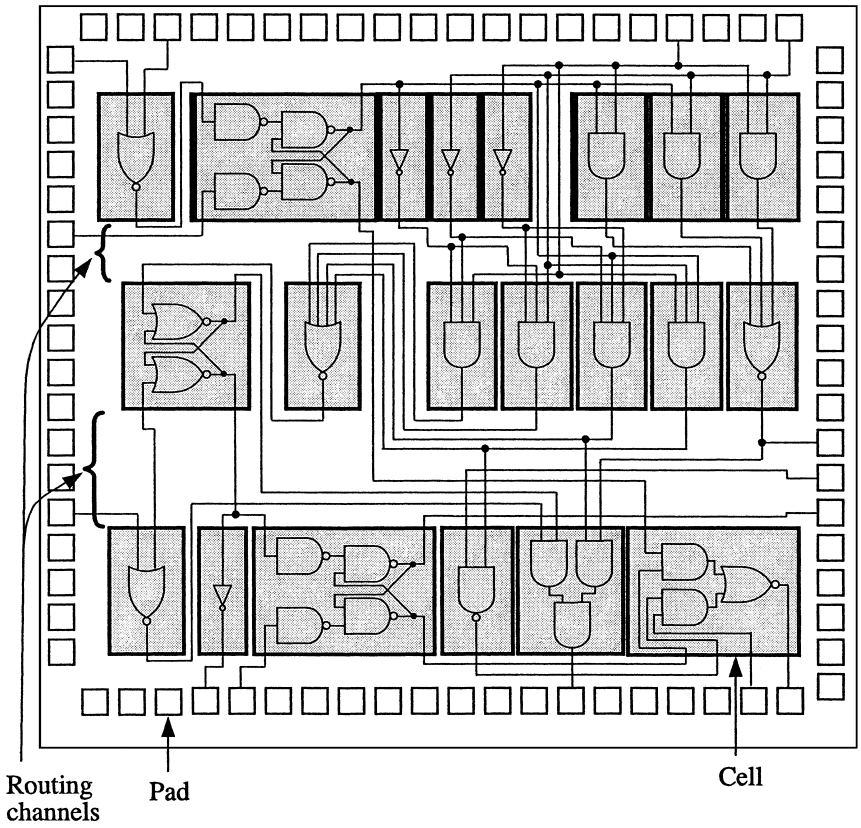


FIGURE 23 Standard cell layout.

for high speed or frequent design changes, so one semi-custom design approach is more suitable for a logic network than other approaches. For example, PLAs are suitable for the control logic unit of a computer but not for adders which require high speed but do not need frequent design changes.

X. RECONFIGURABLE COMPUTERS WITH FIELD-PROGRAMMABLE GATE ARRAYS

Field-programmable gate arrays (FPGAs) are pre-laid transistor circuits with memory that are programmable

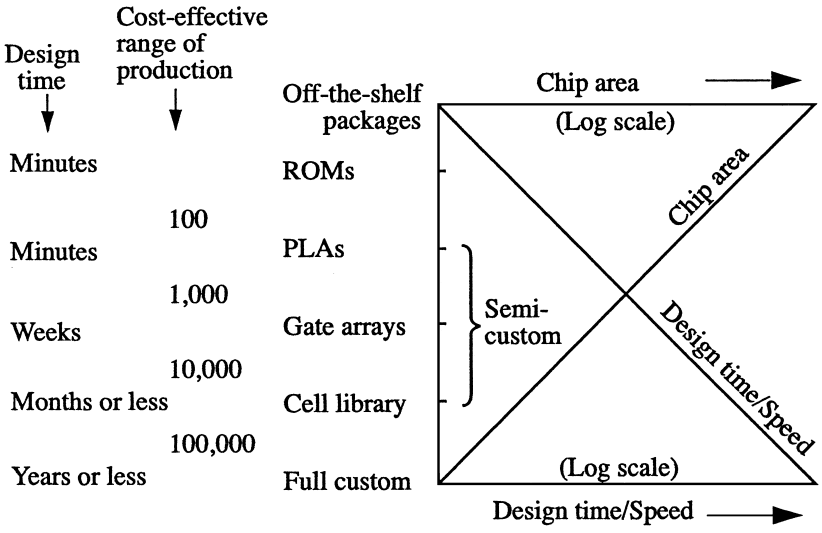


FIGURE 24 Chip area versus design time.

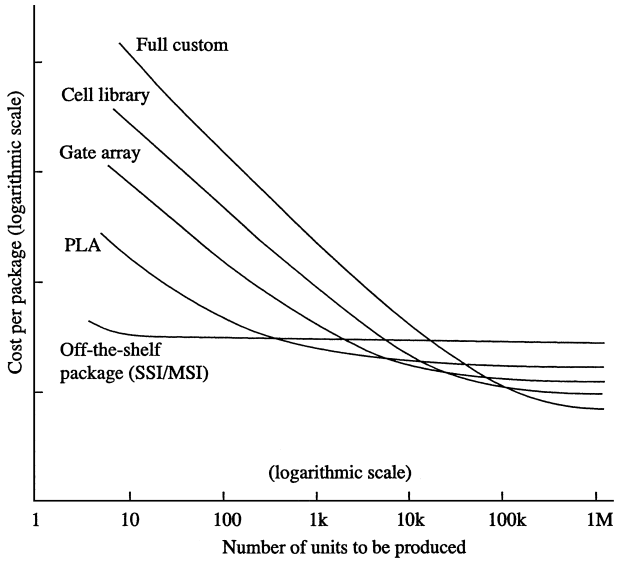


FIGURE 25 Package cost versus production volume.

by users. In other words, with FPGAs, users can realize their own logic networks in hardware just by programming without doing the time-consuming and expensive sequence of design and fabrication illustrated in Fig. 2. As FPGA stands for field-programmable gate array, it sounds like a field-programmable version of gate array, but it is very different from gate arrays, as it is not a semi-custom design approach for IC chip design. It is a completely new means for computation, being different from conventional computation by software run on a general-purpose computer, or by pure hardware specialized for some specific computational tasks. Writing software that lets a set of FPGAs do a desired computational task can be time-consuming and tedious. But the time is still much shorter than the complete design of the entire hardware for the computational task. Once it is completed, designers can realize their design on FPGA chips by themselves in minutes and then the entire set of FPGAs executes the desired computational task. Thus, FPGAs are used for such repeated execution of application programs that have the same type of data. FPGAs are also used as part of a conventional general-purpose computer such that its architecture can be changed, at least partly, whenever necessary. Such a computer is called a reconfigurable computer.

A. Basic Structures of FPGAs

Several different types of structures for FPGAs are available commercially. But all of them have the basic structure that many of the same logic blocks are pre-laid in matrix form, accompanied with a large number of pre-laid line segments for connecting these logic blocks. Thus, FPGAs are sometimes called logic block arrays (LBAs). One structure is similar to a gate array where each logic cell in a gate array is replaced with a logic block, as shown in Fig. 26. The internal structure of logic blocks is far more complex than the cells of a gate array and differs, depending on manufacturers. Unlike the cells of a gate array, each of which consists of a few unconnected components, a logic block consists of SRAMs (used as

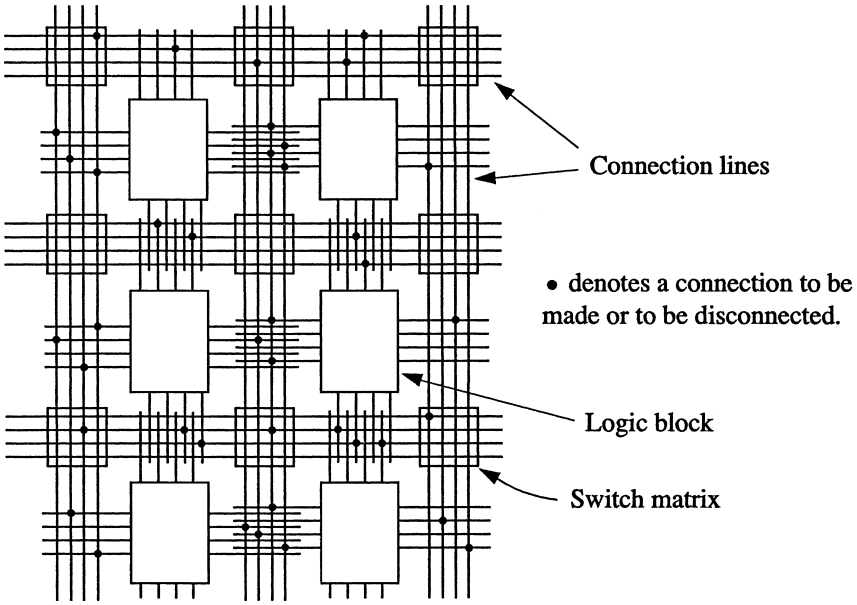


FIGURE 26 FPGA type of gate array with routing channels.

look-up tables that store simple logic functions), PLAs, NAND gates, multiplexers, flip-flops, and others. Line segments are pre-laid horizontally and vertically and are connected to the inputs and outputs of logic blocks by programmable switches. Vertical and horizontal line segments are connected by switch matrices, each of which is shown with a square and is a set of multiplexers controlled by SRAMs or other types of memory. Only necessary line segments are connected in order to make long connections without adding unnecessary delay due to parasitic capacitance by using excessively long line segments.

Each programmable switch or switch matrix consists of many switching elements. It is essential for performance, costs, and sizes of FPGA how switching elements are realized. Switching elements are fuses, antifuses, RAMs or nonvolatile memories (different types of programmable switches are available from different manufacturers), and have all different programming times (writing time). Thus, depending on which of these are used, different FPGAs have significantly different applications. Field-programmable gate arrays based on SRAMs, for example, connect lines by controlling a pass transistor (Fig. 27a), a transmission gate (Fig. 27b), or a multiplexer, each of which is controlled by a flip-flop, i.e., one memory cell of SRAM (Fig. 27c). Any horizontal line and a vertical line shown in Fig. 26 can be connected inside a switch matrix (shown with dots in Fig. 26). If a flip-flop is used, these two lines are connected when the flip-flop is on, and unconnected when it is off.

FPGAs have the following advantages: once designers finish writing software for setting up the set of FPGAs for the execution of a desired computational task such that logic functions be realized by these FPGAs, then the logic functions are realized in these FPGAs in minutes by writing design information into SRAMs, as necessary. When RAM is used, the disadvantage is the following: If the power supply to this FPGA is turned off, the information is lost and consequently, design information must be written each time the power supply is turned on.

B. Features of FPGAs

With FPGAs, design of a new computer can be verified before finalization, and new computers can be introduced to the market quickly. Logic networks in FPGAs perform far faster than verification of designed logic networks by software. The development time of a new computer can be often shortened to less than half. In other words, when logic design is finished, computers can be shipped immediately in FPGAs to major customers for testing. After the customers find bugs, the designers fix them. With FPGAs, the designers can send debugged design to the customers online, and customers can start second tests instantly by revising the contents of the FPGAs. After the repetition of finding and fixing bugs, the production of the finalized computer can be started. In contrast to this, without FPGAs, the computers have to be shipped to customers for testing after manufacturing mask-programmable gate arrays, spending several weeks. Then when the customers find bugs, the designers start the production of the second mask-programmable gate arrays. Thus, a new mask-programmable gate array has to be manufactured, spending a few weeks, for each testing and correction of bugs. The completion of the computer will be significantly delayed with far greater expenses.

An FPGA can pack only one-tenth of the number of logic gates in a mask-programmable gate array because devices for user-programmability, such as SRAMs, nonvolatile memory, and antifuses take up large areas. Mask-programmable gate arrays are faster by two orders of magnitude and far cheaper for large production volume. But the speed of FPGA is still much faster than software run on a general-purpose computer for many application computational tasks, if not all. Field-programmable versions of other semi-custom design approaches have very similar structures to their mask-programmable versions. But FPGAs have completely different structures from

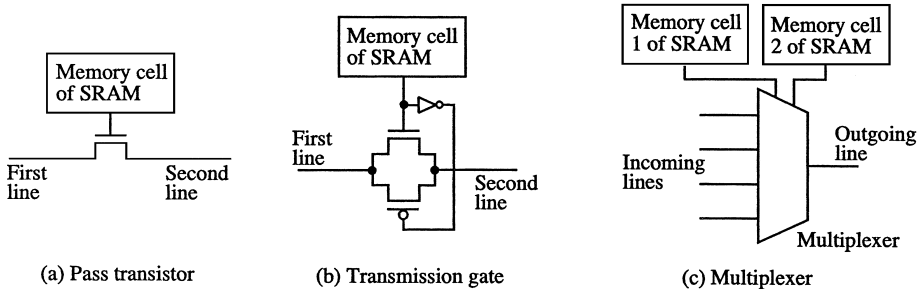


FIGURE 27 Connection of lines by SRAM.

mask-programmable gate arrays, i.e., logic blocks versus cells. Thus, FPGAs are usually used as a new types of computers that can be easily customized for specific computational tasks rather than prototypes for mask-programmable gate arrays.

XI. DESIGN TOOLS AND FUTURE PROBLEMS

Progress in chip design has been essentially a fight with the increasing complexity. To reduce the increase of design time due to increasing complexity, semi-custom design approaches are increasingly often mixed with the full-custom design approach on a VLSI or ULSI chip. Microprocessor chip Z8000 designed in 1979 was probably the last chip designed using the full-custom design approach. This chip has 17,500 transistors and has no PLAs or ROMs. Beyond this number of transistors, use of only the full-custom design approach would be too time-consuming. When more transistors are to be placed on a chip, logic networks designed with cell library approaches, and PLAs and other semi-custom design approaches are mixed on a chip. Computer-aided design tools (CAD programs) are vitally important weapons against long design time and increasing mistakes. By using CAD tools, design time is reduced and chances of mistakes are greatly reduced. The CAD tools, however, cannot always produce results of the same quality as those by the full-custom design approach, depending on application areas such as layout. Yet each CAD program for different purposes is becoming very complex. Interaction among them is becoming essential for further improvement of performance and size reduction of chips. So we need to merge these CAD programs, though this requires formidable efforts.

Improvement of performance and reduction of chip size by scale-down are going to the end, as photolithography is approaching its limit. Also we are going to face complex physical problems, and as channel length is becoming very short, we will probably need different types of new technology.

SEE ALSO THE FOLLOWING ARTICLES

ANALOG-SIGNAL ELECTRONIC CIRCUITS • CIRCUIT THEORY • COMPUTER ARCHITECTURE • EPITAXIAL TECHNOLOGY FOR INTEGRATED CIRCUIT MANUFACTURING • FIELD-EFFECT TRANSISTORS • INTEGRATED CIRCUIT MANUFACTURE • MICROCOMPUTER DESIGN • SIGNAL PROCESSING, ANALOG

BIBLIOGRAPHY

Brown, S. D., *et al.* (1992). "Field-Programmable Gate Arrays," Kluwer Academic, Dordrecht.

Chen, W. K., ed. (1999). "VLSI Handbook," CRC Press, Boca Raton, FL.

Muroga, S. (1982). "VLSI System Design," Wiley-Interscience, New York.

Shoji, M. (1988). "CMOS Digital Circuit Technology," Prentice Hall, New York.

Smith, M. J. S. (1997). "Application-Specific Integrated Circuits," Addison-Wesley, Reading, MA.

Uyemura, J. P. (1999). "CMOS Logic Circuit Design," Kluwer Academic, Dordrecht.

Veendrick, H. (1998). "Deep-Submicron CMOS ICs," Kluwer Academic, Dordrecht.

Weste, N. H. E., and Eshraghian, K. (1993). "Principles of CMOS VLSI Design," 2nd ed., Addison-Wesley, Reading, MA.



Vacuum Switches

Mietek T. Glinkowski
ABB Power T&D Company Inc.

I. Vacuum Arcs
II. Vacuum Interrupters
III. Vacuum-Interrupter-Based Devices
IV. Applications in Power Systems
V. Research and Development Related to Vacuum Switches
VI. Safety

GLOSSARY

Constricted arc Arc where both cathode and anode electrodes contribute to the arc plasma. Arc plasma is confined largely to the cylindrical column between the contacts.

Diffuse arc Arc where only a cathode electrode is active and contains arc spots (cathode spots). Anode electrode is a passive collector of arc plasma.

Vacuum arc An electrical discharge between two electrical contacts in vacuum environment. The more correct term is metal vapor arc.

Vacuum circuit breaker Typically a three-phase device containing three poles with three vacuum interrupters with a mechanical drive and control circuitry to close or open the contacts.

Vacuum interrupter Commonly called "bottle;" set of electrical contacts sealed in a ceramic enclosure and evacuated to vacuum pressure.

Vacuum switches A family of devices based on the vacuum interrupters.

ON FEBRUARY 20, 1890, Oscar Enholm of New York filed a patent No. 441,542 called Device for Transforming and Controlling Electric Currents. This is the oldest known reference to a vacuum switch. It was a combination of a vacuum switch and a transformer, resembling in principle a direct current (DC) driven relay allowing transforming a DC current into alternating current (AC) to send electrical energy to distant electric lighting loads. The part of the vacuum switch consisted of two fixed terminals arranged at an angle to each other and a wedge-shaped bridge piece that could short the two terminals together. The whole arrangement was under a vacuum glass case.

For many years later the benefits of switching electrical currents in a vacuum environment remained largely theoretical. In the mid-1920s Royal W. Sorensen, a professor at California Institute of Technology, Pasadena CA, and his student Hallan E. Mendenhall, designed and tested the first vacuum switches that showed some promise of practical, commercial application. Similarly to the 1890 Patent, Sorensen's switch had two fixed terminals embedded in a glass envelope that was continuously pumped to provide

sufficient degree of vacuum pressure. The movable, third piece was a bridge that could travel up and down to provide electrical connection. The movable bridge piece was connected to a magnetic steel armature that could be actuated by a solenoid. Sorensen's device was tested in a variety of power system conditions successfully passing, for example, tests at an impressive 926 amperes at 41,500 volts.

Although the early experiments with vacuum switches clearly showed superior performance over the conventional air or oil switches, the technology of producing vacuum enclosures, understanding the detailed physical phenomena related to vacuum discharges, and the ability to make commercial products were not sufficient.

In the 1950s General Electric decided to make the next step. After large investment in technology R&D, a focused team of GE experts announced in early 1960 the first commercial three-phase vacuum circuit breaker. It took over 30 years after Sorensen's experiments and almost 70 years after the early U.S. patent to reach the stage at which vacuum switches could become commercially competitive with other switching technologies. This was the beginning of the new industry that is still growing at the turn of the millennium.

I. VACUUM ARCS

Vacuum arcs have been recognized as having unique electrical, physical, and chemical properties much unlike the arcs burning in oil, air, or other gases. The vacuum arc is actually a misnomer. Since there is an arc, there is no longer vacuum. The vacuum arc is referred to an arc burning in the atmosphere of ionized metal vapor originating from the metal(s) of the contact material. When electrical contacts are drawn in a vacuum environment, the arc is ignited at the cathode, the electrode with negative polarity. A so-called "cathode spot" is formed (Fig. 1), a very hot (well above the boiling point of metal) and high intensity (several atmosphere pressure) zone just above the cathode surface. This cathode spot is highly mobile and can produce vapor and the corresponding ionized particles to support the flow of current through the arc plasma to the other electrode (anode). Although the cathode spot has a very small dimension, of the order of 10 μm in diameter, it can support the flow of current of about 100 A for typical copper-based contact materials.

The cathode spot generates plasma, which is quasi-neutral, that is, contains an almost equal number of positively charged metal ions and negative electrons. Both ion and electron fluxes move in the same direction from the cathode spot toward the anode. To support a nonzero net flow of current, the velocities of the lighter electrons have to be greater than the velocity of heavy, metal ions. In a typical arc on copper electrodes for 100% of the net current, the ion component is about 10% of the total, whereas the electron current is about 110%.

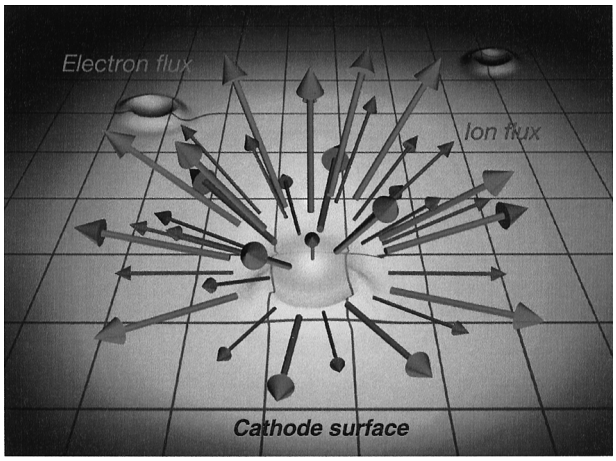


FIGURE 1 Cathode spot of a vacuum arc.

Because the vacuum arc can burn in a minute quantity of metal vapor plasma, and because the differential pressure between the cathode spot and the surrounding ambient (vacuum) is so high, the vacuum arc exhibits unique characteristics attractive for electrical power switching.

During the contact separation, when the arc is drawn, only small energy is needed to maintain the discharge. This makes the vacuum arc very efficient and results in a low erosion rate for the contacts, so small contacts with low mass can perform large number of switching operations. When the external circuit drives the alternating current to zero crossing the arc and its plasma quickly disappear from the electrode gap due to the diffusion to the surrounding vacuum. This process takes only a few microseconds and it is spontaneous: no external puffers, pumps, or other mechanical action is needed. After the plasma is diffused, the electrical current is fully interrupted and the circuit opening operation is completed.

It is remarkable that the contacts in vacuum can interrupt the arc at high rates of change. Because of the rapid diffusion of the plasma, the vacuum contacts can interrupt the current approaching zero line with the high rate of up to 1000 A/μsec. This is more than an order of magnitude higher than the other switching media, including SF6 (sulfur hexafluoride), and air.

A. Diffuse and Constricted Vacuum Arc Modes

Vacuum arcs can exist in two primary modes: diffuse and constricted (columnar arc). The difference between the two modes is significant. Figure 2 illustrates the two kinds. The diffuse arc consists of the cathode spot (mentioned above) or a multiple, depending on the current level. The

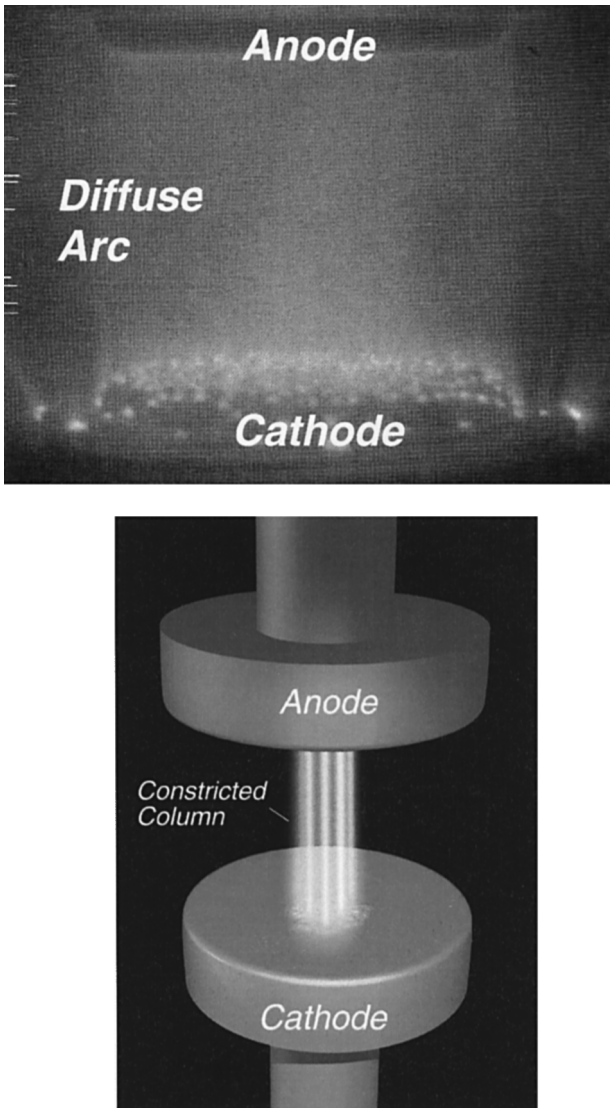


FIGURE 2 Diffuse and constricted vacuum arcs.

anode electrode is a passive contact, which only collects the vapors and the plasma but does not generate it. The arc is generated only at the cathode. When the current is above the ~100 A level (for copper contact arcs) multiple cathode spots will appear. A 2000 A arc, for example, will consist, on average, of approximately 20 parallel diffuse arcs. Each cathode spot "operates" independently. To maintain the arc diffused at the high current levels, an axial magnetic field (AMF) is required (see next section). In such a magnetic field the spots have the tendency to repel each other and spread rather uniformly on the electrode surface. The spots are also highly mobile.

When the higher current arcs are not subjected to the axial magnetic fields the cathode spots will tend to merge, blend, and eventually will form one, larger arc column. This columnar arc (shown in Figure 2) has the large cluster cathode spot that evaporates much more electrode material than the individual, separate diffuse spots. At the same time, the anode electrode also acts very much like the cathode, generating its own large anode spot, metal vapor, plasma, etc. These two spots anchor to the electrodes, much like a stalled hurricane, can cause high, localized contact erosion, thermal overheating damage, and prolong the arc burning even after current zero. To prevent this from happening, a radial magnetic field (RMF) can be applied to the arc region.

The different modes of the vacuum arc and the conditions under which they exist are summarized in Table I. To summarize, in order to control the vacuum arc at high current levels, minimize its damaging effects to the electrodes, and ensure the successful interruption at the current zero, either AMF or RMF fields have to be utilized.

B. Vacuum Arc in the Axial and Radial Magnetic Fields

The effect of the AMF on the vacuum arc is to keep it in the diffuse mode. Although the detailed physical phenomena related to the interaction between the arc and the AMF are still being investigated, it is known that at high currents AMF causes the arc cathode spots to move away from each other. When the field is applied, the arc plasmas from the individual small cathode spots interact with each other. At the same time, the hemispherical pattern of plasma diffusion from the individual spots is subjected to the magnetic cyclotronic effect. The heavy, positively charged ions spin in one direction, the electrons in the other direction. The radii of the ion spins are large (typically much larger than the dimensions of the contacts), and electrons spin around much smaller circles (see Figure 3).

RMF causes the high current, columnar arcs to move (rotate) on the electrodes. The Lorentz force develops:

$$\vec{F}_{Lorentz} \propto \vec{I}_{ARC} \times \vec{B}_{RADIAL}$$

where I_{ARC} is the current flowing through the arc column and B_{RADIAL} is the radial magnetic field (perpendicular to the column). As a result the arc is moved sideways, in the direction perpendicular to the column and perpendicular to the applied radial field. Many vacuum interrupters use clever contact geometry structures to generate radial fields and move the arc around the perimeter of the contacts. In an effective contact design the RMF can spin the arc around the electrode up to 5–7 times per half-cycle of the 60 Hz current, i.e., 8.3 msec. This spinning cools the columnar arc and prevents it from anchoring in one place and causing extensive damage. When the circuit current falls, near the end of the alternating half-cycle, the columnar arc becomes diffuse again so the final stages of the interruption around the current zero are similar to the diffuse arc.

TABLE I Different Modes of Vacuum Arcs

Arc current	Arc mode	Condition for the mode	Arc behavior
Low (less than ~1–5 kA)	Diffuse	No special conditions (free burning)	One to several small diffuse cathode spots, highly mobile
High (more than ~1–5 kA)	Diffuse	Axial magnetic field applied	Large number of uniformly spread, small, diffuse arcs with individual cathode spots
High (more than ~1–5 kA)	Constricted	No special conditions (free burning)	Severe erosion from both cathode and anode spots, damage to the electrodes, arc anchored to one location on the contact surface
High (more than ~1–3 kA)	Constricted	Radial magnetic field applied	Columnar arc moves sideways or rotates, keeping the electrodes cool and minimizing the erosion.

II. VACUUM INTERRUPTERS

The heart of any modern vacuum switching device is a vacuum interrupter (VI). It is a self-contained component, hermetically sealed, housing a contact system capable of electrical switching. Figure 4 shows a schematic example of a vacuum interrupter.

A. Design

The entire VI design is deceivingly simple, although a lot of advanced research and development had to be performed to achieve this level of simplicity. Two contact systems are enclosed in an evacuated ceramic or glass "bottle," one contact being movable with the aid of a metal bellows. Commercial vacuum interrupters have an internal residual gas pressure inside of 10^{-5}–10^{-7} Torr. When the contacts are closed, the current flows through one electrode stem to the contact system and through the other contact system, and to the other contact stem. The contact system is symmetrical. When the electrodes are open under current, the vacuum arc is drawn between the contacts and the current is interrupted at the next upcoming current zero. The plasma generated during arcing is deposited either at the electrodes or at the ion shield that surrounds the contact system. The contact system includes the actual contacts made out of special alloys or sintered materials (see Table II for different applications). For higher current applications, special electrode geometry, utilizing either AMF or RMF principle, is incorporated as well.

In order to generate AMF or RMF for high current applications, different arc control geometries are utilized. A few typical generic geometries found in commercial VI designs are shown in Fig. 5. Many other varieties are found based on several considerations: achieve the lowest cost possible, simplify the design and manufacturing, maximize the mechanical strength of the contacts, maximize the magnetic field generated by the system, and achieve a specific desired distribution of the magnetic field to maximize the arc interruption process and reduce contact erosion.

Besides the special contact material mentioned above, the contact stems and arc control geometries are made of the oxygen-free high-conductivity (OFHC) copper, bellows from stainless steels, ion and bellows shields are

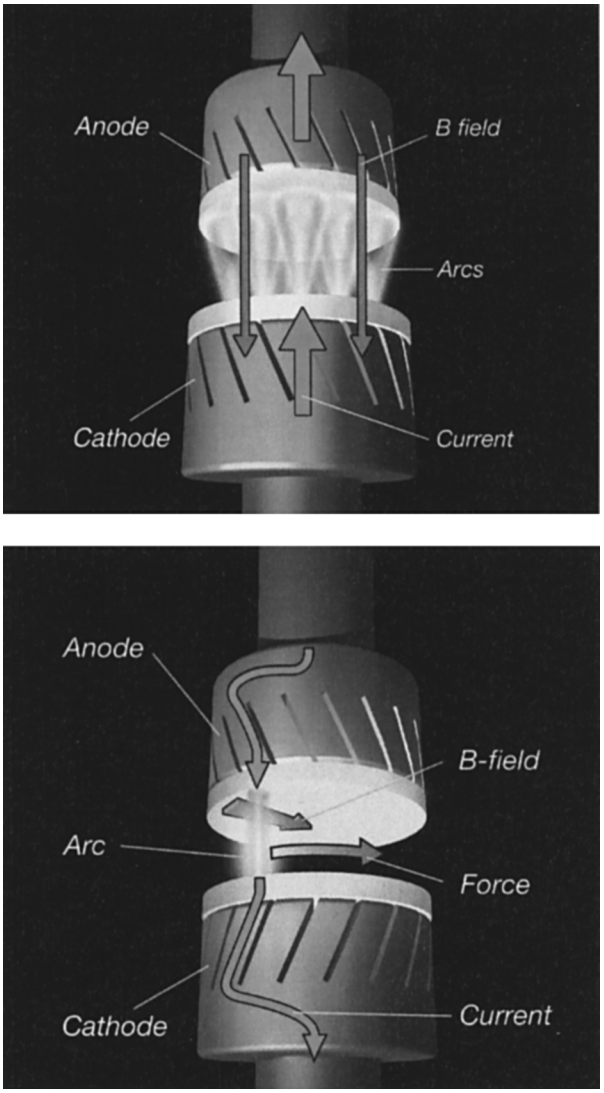


FIGURE 3 AMF and RMF arc control principles.

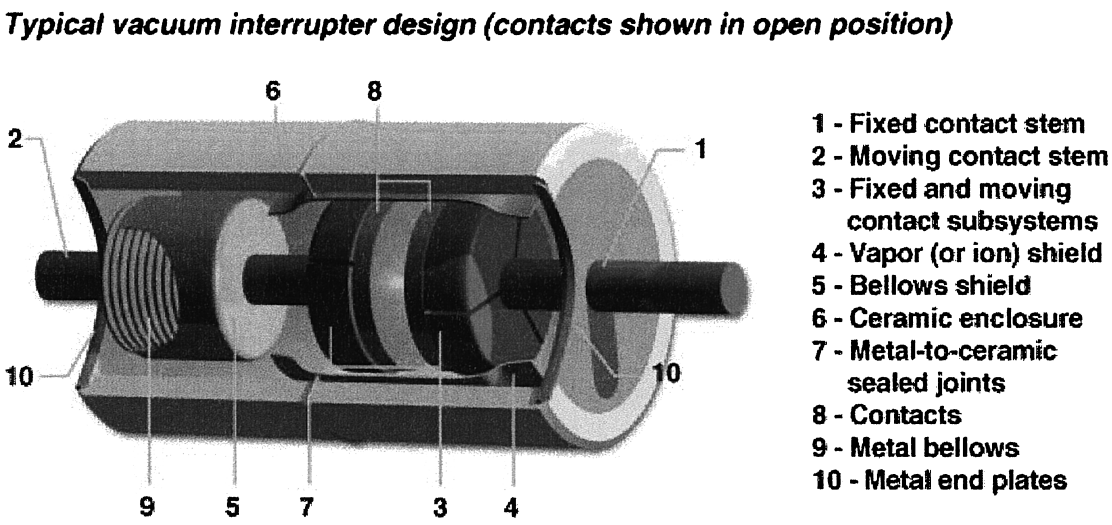


FIGURE 4 Schematic cutaway of a vacuum interrupter.

made of SS, magnetic steels, or other materials. Alumina ceramic (rarely glass) is used for the bottle envelopes.

B. Principle of Operation—Typical Characteristics of the Vacuum Interrupters

1. Electrical Characteristics

a. Closed position. In closed position the current is freely flowing through the interrupter from one contact to the other. The dominant factor affecting the performance of the VI is the contact resistance in the closed state. Typically, the resistance is in the range of several to tens of micro-ohms ($\mu\Omega$). Interestingly, the resistance of the vacuum contacts (R_C) does not depend on the contact surface area (common misconception) but is a function of ρ, electrical resistivity of the material, ζ, empirical factor (mostly 0.1–0.3, but in some cases approaching up to 1.0), H, material hardness, and the contact force F,

$$R_C = \frac{\rho}{2}\sqrt{\frac{\zeta \pi H}{F}}$$

TABLE II Typical Contact Materials Used in Vacuum Interrupters

Contact material	Typical application
Cu-Cr (with additives)	Medium voltage high current switching
WC-Ag	Low voltage, low current, frequent duty switching
W-Cu	Medium voltage, low current switching
Cu-Bi	Medium voltage, high current switching

b. Opening operation. The arc drawn between the contacts during opening operation has relatively low voltage, which is almost constant and ranges from approximately 15 to 45 V depending on the material. At currents higher than few kA the arc voltage could increase somewhat. For RMF contacts the arc voltage at higher currents could reach a few hundred volts, for AMF contacts the

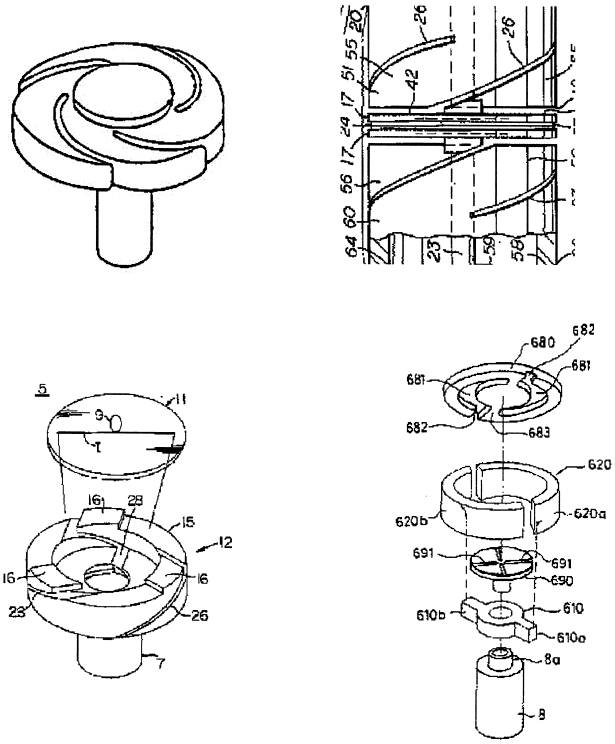


FIGURE 5 Different types of arc control geometries.

increase of only several volts is observed. In both cases the arc energy, which is a product of arc voltage, arc current, and duration of the arcing is quite low compared to other switching arc media. This is a definite advantage since it allows the designers to keep the VI dimensions to minimum without the need for extensive heat sinks, puffers, heat exchangers, etc. When the arc current reaches zero the interruption occurs. Because of the natural diffusion of plasma from the arc the VI rapidly regains dielectric strength and can support the voltage of the external electrical circuit. On rare occasions when the voltage applied by the system during the opening (TRV—transient recovery voltage) exceeds the dielectric capability of the VI, a reignition or a restrike of the arc can occur. However, since the interrupting ability of the vacuum arcs is high, the restrike or reignition current will quickly be interrupted as well.

When an external circuit drives the arc current to zero with a slow rate of change, with dI/dt of less than several A/μsec, it is likely that during the last few amperes before current zero the arc will spontaneously extinguish. These last few amperes of current cannot provide enough energy to the arc to sustain stable plasma production in the cathode spot and maintain the conductivity between the electrodes. Such a spontaneous event is called current chopping. Chopping is typical for all kinds of VIs and depends primarily on the contact material composition and slightly on processing the material and on external circuit capacitance and inductance. Typical values for contact chopping are between 1 and several amperes.

c. Open position. In fully open position the vacuum contacts can support system voltages depending on the gap length. Typical dielectric strength of the vacuum gap could reach 30 kV per mm for gaps of 0 to 1 mm. For longer gaps the relationship is not linear but rather proportional to the square root of the gap length, $V \sim (\text{gap})^{\sim 0.5}$, with the exponent factor ranging from $\sim$0.4 to $\sim$0.6.

Another important effect occurs when the open gap is subjected to an increasing voltage stress. The first time the gap breaks down, the dielectric stress could be quite low, of the order of 10 kV/mm, but at successive reapplications of voltage the strength will increase gradually to $\sim$39 kV/mm. This is known as the conditioning or cleaning effect. There is a simple physical explanation for that. On the microscopic level, typical contacts often contain minute metallic protrusions or "whiskers" which stand out from the bulk of the contact surface. When the voltage is applied these protrusions increase the electric field stress as compared to the perfectly flat contact surface. This electric field enhancement lowers the effective breakdown voltage. However, every time the voltage is applied the protrusions are melted down or evaporated by the action of the electron emission or the preceding breakdown. Thus after 30 to 50 tests the contacts are "clean" from protrusions and other impurities.

d. Closing operation. During closing of the vacuum contacts, the power system voltage is supported all the time until the gap is so short that it can no longer tolerate the stress. Typically, in the last millimeter or so depending on the circuit voltage, the breakdown occurs, and depending on the external circuit parameters, the conduction of current begins. When the contacts finally mate together the arc is no more and the change from the arcing to the fully closed state is hardly noticeable. The only concern for the designers and users of the VI is the welding that might occur when the two contacts close. If the arcing current prior to closing was too high or the contact closing was slow the arc might generate so much molten metal on the surfaces that it can weld the contacts together. Again, in typical applications, a certain limited amount of welding is possible and expected but the mechanical design of the VI drives has to take this into account. The stronger the welding during closing, the stronger the mechanism has to be to break the weld and open the contacts. Typical values for the welding stress for power vacuum interrupter contacts are 30–40 kg $*$ m/sec.

2. Mechanical Characteristics

a. Contact popping. It was mentioned that a mechanical drive to operate the VI contacts has to be strong enough to break normal welds expected when the contacts are closed. It was also explained that, in order to obtain low contact resistance in closed state, certain contact force has to be applied. There is another mechanical requirement to keep the contacts closed when a sudden surge of high current, such as a fault current, flows through the VI. Current flowing from one contact system to the other has to pass through an interface area where the actual metallic contacts touch. This area is typically very small and limited to only a few such interface points. When the current "squeezes" through these interface points, its own, local magnetic field will tend to repel the contacts. This can lead the contacts to instantaneously pop open. The magnetic forces causing this behavior are called the popping forces. The popping force can be represented by the empirical expression:

$$F_p = C * 10^{-5} * I^x$$

where F_p is in newtons, C and x are constants, and I is current in amperes. Two sets of values for C and x have been in use for power vacuum interrupters: $C = 8.0$ with $x = 1.54$ and $C = 23.6$ with $x = 1.45$.

b. Contact bounce. Every modern design of a vacuum-switching device calls for as fast a closing as possible. In actuality there is a limit on the closing speed of the vacuum interrupter contacts. A typical value for closing speed is 1–3 m/sec. This value is a compromise between the high expectations of application and design engineers and physical limits of the VI. An important consideration is a sudden exchange of momentum when two masses (two contacts) collide on closing. The higher the closing speed the more energy has to be dissipated and the more damage (deformation due to the inelastic collision) the contacts may suffer. The elastic part of the collision process can also cause contact bounce. If the bodies were perfectly elastic, the contact would never close on first impact but rather rebound at the same relative speed at which it collided. Minor bouncing is almost inevitable and acceptable. By detailed mechanical analysis of the momentum transfer between the contacts it is possible to minimize the bouncing in a given design. This is accomplished by finding a compromise between the kinetic energy at impact (by lowering the closing speed and/or lowering the contact mass) as well as balance between the elastic and inelastic collision processes (by optimizing stiffness coefficients of the contact system). The OFHC copper that is used for most of the contact systems is relatively soft, so this, by itself, can absorb some of the energy and help reduce the bouncing.

c. Contact wipe. The so-called contact wipe can take care of most of the mechanical requirements discussed above. An example of this concept is shown in Figure 6. A small spring is enclosed in a box-like enclosure. One end of the assembly is attached rigidly to a box and the other end attached to a spring inside. This wipe assembly is then mounted between a moving contact of the VI and a mechanical drive. When the contacts close, the wipe assembly acts as a cushion, minimizes bouncing, and reduces the contact deformation. In the closed position, the wipe spring provides the contact force required for the low electrical resistance. During opening, the wipe spring helps break the minor welds that develop between the contacts. This is accomplished when the drive end of the wipe assembly (lower end in Figure 6) accelerates first, the wipe spring decompresses inside the box, and after some velocity is achieved, the upper end (the VI contact side) is struck on impact breaking the contact welds.

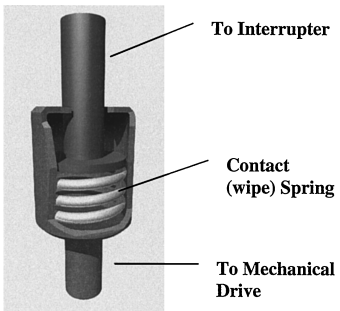


FIGURE 6 Example of contact wipe.

3. Thermal Characteristics

The main source of heat inside the vacuum interrupters is the contact resistance. Assuming a representative value of 15 $\mu\Omega$ and the rms current of 600 A, the power loss generated by the current will be $I^2 * R = 5.4$ W. At currents of 2000 A, the same interrupter will have to dissipate ~60 W. Although it seems like a small amount, all this power will be dissipated as heat and, due to the lack of gas inside, will have to be removed from the interrupter by heat conduction through the contact stems. This heat conduction is governed primarily by the temperature gradient and can be expressed by

$$\frac{dQ}{dt} = kA\frac{dT}{dx},$$

where Q is the heat energy, t is time, k is a thermal conductivity of material (OFHC Cu), A is a cross section of a conducting material, (contact stem, for example), T is temperature, x is linear distance, and dT/dx is the temperature gradient. At extremely high temperatures, above approximately 100°C, heat radiation may also play some role in the total heat performance of the vacuum device. In these cases, Stefan's law describes heat radiation, $W = e\sigma T^4$, where W is the radiated energy per unit area, e is the constant of emissivity, and σ is Stefan–Boltzmann constant. Therefore, the main consideration for keeping the interrupters cool is to provide an adequate heat path from the contact area to the outside of the interrupter. In most power applications, for interrupters rated at currents up to ~1000 A no special provisions have to be made. Existing current carrying buses, clamps, and other metallic components attached externally to the interrupter stems suffice. However, for higher current devices heat sinks, similar to the ones used in power electronics, are frequently used. In extreme applications forced cooling by means of fans and blowers can be used.

C. Typical Ratings

Commercial vacuum interrupters are used in a wide variety of switching applications. Table III summarizes the typical ratings.

TABLE III Typical Ratings of Vacuum Interrupters

Rating	
Voltage	Few hundred volts to 70–80 kV
Continuous current	Few amperes to ~4000 A
Short-circuit current (interrupting current)	Few tens of amperes to ~80 kA

D. Maintenance and Longevity of Vacuum Interrupters

Vacuum interrupters require no maintenance. There are no replacement components. No contact inspection is necessary, and not even possible, since the units are hermetically sealed for life. The only maintenance that can be performed is external. From the electrical point of view, contact resistance in the closed state and the dielectric (voltage) capability in the open state could easily be evaluated. From the mechanical point of view, an inspector can measure an incremental change in the contact wipe distance and therefore, assuming no change in the linkages and other mechanical drivers, determine the cumulative contact erosion inside the bottle.

Because of excellent electrical switching properties and rather low mechanical requirements, typical commercial vacuum interrupters are rated for a few thousand to a few million mechanical operations and from a few tens to a few thousands of short-circuit current interruptions. When required by special applications, the VIs can be strengthened even further at an appropriately higher cost. Modifying the design by replacing the bellows with longer mechanical life can greatly increase VI mechanical endurance, while increasing the thickness of the contact material will increase the electrical life.

III. VACUUM-INTERRUPTER-BASED DEVICES

A. Vacuum Circuit Breakers

The most common devices utilizing the vacuum interrupters are vacuum circuit breakers (VCB). Two categories are widely used, indoor and outdoor breakers. In a majority of cases, these devices contain three phases. Figure 7 illustrates examples of both types. The primary difference between the two is that the outdoor breaker, commonly called a dog-house breaker (for the shape of the main housing), has to protect the vacuum interrupters from adverse outdoor conditions. Electrical connections have to be provided via weatherproof, watertight bushings. Overhead wires typically bring the current to the bushings. The indoor breakers have less rigorous mounting and connections. They are used inside the cells, cube-like metal boxes

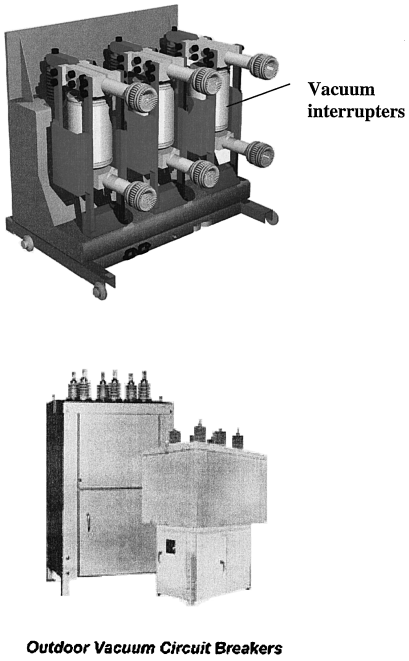


FIGURE 7 Examples of indoor and outdoor vacuum circuit breakers (Courtesy of ABB Power T&D).

that in turn go into metal frames of switchgear assemblies. Depending on whether the indoor breakers are fully withdrawable from their frames or not, such an arrangement is called metal-clad or metal-enclosed, respectively.

Individual cubicles may house breakers or other auxiliary devices, such as voltage and current transformers, fuses, buswork, etc. The cubicles can be arranged one high (i.e., one per vertical frame), or two high, two per vertical frame to save floor space. Figure 8 shows a two high design of metal-clad switchgear. Circuit breakers are designed to interrupt up to the highest short-circuit currents in the system. Their speed of operation is critical to protect the power system against the damaging effects electrical faults can have on other apparatus. Typical modern vacuum circuit breakers can clear the faults in 3 cycles of power frequency (i.e., ~50 msec) or less. The short-circuit current rating of vacuum breakers can reach up to 63 kA_{rms}.

Mechanical drives of circuit breakers are typically of a spring-type design. Two sets of springs are usually found, opening and closing. During opening, while the opening spring provides the energy for the opening action of the interrupters, it also charges the closing spring, usually not as strong. After a mechanical operation, a charging motor recharges the opening spring via an elaborate mechanical system of gears, ratchets, sprockets, etc. This complexity is necessary so the devices can meet rather rigorous industry standards (IEEE, ANSI, IEC, and others) for timing of

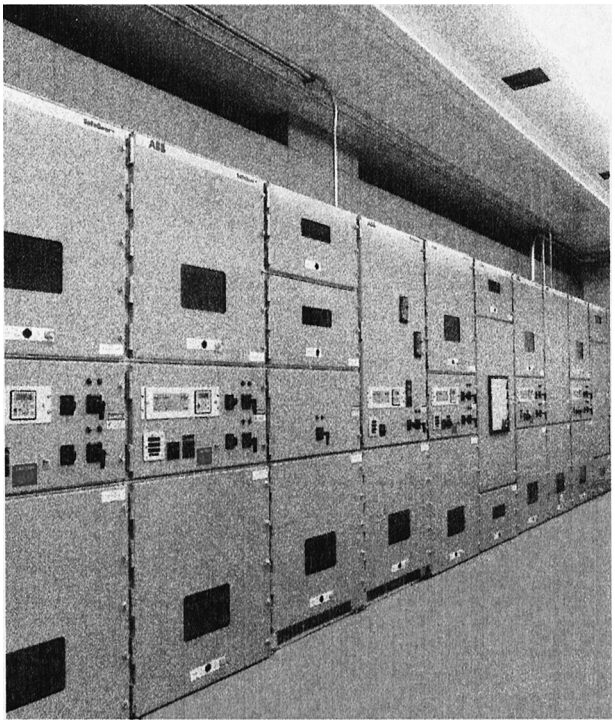


FIGURE 8 Example of two-high design of indoor vacuum switchgear (Courtesy of ABB Power T&D).

the closing and opening, trip free, antipumping, and other features. When the auxiliary power is lost to the charging motor, circuit breakers are able to perform up to three operations (O-C-O) without recharging.

B. Vacuum Reclosers, Sectionalizers, and Load-Break Switches

Vacuum circuit breakers are the most sophisticated of the vacuum-switching devices. Their mechanisms are rather complex and precise. They also require the most maintenance. The full functionality and capability of a circuit breaker is not needed in all power system applications. To save money and make power systems as simple as possible, less complex devices are often used.

For outdoor, overhead power distribution, vacuum reclosers provide an excellent lower-cost substitute for circuit breakers. Reclosers are typically mounted on the pole tops (see Figure 9). Their short-circuit power ratings are typically less than the circuit breakers (of the order of 12–16 kA_{rms}), but in widely dispersed overhead systems this can provide adequate protection against faults. Although the recloser does not have all the characteristics of the circuit breaker, it has the capability of automatic reclosing after it has been tripped by a fault. This feature is extremely useful for distribution system operators since in most cases the faults in overhead lines are temporary and clear themselves in a matter of a few to tens of seconds (lightning strike, tree branch falling on the line, an animal climbing on an insulator, etc.). A special, programmable timer can be set to reclose the vacuum interrupter up to three times. The sequence might look like this: fault trips the recloser open, 300 msec later the recloser attempts to reenergize the circuit, if the fault persists the recloser trips again. It tries a second time after 3 sec, if the fault remains, the recloser trips again. The third attempt to energize might be after 30 sec. When even this attempt fails the recloser locks out and the electrical crew has to be sent to diagnose and clear the fault.

Vacuum technology fits very well with the recloser requirements since it can easily handle frequent operations (reclosures). Vacuum interrupters do not need any extra time to recover, so even the first reclosing operation (after 100–300 msec) does not present a problem.

In the Figure 9 the recloser's three poles are made by casting vacuum interrupters in a solid dielectric and mounted on the cabinet that houses the mechanical drive. Since reclosers need to automatically sense the faults and decide on tripping sequence, the casted poles also contain instrument transformers that provide the necessary signals about the currents in the lines. That makes for a compact device, which is self-sufficient and complete. The control electronics (timer and other logic) is housed in a separate enclosure mounted at the bottom of the pole.

A vacuum sectionalizer is basically an outdoor switch, designed to automatically break (sectionalize) a long distribution line (feeder) into smaller sections when the fault occurs. Sectionalizers do not have fault handling capability; their interrupting current rating is typically few hundred amperes, corresponding to the magnitude of the load current only. Their mechanical design is quite

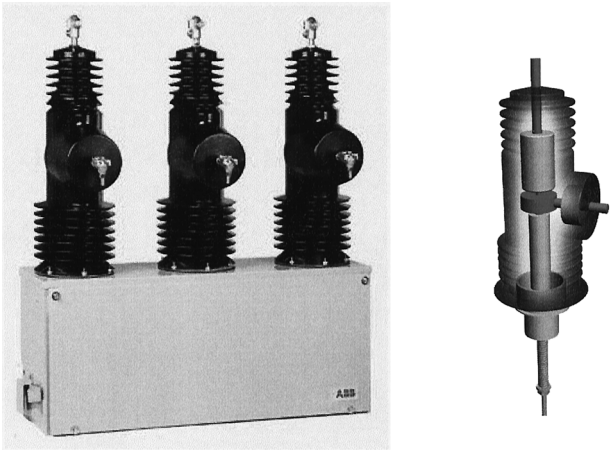


FIGURE 9 Example of a vacuum recloser and its pole assembly (Courtesy of ABB Power T&D).

simple and their vacuum interrupters often contain the simplest butt contacts without any arc controlling features. Because of that, sectionalizers are often less expensive than reclosers. Sectionalizers and reclosers often look alike from the outside. Without seeing the vacuum interrupters and knowing the logic of operation, they are almost indistinguishable.

Vacuum load break switches provide the lowest functionality in the system. They are capable of switching the load current (up to few hundred amperes) and may or may not have mechanical drives. If a drive is included, the vacuum switch can be operated remotely and be part of an automated system. In the simplest version, the vacuum switch has to be operated manually by a hot stick, handle, or similar means. Vacuum switches are found frequently in overhead and underground installations. Since vacuum switches do not provide protection in the system against short circuits, they are frequently used in combination with other devices like fuses and circuit breakers. This divides the duties: switches are used for load switching under normal operating conditions, the fuses and breakers protect in abnormal situations.

C. Vacuum Contactors

In some applications, motors and other electrical loads might have to be switched very frequently. It is not unusual for a motor (of an elevator or a conveyor belt, for example) to be switched on and off tens of times during a single day. Even the most rugged switches and circuit breakers would not last long under those kinds of conditions. What complicates the situation even further is that, during starting, motors draw high current, comparable to the short-circuit current. The need is therefore to have a device that can switch on high momentary currents, switch off normal load currents, and do this repeatedly. The unique features of vacuum interrupters come to the rescue. With low mechanical energy required to move the contacts, and excellent, low energy arc characteristics, vacuum interrupters can be designed into contactors capable of meeting all the requirements. With simple solenoid drives, very few moving parts, and small but durable interrupters, contactors' mechanical longevity can easily exceed a million operations without maintenance. Special contact materials used in contactor bottles, like WC-Ag or W-Cu, provide excellent switching characteristics with low chopping current and minimum erosion.

Vacuum contactors are popular in mining and petroleum industries where fire hazard is a concern. Vacuum bottles seal the arcing securely and eliminate the risk of accidental sparking or ignition of nearby flammable materials.

Vacuum contactors can be of locked or unlocked type depending on specific requirements. When the auxiliary power is lost to the drive solenoid the unlocked contactor will drop its contacts. In the locked version the contacts will stay open unless the lock is released or tripped.

D. Triggered Vacuum Gaps and Triggered Vacuum Switches

Mechanical closing of any contacts, even in vacuum, takes time. Even the strongest, most powerful mechanisms require at least milliseconds to bring contacts from their fully open position to their touch. It might not be fast enough for applications where extreme speed is required. Such might be a case of protecting serious capacitor banks used to compensate reactances of transmission systems. Also, fusion reactors have to be pulsed with electrical currents with the precision of a few tens or hundreds of microseconds ($1\ \mu s = 10^{-6}$ sec). A set of vacuum contacts, stationary or movable, can be equipped with a special trigger (see Figure 10) typically embedded in the contact body with the electrical connection brought outside. This three-terminal device can act very much like a silicone controlled rectifier (SCR) or similar power electronics device, except that it can handle much higher currents and voltages ratings. When a pulse of a few hundred volts and duration of a few microseconds is applied to the trigger electrode, a small discharge is established between it and the main contact. A special substance, like a titanium hydrate, acts as a release agent of ions and electrons to aid the arc ignition process and help spread the discharge plasma into the main intercontact gap. With the voltage between the main contacts, the main vacuum arc ignites within microseconds and commutates the circuit. With the vacuum arc voltage only of the order of several volts, the main power circuit does not even know that the arc is there.

In triggered vacuum gaps (TVG) the main contacts never move (close), so the conduction of current ceases when the main circuit current reaches zero. To reestablish the flow, TVG has to be triggered again after every current zero. Triggered vacuum switch (TVS) has one of the contacts movable, very much like the ordinary vacuum

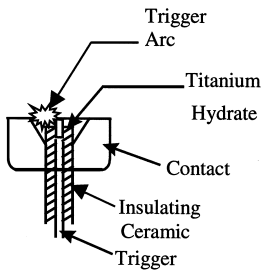


FIGURE 10 Trigger of a triggered vacuum gap or triggered vacuum switch.

interrupter, so simultaneously with the triggering operation a signal is sent to the mechanical drive to start closing on the contacts. This way, after a few milliseconds a metallic connection is established and the arc eliminated. TVS will conduct the current until the mechanism receives the signal to open and drives the contacts apart.

Both TVG and TVS are rather unique devices and are used only if all other means of electrical switching are exhausted. Commercial TVGs are available on the market today with voltage ratings of up to several kV and momentary current ratings of several tens of kA.

IV. APPLICATIONS IN POWER SYSTEMS

A. Application Related Characteristics of Vacuum Devices

Rating structure of vacuum circuit breakers can serve as a good example of various specifications that vacuum devices meet in power systems. Table IV illustrates the most commonly used rating structure.

B. Electrical Interaction between Vacuum Devices and Other Power System Components

Electrical transients caused during closing and opening of an electrical circuit result from the interaction of the switching device and its surrounding system. Power system planners and designers calculate electrical transients to predict the worst stresses that the system has to be able to withstand. To accomplish this task the electrical switching characteristics of the vacuum device and the transient equivalent circuits of all other surrounding apparatus have to be jointly assessed. Most of the time only two main characteristics of the vacuum switching have to be known: dynamic dielectric behavior of the contacts in the nonconducting state and dynamic interrupting behavior of the vacuum arc in the conducting state. Those two characteristics fully define how the switch will behave in any given power system and how its interaction with the system might influence the generation of the switching transients.

TABLE IV Rating Structure of a Vacuum Circuit Breaker

Rating	Sample value
Nominal system voltage	15 kV
Rated maximum voltage	15.5 kV
Continuous current	2000 A
Rated frequency	50 or 60 Hz
Power frequency voltage withstand	50 kV
Power frequency voltage withstand (when wet, for outdoor devices)	45 kV
Rated impulse voltage withstand	110 kV
Short circuit interrupting current	25 kA
Rated interrupting time	50 msec
Rated close-and-latch current	65 kA
No-load mechanical endurance	30,000 operations
Full-load-current switching endurance	10,000 operations

Let us examine one possible switching scenario shown in Figure 11. A vacuum switch is installed between the power system (voltage source on the left) and an inductive load, such as a motor, or transformer (reactor on the right). When the switch is called to operate it might happen that the actual, physical contact separation occurs just before the natural current zero of the power frequency. In that case, since the current is always interrupted at zero, the contacts at that instant of time will be apart by only a fraction of a millimeter. Vacuum contacts can handle the interruption with that contact spacing but the resultant transient recovery voltage (TRV) generated by the inductive load might be too severe for that gap. Inductive circuit TRV has a rather fast rate of increase and an amplitude of up to 2 pu, i.e., twice the system peak voltage. This TRV impinged on the short contacts of the VI may cause the interrupter to break down. This, in turn, will re-establish the vacuum arc and the flow of current. This re-ignition current might have high frequency components that will oscillate around the current zero value. Again, because of the excellent interrupting capability of vacuum these high frequency currents will be quickly interrupted at one of their current zeros. During this short re-ignition time (of the order of

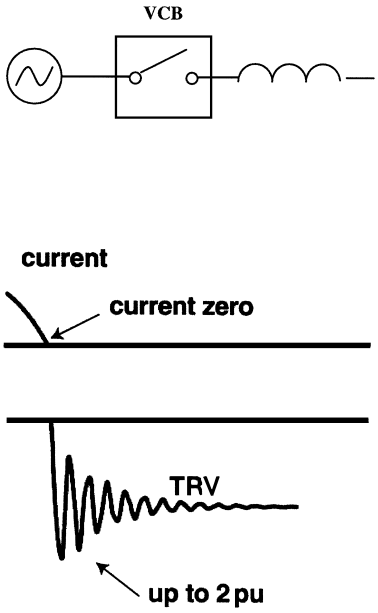


FIGURE 11 Switching an inductive load in the power system with current and voltage waveforms of the VCB.

tens of μsec) a small amount of magnetic energy may build up in the inductive load. After the re-ignition current is interrupted the new TRV might now be of even higher magnitude (more than 2 pu) since the extra energy in the inductive load may contribute to the build up. In the worst-case scenario, this build up of magnetic energy in the inductive parts of the power circuit may be greater than the natural damping (dissipation) of the high frequency transients that the power system can provide. As a result, the vacuum switch might continue re-igniting under the increasing TRV stresses and re-interrupting the high frequency surges of currents. This escalation process can be dangerous and either the vacuum device or the other components of the power system will eventually suffer damage. Fortunately, these situations are very rare although possible. It is a good application of engineering practice that, when the behavior and the parameters of the power system are unknown, the surge suppression (like voltage arresters, surge capacitors, surge R-C circuits, etc.) be used.

V. RESEARCH AND DEVELOPMENT RELATED TO VACUUM SWITCHES

Commercial vacuum switching devices are far from being an old technology. Continuous research and development activities in this field demonstrate that the potential for improvement and new solutions is significant. A few examples illustrate the most significant R&D opportunities.

A. Continuous Optimization

The industry continues to find better, more efficient way of utilizing the concept of vacuum interruption and switching in power systems. One of the recent trends is to use a standard, ceramic-enclosed vacuum interrupter and encapsulate it in a solid dielectric, like epoxy or polyurethane. Modern VIs are very compact and use as little ceramic as possible to keep the cost down. Since the weakest dielectric path for breakdown is in air, along the outside of the ceramic, from one contact end to the other, encapsulating the VI in a solid dielectric removes this weakness and can noticeably improve the voltage rating of the entire assembly. Many commercial devices are already available with these integrated, encapsulated pole assemblies.

Advances in metallurgy, powder chemistry, and process technologies result in improvements in new contact materials. Although the fundamental formulations remain unchanged, small amounts of additives, such as Te, Bi, and others, can alter electrical and mechanical properties, lower the current chopping levels, lower the welding strengths, increase the contact toughness or reduce the material erosion due to arcing.

Great emphasis is placed on designing vacuum-switching devices that would require less maintenance by the customers. Vacuum interrupters themselves are maintenance free but the operating mechanism, mounting of the poles, auxiliary circuits, etc., require periodic inspection. The most typical maintenance requirement for vacuum circuit breakers is lubrication of moving parts to ensure that the timing of the breaker remains within the tolerances. Also the cleaning of high voltage parts, like encapsulated poles and bushings, is a standard procedure. New technologies are being researched, by which no lubrication might be required to operate mechanical drives. From the electrical point of view, the materials sciences provide new opportunities for insulating materials that are self-cleaning, and recover their original surface properties even after exposure to contaminants and severe weather conditions.

Although vacuum switches are simple on–off devices, more sophisticated operating logic can add functionality and therefore improve the overall value of the technology. A perfect example of this is a vacuum recloser, mentioned earlier, by which a vacuum interrupter can be programmed to trip, and reclose three times to automatically restore the power circuit. The operating logic of reclosers has been improving with more coordination between the different units, self-diagnostics, and smart recloser timing depending on the actual state of the power system.

B. New Designs of Contact Geometries

The physics of the beneficial effect of the magnetic field on arcing in vacuum is still not fully understood. A great deal of research activity is focused not only on understanding the phenomena but also on utilizing them to achieve new higher ratings within smaller volumes and dimensions. The clear objective is to design smaller, lighter contacts, to minimize the mechanical energy requirements to operate them and to increase the current interrupting capability. What the actual physical limit is of the maximum current density that perfect contacts can handle, researchers do not agree. Many feel that the technology of today is far from that physical limit. Since most agree that the secret is in the magnetic field, R&D efforts focus on designing more sophisticated magnetic field contact geometries generating quadrupole fields, and nonlinear and nonuniform field distributions.

C. Low Energy Mechanical Drives (Magnetic Actuators)

One of the more significant recent technological accomplishments is to introduce magnetic actuators to power circuit breakers to replace the conventional, spring-motor

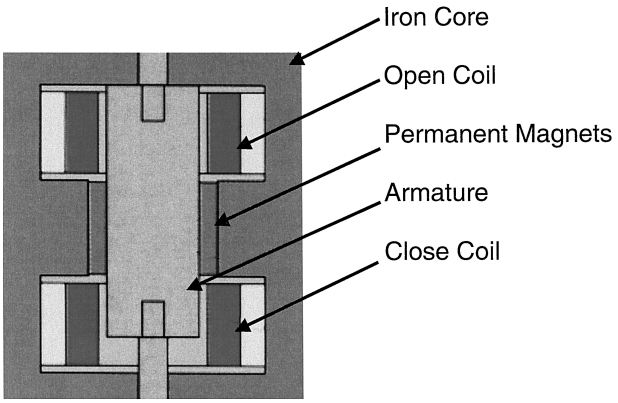


FIGURE 12 Magnetic actuator used in vacuum circuit breakers (Courtesy of ABB Power T&D).

mechanical drives. Fewer moving parts, higher reliability, more functionality, and smart electronic control logic are clear advantages of magnetic drives. Their principle is similar to the well-known solenoids found in all electrical relays but with the addition of a permanent magnet that is embedded in the solenoid (see Figure 12). The permanent magnet provides a holding force that can keep the contacts in either closed or open position with no external energy required. This is a significant improvement over the relay or a contactor, where the solenoid has to be continuously energized in order to keep the main contacts closed. The energy is required only to change the state of the actuator from open to closed or vice versa. This is provided by pulsing a coil of the actuator with a short surge of current, typically several milliseconds. The coil generates a magnetic field that instantaneously opposes the field of the permanent magnet and reverses the net field. This moves the armature, which is connected to the vacuum interrupters. Two main types of magnetic actuators can be found today, single coil and dual coil. In a single coil design the closing of the actuator is performed with the energy of the pulsed coil and the permanent magnet, but the opening is performed by the opening spring. On opening the coil is pulsed only to cancel the permanent magnet force holding the contacts in closed position. The dual-coil design, as shown in Figure 12, is symmetrical, in a sense, and uses two coils to deliver both the closing and opening energy.

The new, stronger, rare-earth magnets can provide holding forces even for the high short-circuit current ratings of vacuum circuit breakers, making the magnetic actuator concept attractive to the entire family of vacuum devices.

D. Other R&D Activities

All three phases of vacuum circuit breakers are typically operated in a ganged manner, that is, all three either close or open with the same speed and the same relationship to each other. However, the three phases of the power system are 120 electrical degrees apart, about 5.6 msec for 60 Hz. To do the best switching possible and minimize its impact on the system, the three phases of the breaker have to switch independently but in synchrony with their respective power system phases. This is called synchronous switching. A great deal can be gained from synchronous closing and opening. For example, the capacitor banks can be energized with less stress to the breaker and the power system if the contacts close around the zero value of the voltage across the respective phases. On opening one can eliminate even the possibility of re-ignitions by avoiding separating the contacts just before current zero. Many more examples are well known to power system designers and operators.

The ability of precisely switching the poles of the vacuum device can be utilized in even a more advanced context. Imagine a "smart" breaker that can read and interpret its voltage and currents and make the switching decision automatically depending on the state of the power system. The trend is to combine the capabilities of the vacuum breaker technology with the protective relays of power systems in one functional unit, either centralized or decentralized, depending on the application. Such an integrated device or system can then perform other functions of self-monitoring and self-diagnostics and communicate automatically with the neighboring systems. It is expected that in the next few years these concepts will become reality.

VI. SAFETY

When high voltage appears across two contacts in vacuum, X-rays can be emitted from the contact surface due to the electron bombardment. These electrons are the result of so-called cold emission. The electrons from the negative cathode accelerated in the electric field between the contacts strike the anode and cause secondary electron, and X-ray emission. Under all normal operations, with voltages within the ratings, no measurable X-rays are ever emitted from commercial vacuum interrupters. However, when the contact separation is less than normal and the voltage across the vacuum device is excessive, some X-rays may be generated. Such might be the case when a 38-kV vacuum interrupter is subjected to 70 kV during routine, ANSI-standard-required voltage withstand tests. In such cases, distance of a few meters is required for safety. This requirement coincides with the normal safety distances due to electrical hazards around high voltage equipment. If required, added protection of thick, soft-metal shields can be installed around the test cell areas where the vacuum devices are tested.

SEE ALSO THE FOLLOWING ARTICLES

POWER ELECTRONICS • POWER TRANSMISSION, HIGH VOLTAGE • VACUUM TECHNOLOGY

BIBLIOGRAPHY

Boxman, R., Martin, P., and Sanders, D., eds. (1995). *Handbook of Vacuum Science and Technology. Fundamentals and Applications*, Noyes Publications, Park Ridge, NJ.

Browne, T. E., Jr., ed. (1984). *Circuit Interruption. Theory and Techniques*, Dekker, New York.

Cobine, J. D. (1958). *Gaseous Conductors. Theory and Engineering Applications*, Dover, New York.

Flurscheim, C. H., ed. (1985). *Power Circuit Breaker Theory and Design*, Peter Peregrinus, on behalf of the IEE, London.

Greenwood, A. (1994). *Vacuum Switchgear*, Institute of Electrical Engineers, London.

Lafferty, J. M., ed. (1980). *Vacuum Arcs. Theory and Applications*, John Wiley & Sons, New York.

Latham, R. V. (1981). *High Voltage Vacuum Insulation. The Physical Basis*, Academic Press, New York.

Proceedings of the International Symposia on Discharges and Electrical Insulation in Vacuum (IS DEIV) can serve as an excellent source of up-to-date and technical information on vacuum arcs and devices.