Physical limitations on miniaturization in microelectronics

Yu. V. Gulyaev, V. B. Sandomirskiĭ, A. A. Sukhanov, and Yu. Ya. Tkach

Institute of Radio Engineering and Electronics, Academy of Sciences of the USSR

Usp. Fiz. Nauk **144**, 475–495 (November 1984)

The various types of physical limitations on the miniaturization of active elements are classified. Those physical limitations on the sizes of elements which arise in the manufacture of integrated circuits are analyzed. The various devices are classified by size. A scaling model which describes the changes in the parameters of the devices in the course of miniaturization is discussed. The limitations of scaling are discussed. Physical limitations which are set on the sizes of elements by their operating mechanism are examined. The most important limitation on the degree of integration of elements on a chip and on the working speed—the product of the number of elements and the reciprocal of their switching time—is imposed by the heating of the chip. It is concluded that the physical limitations are unimportant down to dimensions of the order of 0.1 μm; that in order to reduce the size of elements below 0.1 μm it is necessary to produce elements which have no p-n junctions and which are heterostructures; and that the change in conduction mechanism in the structures with dimensions less than 0.1 μm will require the development of new types of elements: ballistic and tunnel transistors.

1. INTRODUCTION

Developments in microelectronics are now coming so rapidly and are being so effectively adopted in various spheres of human activity and in meeting the needs of mankind that there is talk of an "electronic revolution."[1–7] Just what is microelectronics?

Microelectronics is that field of electronics which combines a complex of physical, engineering, and technological problems aimed at developing complex electronic circuits for processing and transmitting information.

Modern microelectronics is based on solid-state microcircuits. They are fabricated by an integrated planar technology which makes it possible to fabricate, in a single technological cycle, a large number of elements of the same type near or on the surface of a solid. In practice, microcircuits are fabricated on an individual crystal or "chip," generally of silicon, and consist of discrete elements which interact with each other only through a system of special "interconnections," which unambiguously determine the functional properties of the integrated circuit. The fact that an integrated circuit can be represented as a set of individual elements, each of which in turn has clearly distinguishable homogeneous regions (a drain, a source, a gate, etc.) is the content of the so-called partition principle.[6]

According to this principle, an integrated circuit and the elements of which it consists are characterized by the following parameters:

N, the number of active elements on the chip;

τ, the time required for the switching of an element between two electrical (or logic) states by a signal;

$\nu = 1/4\tau$, the cycling frequency;

$N\nu$, the working speed of the circuit;

P, the power dissipated during the switching of an element;

$P\tau$, the quality index of an element, which determines the amount of energy dissipated by the element per switching event; and

d; the scale dimension of the active region of the element (the length of a channel or the width of a base), which is determined by the resolution of the technological process (the smallest line width).

Equally important characteristics of a microcircuit, which determine whether it can be used extensively, are the cost per bit of data (C) i.e., in practice, the cost of one element, and the operating reliability of the circuit, which is determined by the probability of its failure.

On the basis of these parameters we can draw a picture of the present state of microelectronics and of the pace of its development[7,8] (Table I). Table I is supplemented by P,τ and N,ν diagrams in Figs. 1 and 2, which show the regions of parameter values characteristic of the various types of devices.

It can be seen from Table I and Figs. 1–3 that progress in microelectronics is actually being achieved through a reduction of the sizes of elements and a corresponding increase in their packing density. As an example of the present trends

TABLE I. Parameters charcterizing the state and the dynamic development of microelectronics

Years	1980	1985–1990
N, elements/chip	10^4—10^5	$3\cdot10^5$—10^6
τ,s	10^{-5}—10^{-6}	10^{-8}—10^{-9}
ν, Hz	10^5	$2.5\cdot10^7$
Speed, element·Hz	10^{10}—10^{11}	10^{13}
P,W	10^{-4}	10^{-5}
$P\tau$,J	10^{-9}—10^{-10}	10^{-13}—10^{-14}
c, dollar/bit	10^{-4}	10^{-5}
d, μm	3,5	0,5

0038-5670/84/110868-13$01.80

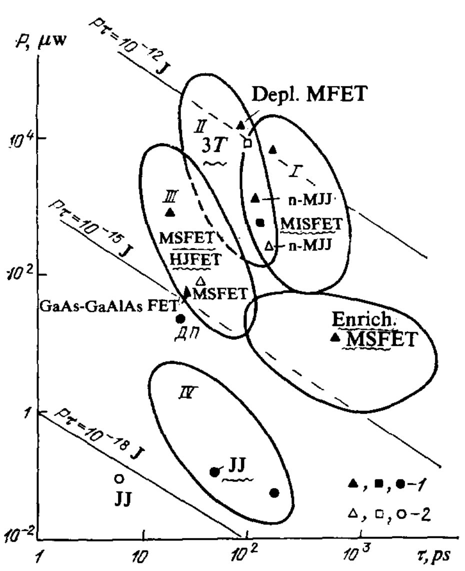

FIG. 1. Power dissipation versus the switching time (*P*, τ diagram for devices of various types. 1: Characteristic points for existing devices. FET) Field-effect transistors; BT) bipolar transistors; JJ) Josephson junctions. 2: Calculated points for the same devices. MIS) Metal-insulator-semiconductor; MS) metal-insulator; HJ) heterojunction. The outline regions are the parameter regions characteristic of various devices. I—Metal-insulator-semiconductor field-effect transistors (MISFET); II—bipolar transistors; III—metal-semiconductor field-effect transistors (MSFET); IV—Josephson devices.

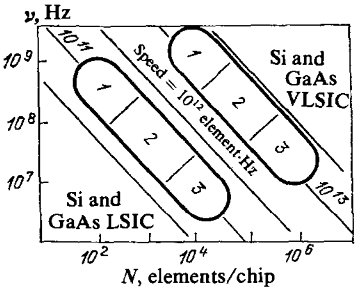

FIG. 2. Present state (1980) and future (1985–1990) of very large-scale integrated circuits. 1—Integrated circuits for control and computation; 2—radar digital processor; 3—sensitive signal processor. Working speeds in the various fields of application.

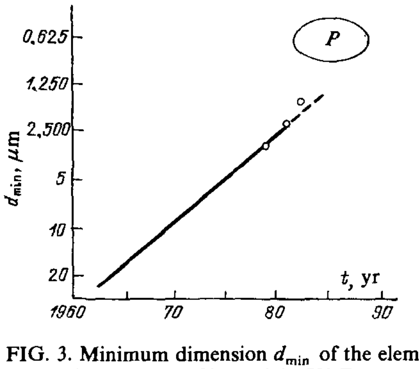

FIG. 3. Minimum dimension $d_{\min}$ of the elements of integrated circuits versus the time *t*. *P*—Plans of the US Department of Defense.[10,36]

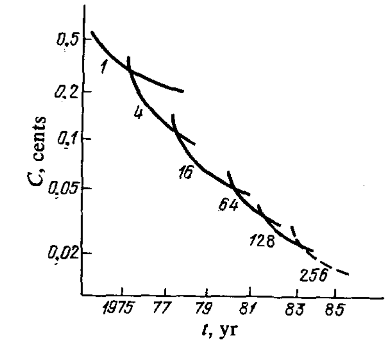

FIG. 4. Lowering of the cost per element (*C*) of memory circuits of various capacities over the time *t*. The curves are labeled with the capacity of the memory circuit in Kbits.

in the development of microelectronics we note that a 20-fold improvement in the operational characteristics of an integrated circuit on a crystal consists of a twofold improvement due to refinements of the circuit technology and an increase in the dimensions of the crystal and a tenfold improvement due to miniaturization of the elements and an increase in their packing density.

It is important to note that an increase in the number of elements on a crystal at essentially no change in its cost is accompanied by an equally rapid decrease in the cost of an individual element (Fig. 4).

Over the past twenty years we have witnessed an annual doubling of the number of elements on a crystal, *N* (Moore's law).[8] It is expected that the number of elements on a chip will continue to increase exponentially in the future, but the exponent will be lower; specifically, *N* will double every two years[9] and will reach 10^9 by the year 2000 (Fig. 5).[2]

It is pertinent to note some of the programs for developing very large-scale integrated circuits (VLSIC). The US Department of Defense has developed a six-year program (1979–1985) for developing ultrafast integrated circuits for armament systems with elements of submicron size (down to 0.5 μm by 1985) and with a working speed of[10] 10^{13} element·Hz. Processors made from ultrafast integrated circuits are to perform 10^9 operations per second and will be used in real-time electronic systems.[1)] The Pentagon originally planned to spend $210 000 000 on the ultrafast-integrated-circuit program, but appropriations have subsequently been increased to[11] $320 000 000.

Another program in the USA is the program on Space-Hardened Advanced Processing Elements (SHAPE), with the goal of producing radiation-resistant very large-scale integrated circuits at Cornell University. Work being carried

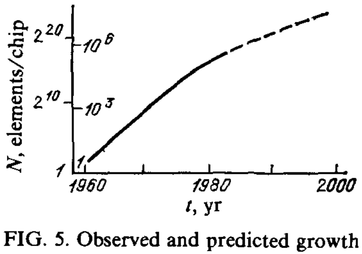

FIG. 5. Observed and predicted growth of the number of elements on a crystal (chip).

[1)] If they were constructed from large-scale integrated circuits, they would weigh 10^4 kg and would draw a power of 100 kW. Since the cost of 1 kg of excess weight for a satellite, for example, is $5000, and the cost of energy is $5000/W, it becomes clear that they could not be constructed from large-scale integrated circuits.[10]

out as part of this program is intended to improve the radiation resistance of very large-scale integrated circuits (the goal is to raise the tolerable dose to $5 \cdot 10^4$ rad) for use in space applications.[12]

A government program to develop very large-scale integrated circuits has been developing in Japan since 1976. This program was developed with the goal of winning the market for electronic apparatus and is being financed by the Ministry of External Trade and Industry. The primary goal is to develop by 1989 the world's fastest compter, with a speed of 10^{10}–10^{11} operations per second and with a memory of 10 Gbits, i.e., with characteristics two orders of magnitude better than those of present-day IBM 370*s*. The plan was to develop by 1983 single-chip memories with a capacity of 256 kbits and 32-bit microprocessors with a random-access memory of 32 kbits. Only half of this program is being financed by the government, and after the technology is developed the companies involved are obligated to return to the government the money it will have spent. Three joint ventures are implementing this Japanese program for developing very large-scale integrated circuits: a reasearch laboratory established especially for the purpose, a computer development laboratory, and an independent company, NEC-Toshiba Information Systems.

A question which arises along with the development of integrated circuits is that of the physical limitations on the minimum dimensions of the elements in microelectronics and on the greatest possible degree of integration. The problem of physical limitations has become particularly acute in connection with plans to develop very large-scale integrated circuits with more than 10^5 elements on a chip. For very large-scale integration, the size of an individual element is below the micron level and is approaching the characteristic lengths which determine the operation of an element, such as the width of the space-charge region at the p-n junction and the mean free path.

In this paper we will be focuing on the physical problems which limit the dimensions of elements in microelectronics. In this connection we wish to stress the following point: The ultimate and unchanging goal of microelectronics from the moment of its birth has been to increase working speed. From this standpoint, the problems which we will discuss here are by no means the only pertinent problems, although they are the most urgent. We will have essentially nothing to say about, for example, the problems of new materials, new principles, and the associated effects; we will touch only briefly on the important problems of interconnections (which become particularly acute for ultrafast integrated circuits), radiation resistance, etc. Each of these questions contains some important physical problems, but in the present paper we will be discussing only those questions which are pertinent to the physical operating principles of the active elements of modern microelectronics.

The question of physical limitations arises only in a discussion of specific elements of integrated circuits, existing or proposed.

The limitations on the sizes of elements can be classified on the basis of the partition principle. Three requirments must be met in order to achieve a high degree of integration: First, we must be able to fabricate elements of sufficiently small size. Second, each separate element must operate normally. Finally, all the elements must operate normally in an integrated circuit. It follows that the possible limitations on the sizes of elements can be classified in the following three groups:

1) The physical limitations which arise in the manufacturing technology of integrated circuits (the limitations which stem from the scattering of the exposing beam, the fluctuations of dopants, etc.).

2) The physical limitations on the operation of the individual elements (the limitations which stem from the joining of p-n junctions, breakdown due to heating of electrons, etc.).

3) Those physical limitations on the degree of integration N and the sizes of the elements which result from interactions between elements (Joule heating).

The physical limitations related to the operation of the individual elements for field-effect triodes and bipolar transistors have much in common, so we will discuss these limitations for only one case, that of field-effect triodes.

2. PHYSICAL LIMITATIONS IN THE MANUFACTURE OF VERY LARGE-SCALE INTEGRATED CIRCUITS

Integrated-circuit elements are a set of heterogeneous and generally nonequilibrium regions with sharp boundaries. About a hundred methods for fabricating them are now available in integrated-circuit technology. These methods differ primarily in the isolation of the devices from each other (by means of p-n junctions or insulating layers), in the method used to fabricate the p-n junction (diffusion, ion implantation, or layer-by-layer epitaxial growth), in the structure of the system of interconnections (metal connections, lines of polycrystalline silicon, or channels of the appropriate conductivity type), in the insulation between levels, etc.[13]

The methods used most widely at present are[13] lithography, doping by diffusion through oxide masks, local oxidation through nitride masks, epitaxial growth, ion implantation, anodic oxidation, and metallization. In particular, the lithographic method makes it possible to produce a raised image on a surface—"windows"—for subsequent selective processing of the material in the windows. The lithographic process includes the deposition of an organic resist; the exposure of this resist to some type of radiation through a template or through direct writing by a beam, which results in polymerization of depolymerization of the resist; and the removal of the unexposed regions (for so-called negative resists) or of the exposed regions (for positive resists) by dissolution, heating, or ultraviolet radiation. The lithographic process is usually followed immediately by a local etching of insulating or metal films which transfers the pattern of the phototemplate to an inorganic and insensitive material deposited before-hand on the surface of a semiconductor wafer.

As an example we will briefly describe the sequence of operations in the fabrication of an *n-p-n* bipolar transistor with isolation by a p-n junction.[13]

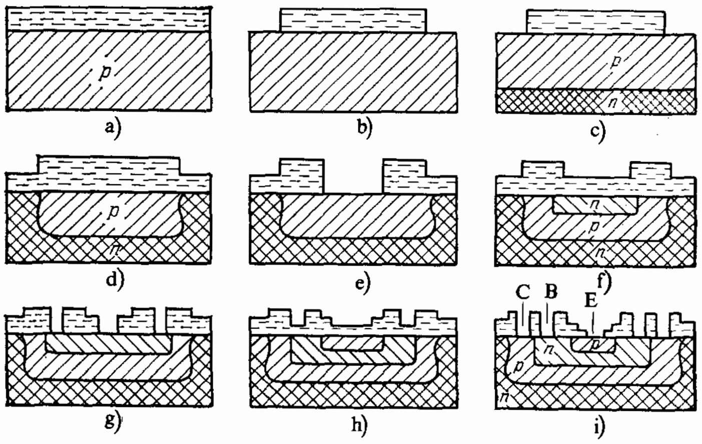

FIG. 6. Sequence of operations in the fabrication of an *n-p-n* bipolar transistor with insulation by a p-n junction.

1. The initial p-type wafer is oxidized, and windows are produced by lithography in the oxide for the production of the isolating p-n junction (Figs. 6a and 6b).

2. An n-type impurity is introduced in the windows and on the backside of the wafer, and a diffusion annealing is carried out in order to form an isolating p-n junction (Figs. 6c and 6d). It should be noted that the diffusion process is accompanied by the oxidation of the silicon surface (Figs. 6c, 6d, and 6g).

3. After the appropriate windows have been fabricated in the oxide film, a doping is carried out in the n-type surface regions through thermal diffusion for the purpose of forming the base regions (Figs. 6f and 6e), emitter regions, and the collector contact regions (Figs. 6g and 6h).

The following methods have been proposed for fabricating submicron elements and very large-scale integrated circuits: electron beam, x-ray, and ion lithography; ion- and electron-beam etching; laser and ion processing; molecular-beam epitaxy; and ion implantation. Progress in the development of these methods, primarily electron-beam lithography, has made it possible to achieve a resolution of 100 Å in the plane of the crystal and a resolution of only a few angstroms in the perpendicular direction.[14–17]

Comparing this figure ($\approx 0.01\,\mu$m) with Table I, we thus see that the resolution o the lithographic methods will run into essentially no limitations on element size in the immediate future. As will be seen below, however, other and more important limitations arise in the fabrication of semiconductor integrated circuits, e.g., limitations stemming from fluctuations in the concentration of a dopant and from its surface diffusion.

Let us summarize the main technological limitations.

a) Smearing of the edge of an exposed region

Until quite recently it was believed that the primary limitation in the technology stemmed from the photolithography and was due to the diffraction blurring of the edge of the exposed region (Δx) over a distance greater than the wavelength λ (Refs. 17-19):

$$\Delta x \geqslant \lambda. \tag{1}$$

For visible light, λ is about $1\,\mu$m, and here photolithography is approaching its limiting capabilities: Line widths of 3–5 μm have been achieved. The use of x-ray or electron lithography will substantially reduce λ and thus the minimum line width; for electrons with an energy $E = 10$–10^3 eV, for example, we would have

$$\Delta x \geqslant \lambda = \frac{\hbar}{\sqrt{2mE}} \approx 1-0.1 \ \text{Å}. \tag{2}$$

b) Scattering of the beam in the resist and in the semiconductor

One of the limitations in lithography results from the scattering of the (electromagnetic or electron) beam in the resist and in the semiconductor.[18,20–23,25] The organic resists in use (such as polymethyl methacrylate) have a minimum thickness of about 50 Å (a few molecular diameters). The scattering of the beam in the resist results in the smearing of a line by an amount of the order of the thickness of the resist, i.e., 50 Å. In this stage of the development of microelectronics, therefore, the limitation resulting from the scattering of the beam in the resist is unimportant. Furthermore, the resolution can be brought down below 10 Å by using a glassy chalcogenide semiconductor as resist.[24]

Limitations of the same order of magnitude arise in the etching of the resist. In this case the scale dimension determining the smearing is the length of the resist molecules.[16,23]

During the exposure, the electrons which have struck the semiconductor may bombard the resist as a result of back-scattering effects; as a result, there will be a further blurring of the edge of the exposed region of the order of the electron mean free path in the semiconductor (l). For electrons at an energy of 25 keV the mean free path reaches $3\,\mu$m. The back-scattering blurring can be reduced by using lower-energy electrons.

There are plans to use ion lithography in the near future and also to combine electron and x-ray lithography. In the latter case, low-energy electrons would be used to prepare a high-resolution mask, which would then be used for the x-ray lithography of the samples. This approach will significantly increase the cost of the fabrication of integrated circuits, since electron lithography is presently the most expensive technological process. The secondary electrons which are excited by the x radiation and which return to the resist will cause a comparatively slight blurring, of the order of 100 Å.

c) Restrictions due to the spherical aberration of the electron beam[25]

An increase in the electron-beam intensity I increases the size of the spot as a result of the Coulomb repulsion of the electrons in the beam (spherical aberration), while a decrease in I increases the required exposure time and thus the cost of the lithographic process. We will discuss how the combination of these factors leads to lower limits on the size of the exposing beam and thus the minimum size of an element.

The radius of the electron-beam spot (R) must be greater than its transverse aberrational blurring:

$$R > \frac{f}{4}(BS)^{-3/2} I^{3/2} = kI^{3/2}, \tag{3}$$

where f is the spherical-aberration constant, B is the brightness of the source, and S is the area of the emitting surface. Since electrons reach the target in a random manner over time, the total number N_e of electrons which reach the target during the exposure time t_{exp} must be quite large: $N_e > N_{\mathrm{m}}$. Here N_{m} is set by the condition that at $N_e = N_{\mathrm{m}}$ the standard deviation must be much smaller than N_{m}. Knowing N_{m}, we can estimate the minimum intensity; substituting the latter into (3), we find

$$R > R_1(t) = k\left(\frac{eN_{\mathrm{m}}}{t}\right)^{3/2}. \tag{4}$$

On the other hand, an increase in the time over which the spot is exposed will increase t_r, the total time required for the exposure of a chip of area S_r:

$$t_r = \frac{S_r}{\pi R^2}\, t. \tag{5}$$

We denote by q the cost of operating the exposing apparatus per unit time. The cost of the exposure process, $C_{\mathrm{exp}} = t_r \cdot q$, must not exceed a certain maximum C_{M}. Combining this condition with (5), we find the following restriction on the spot radius:

$$R \geqslant R_2(t) = \left(\frac{S_r\, tq}{\pi C_{\mathrm{M}}}\right)^{1/2}. \tag{6}$$

The region bounded by inequalities (4) and (6) is the unhatched region in Fig. 7. From (4) and (6) we find the minimum spot radius R_m:

$$R_m = \left(\frac{f}{4}\right)^{1/4}\left(\frac{S_r qeN}{\pi C_{\mathrm{M}} BS}\right)^{3/8}. \tag{7}$$

Setting $C_{\mathrm{M}} = \$10$, $S_r = 1$ cm^2, $f = 5$ cm, $S = 10^{-10}$ cm^2, $B = 10^6$ A/cm^2·sr), $q = \$10$/hr, and $N_{\mathrm{M}} = 200$ (this value corresponds to an appearance probability $N = N_m/2$ of 10^{-14} in the case of a Poisson distribution), we find[14]

$$R_{\mathrm{m}} = 2 \cdot 10^{-6}\,\mathrm{cm}.$$

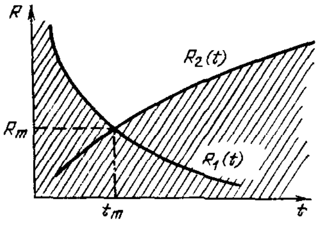

FIG. 7. Curves of $R_1(t)$ and $R_2(t)$ which describe the limits on the size of elements imposed by spherical aberration and the increasing cost of the exposure process, respectively. t—Exposure time; R_m—minimum element size; t_m—the exposure time corresponding to R_m. The unhatched region is the region of allowed dimensions of the electron-beam spot ($R > R_1, R_2$).

d) Effect of impurity fluctuations[20]

Fluctuations of the dopant lead to restrictions on the dimensions of the working regions of the elements of integrated circuits whose characteristics are sensitive to the doping level, particularly elements with p-n junctions.[20]

Fluctuations in the impurity concentration are unavoidable and are especially important when the size of the active region of the elements (d) and the average density of the dopant impurity (n) are low. We denote by $\bar{N}_i$ the average amount of impurity in the active region of an element ($\bar{N}_i = \bar{n}d^3$), and we deonte by ε_{M} the maximum tolerable relative deviation of the amount of impurity from its average value: $\varepsilon = (N_i - \bar{N}_i)/\bar{N}_i$. The probability that in a given cube of volume d^3 the value of ε will exceed ε_{M} is then, for a Gaussian impurity distribution,

$$P = 1 - \sqrt{\frac{2}{\pi}} \int_0^{\varepsilon_{\mathrm{M}}\bar{N}^{1/2}} e^{-y^2/2}\,dy. \tag{8}$$

The product of the probability P and the number of elements on a chip, $N = S/d^2$, determines the number of defective elements. If we require that this value be less than unity for a chip, we find the following restriction on the size of the elements:

$$\frac{d^2}{S} < 1 - \sqrt{\frac{2}{\pi}} \int_0^{\varepsilon_{\mathrm{M}}(\bar{n}d^3)^{1/2}} e^{-y^2/2}\,dy. \tag{9}$$

Substituting the typical values $\varepsilon_{\mathrm{M}} = 0.1$, $\bar{n} = 10^{19}$ cm^{-3}, and $S = 10^{-2}$ cm^2 into (9), we find $d > 10^{-5}$ cm.

e) Other types of restrictions in the technological process

There are two other restrictions which are important in practice. Elliot *et al.*[26] have shown experimentally that the width of the p-n junctions of the source and the drain of a field-effect transistor is determined by surface diffusion and has a lower limit of 0.1 μm.

If the structure of the elements is complex, the repeated positioning of the masks will cause an accumulation of errors, which will correspondingly limit the minimum dimensions of the elements.[23]

It follows from this discussion of the technological limitations on dimensions that

1) the restictions which arise in lithography are not important for devices of the immediate future, with dimensions greater than 10^2 μm, and

2) the fluctuations of the dopant and the surface diffusion restrict the dimensions of bipolar and field-effect transistors with p-n junctions to a value on the order of 10^{-1} μm. These restrictions can be overcome by (first) changing from devices with p-n junctions to other devices, in particular, devices with heterojunctions and (second) reducing the temperatures of the physical and chemical processes.[27]

3. PHYSICAL RESTRICTIONS IMPOSED ON THE DIMENSIONS OF ELEMENTS BY THEIR OPERATING MECHANISM

a) Classification of devices on the basis of dimensions

Elements can be classified on the basis of their geometric dimensions by comapring the dimensions of the working regions with the scale lengths which determine the operation of the device.[6] Among the scale lengths are the width L of the space-charge region, the carrier mean free path l, and the electron wavelength λ. In real devices these lengths are related by $L > l > \lambda$. Using these lengths, we can define four groups of devices:

1. Bulk devices with $d > L, l, \lambda$.
2. Devices of an intermediate group with $L \gtrsim d > l, \lambda$.
3. Ballistic devices with $l > d > \lambda$.
4. Quantum devices with $d \leqslant \lambda$.

The elements of the existing large-scale integrated circuits fall in the first of these groups. The elements of the very large-scale integrated circuits of the immediate future, in particular, field-effect transistors with submicron-size channels, fall in the second group or even the third.

In the devices of the first and second groups, the conductivity is described by introducing a mobility, as it is in bulk crystals.

Active research is also being carried out on the devices of the third and fourth groups, in which the transport of carriers is a quasiballistic[28-30] or tunnelling transport. Among such devices are submicron-size GaAs diodes,[31] tunnel diodes, and Josephson junctions.[32]

b) Field-effect transistor; scaling of parameters

1) *Operation of a field-effect transistor.* Figure 8 shows the standard geometry of an n-channel field-effect transistor with an insulated gate. This is a metal-insulator-semiconductor structure with a metal gate and a p-type semiconducting substrate, on whose surface two n^+-type regions are fabricated: the source and the drain. A voltage V_g is applied between the gate and the source, and a voltage V_D is applied between the drain and the source. In addition, a voltage V_{SS} may be applied between the source and the substrate. The operation of a field-effect transistor is based on a field effect: a modulation of the surface conductivity of the semiconductor between the source and the drain upon a change in the metal gate potential V_g. If there is no voltage between the gate and the semiconductor ($V_g = 0$), the current which can flow from the source to the drain is the reverse current of the p-n junction of the drain and is correspondingly low. As V_g is raised to a threshold value V_t, an n-type inversion layer forms on the surface of the semiconductor, and the source and the drain in the n^+-type region are connected by a conducting channel.

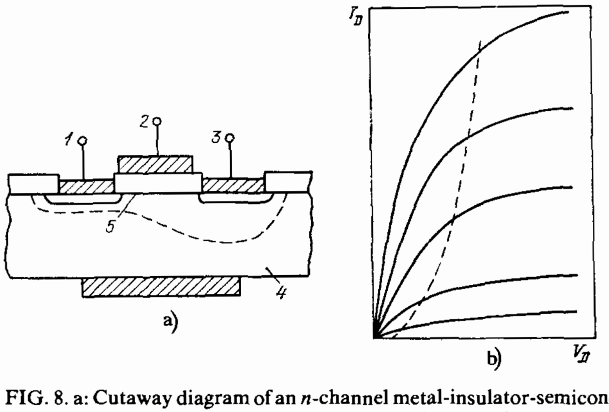

FIG. 8. a: Cutaway diagram of an n-channel metal-insulator-semiconductor field-effect transistor under working conditions. 1—Source; 2—gate; 3—drain; 4—substrate; 5—channel. b: Idealized output characteristics $J_D(V_D)$ of a metal-insulator-semiconductor transistor for various values of V_g (larger currents correspond to larger voltages V_g; the saturation region is to the right of the dashed curve).

The threshold voltage is determined from the requirement that the conductivity of the inversion channel be comparable to the bulk conductivity of the semiconductor:[33]

$$V_n = 2\Psi_B + \frac{Q_B}{C_1} = 2\Psi_B + \frac{2}{C_i}\sqrt{\varepsilon_S e N_a \Psi_B}\,; \qquad (10)$$

here Q_B is the charge in the depletion region of the semiconductor, N_a is the density of acceptors in the semiconductor, $e\Psi_B$ is the difference between the position of the Fermi level for the semi-conductor of the substrate and that of the intrinsic semiconductor, $C_1 = \varepsilon_i / a$ is the capacitance of the insulator, a is the thickness of the insulator, and ε_i and ε_s are the dielectric permittivities of the insulator and the semiconductor, respectively.

The voltage-current characteristic of the field-effect transistor under the condition $V_D < V_{DS}$ ($V_{DS} = V_g - 2\Psi_B$ is the saturation voltage for the characteristic) is described by

$$J_D = \frac{w}{d}\mu_n C_1 \left\{ \left(V_g - 2\Psi_B - \frac{V_D}{2} \right) V_D - \frac{2}{3C_1}\sqrt{2\varepsilon_S e N_a}\left[(V_D + 2\Psi_B)^{3/2} - (2\Psi_B)^{3/2} \right] \right\}, \qquad (11)$$

where w is the channel width, and μ_n is the electron mobility in a channel.

If the voltage between the source and the drain, V_D, is low, the resistance of a channel remains constant, and the channel current J_D is a linear function of V_D (Ref. 33). A further increase in V_D leads to a decrease in the depth of the channel near the drain, an increase in the resistance of the channel, and saturation of J_D (Fig. 6).

The voltage V_D at which the current reaches saturation corresponds to zero channel depth near the drain.

2) Changes in the parameters of devices as their dimensions are reduced: a scaling model for a field-effect transistor. A decrease in the geometric dimensions of a device entails changes in its working characteristics. Strictly speaking, a reduction in dimensions poses new problems of designing and optimizing the working parameters of the devices. There is, however, a rather broad range of geometric dimensions in which simple scaling relations can be used to estimate the changes in the parameters of a device when all its geometric dimensions are reduced by a factor of K and thus to assess the advantages of miniaturization.

Scaling means determining the scaling factors $F(K)$ by using which the parammeters of a device which has been reduced in size by a factor of K are expressed in terms of corresponding parameters of the initial device.

There are several scaling models, which differ both in the entity to which they are applied and in the naure of the quantities which are required by the theory to remain constant. For example, there is a scaling of field-effect transistors with constant electric fields in the semiconductor and in the insulator,[34] and there is a scaling of bipolar transistors with a constant voltage.[35,36]

Let us describe scaling at a constant electric field for field-effect transistors.[34] We assume that the dimensions of the transistor (the channel length d, the channel width w, and the oxide thickness a) are reduced by a factor of K. If the fields in the working regions of the transistor are to remain constant, all the voltages must be reduced by a factor of K, and the impurity density in the semiconductor must be increased by a factor of K. The latter requirement follows from the requirement that the field E_S in the space-charge region of the semiconductor must remain constant:

$$E_S \sim \sqrt{(V_g - E_i a)\, N_a},$$

where E_i is the field in the insulator.

According to (11), the current can be written as follows for gate voltags well above the threshold ($V_g \gg V_n$):

$$J_D = \frac{w}{d} \mu_n \frac{\varepsilon_1}{a} V_g V_D. \tag{12}$$

From (12) we see that J_D is inversely proportional to K.

Table II shows how the parameters of field-effect transistors and integrated circuits (the packing density and the power dissipation) vary with the scaling factor K.

A reduction in the dimensions of a device means that the dimensions of the interconnections and contacts will be reduced, and their characteristics will be changed (Table III). We see from this table that a decrease in the dimensions of the interconnections degrades the overall characteristics of integrated circuits.

TABLE II. Scaling of elements (changes in the parameters of metal-insulator-semiconductor field-effect transistors as the dimensions are reduced by a factor of K)

Parameter	Expression	Scaling factor
1. Geometric dimensions d, w, L, a		K^{-1}
2. Voltages V_D, V_g	$E = \text{const}, \quad V_D = E_D d$	K^{-1}
3. Impurity density in substrate, N_a	$E_S \sim \sqrt{(V_g - E_1 a)\, N_a} \approx \text{const}$	K
4. Current in linear region of I, V Char-ic, J_D	$J_D = \frac{w}{d} \frac{\varepsilon_S}{a} V_g V_D$	K^{-1}
5. Gate area S_g	$S_g \approx wd$	K^{-2}
6. Gate capacitance C_g	$C_g \approx \frac{\varepsilon_1 wd}{a}$	K^{-1}
7. Maximum density of elements on a chip	$\overline{N} = \frac{1}{S_g}$	K^2
8. Switching delay time	$\tau = \max\left(\frac{d^2}{\mu V_D};\ \frac{C_g V_g}{J_D}\right)$	K^{-1}
9. dc Joule power dissipation	$P = J_D V_D$	K^{-2}
10. Joule power dissipated on switching control	$P = \frac{1}{2} C_g \Delta V_g^2$	K^{-3}
11. Quality parameter of device, Q	$Q = P\tau$	K^{-3}
12. Speed of integrated circuit	$\text{Speed} = \overline{N}\tau^{-1}$	K^3

3) Reasons for failure of scaling. Formally, scaling tells us that we can reduce the dimensions of devices without limit. In reality, however, scaling begins to break down at some point. Let us analyze the reasons which force us to abandon the scaling relations in the miniaturization of devices. We stress that such problems actually arise in the devices of silicon technology only for elements of submicron size.

1. The operation of a field-effect transistor, i.e., the formation of an inversion channel, requires that the gate voltage exceed a threshold value V_t ($V_g > V_t$), which is essentially independent of K. Beginning at certain K values, V_g should thus remain constant, but this constancy is inconsistent with relation 2 in Table II.

2. The voltage between the drain and the source must be much larger than kT/e ($V_D > 10kT/e$). This condition is required to keep the thermal fluctuations of the current small in comparison with the current itself.

3. If the space-charge regions of the source and drain junctions are to be kept from reaching each other (if puncture is to be avoided), the length of the channel must be considerably larger than the thickness of the p-n junction ($d > 2L$).

4. The breakdown voltage of the drain-substrate p-n junction decreases sharply with decreasing channel length[37,38] (Fig. 9). This effect can be explained in the following way. The system consisting of the source, the substrate, and the drain constitutes an *n-p-n* transistor. Before its breakdown, the holes generated in the drain depletion region cause a current through the substrate. As a result, the potential of the substrate increases near the source, so that there is an increased injection of electrons from the p-n junction of the source into the substrate. The current through the substrate, which is a parasitic current for the operation of a

TABLE III. Scaling of interconnections: changes in the parameters of interconnections as their dimensions decrease

Parameter	Expressions, comments	Scaling factor $F(K)$
1. Resistance of interconnection lines (ICL)	$R_l=\rho_l\frac{l}{w_l\cdot h_l}$; l, w_l, h_l: length, width, and thickness of line	K
2. Normalized voltage drop across ICL	$J_D\cdot R_l/V_D$	K
3. Response time of ICL, $\tau l\ l$	$\tau_l=\begin{cases} R_lC_l, \\ \frac{l}{v_0} ; \end{cases}$ c_l (capacitance of connection), v_0 (velocity of an electromagnetic wave)	1
4. Current density in ICL, j_l	$j_l=\frac{J_D}{w_lh_l}$	K
5. Contact resistance R_k	$R_k \backsim w_lh_l$	K^2
6. Contact voltage drop V_k	$V_k \sim J_DR_k$	K
7. Normalized contact voltage drop	$\frac{V_K}{V_D}$	K^2
8. Normalized response time of ICL	$\frac{\tau_l}{\tau}\approx\frac{R_lC_l}{\tau}$	K

metal-insulator-semiconductor transistor, increases with decreasing channel length and leads to an anomalous decrease in the breakdown voltage.

5. When the channel length (d) becomes comparable to the width of the source and drain depletion regions, the threshold voltage V_t decreases (the short-channel effect).[39,40] The reason is that with decreasing channel length the fraction of the lines of force which begin at the charge in the channel and the depletion layer of the semiconductor and terminate not at the charge at the gate but at the charge of the drain and the source increases (Fig. 10). There is accordingly a decrease in the charge Q_B in the depletion region of the semiconductor, so that there is a decrease in V_t, according to (10). Calculations on this effect require a numerical solution of the two-dimensional Poisson equation. According to Ref. 36, however, a simple estimate of the short-channel effect reduces to replacing Q_B by the effective charge of the depletion region: $Q_B^* = fQ_B$. The coefficient is found from geometric consideration regarding the charge distribution to be

$$f=1-\left(\sqrt{1+\frac{2L}{X_j}}-1\right)\frac{X_j}{d},$$

where L is the width of the depletion region upon inversion, and X_j is the width of the region of the p-n junction of the source and drain.

We see from Fig. 11 that there is a good agreement between the theoretical and observed behavior. Figure 12 shows the functional dependence $V_t(d)$ predicted by this model.[40]

6. As the channel width w is reduced, there comes a point at which the threshold voltage V_t increases (the narrow-channel effect; Fig. 13).[41] This effect results from an effective increase in the charge in the depletion region of the semiconductor, Q_B. A simple geometric model for this effect was developed in Ref. 42. From the cross section of the metal-insulator-semiconductor transistor shown in Fig. 14 we can easily calculate the effective increase in Q_B and then, from (10), the change in the threshold voltage:

$$Q_B^*=-eN_awX_j\left(1+\frac{\pi X_j}{2w}\right)=Q_B\left(1+\frac{\pi X_j}{2w}\right).$$

Figure 15 (Ref. 43) shows the observed and predicted

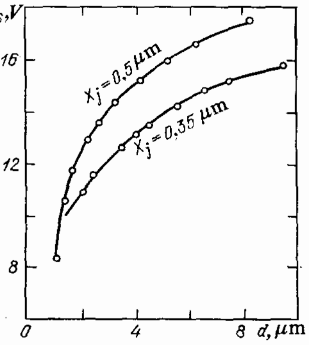

FIG. 9. Breakdown voltage of the source-substrate p-n junction, V_{DS}, versus the channel length d.

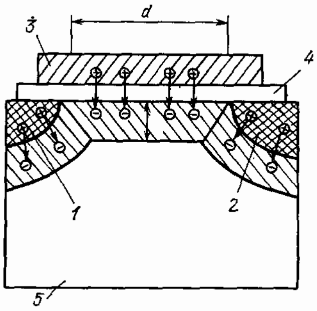

FIG. 10. Cutaway diagram of a metal-oxide-simiconductor field-effect transistor with a short channel.

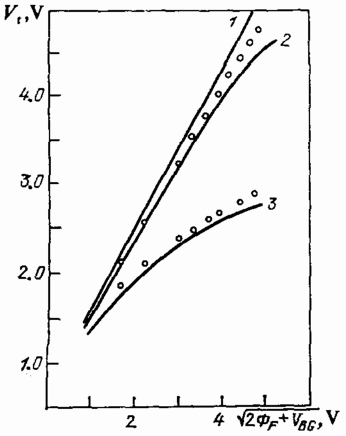

FIG. 11. Threshold voltage versus the potential of the semiconductor for a p-channel metal-oxide-semiconductor transistor with various channel lengths.[40] 1—$d \approx 7.4\ \mu$m; 2—$d \approx 3.8\ \mu$m; 3—$d = 1.4\ \mu$m.

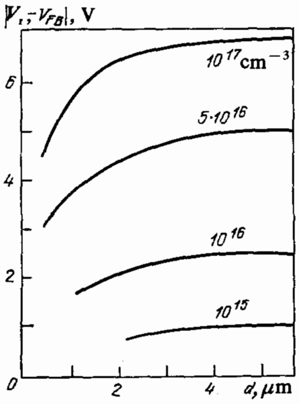

FIG. 12. Theoretical plot of the threshold voltage V_t versus the channel length d (Ref. 38). V_{FB} is the flat-band voltage. The thickness of the insulator is 500 Å; $S_j = 0.5\ \mu$m. The curves are labeled with the donor density in the substrate.

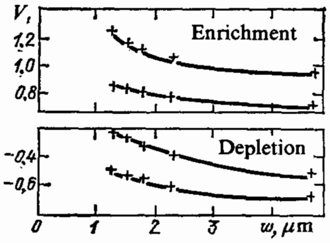

FIG. 13. The threshold voltage V_t versus the channel width w for an n-channel metal-insulator-semiconductor transistor with a channel length $d = 15\ \mu m$ (Ref. 37). V_{SS} is the source-substrate voltage.

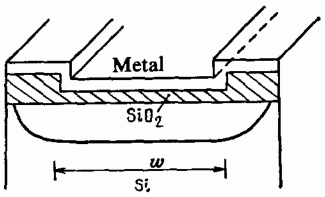

FIG. 14. Cutaway diagram of a narrow metal-insulator-semiconductor transistor showing the effective increase in the width of the depletion layer in the semiconductor.

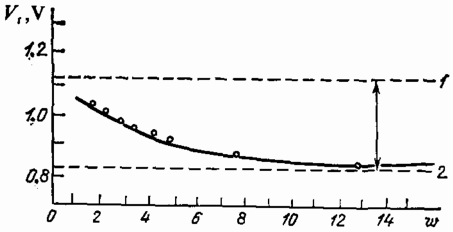

FIG. 15. Experimental and theoretical results on the threshold voltage V_t versus the channel width w for $V_D = 0.1$ V and $V_{sx} = 3$ V. 1—V_t for a long, wide channel with $d = 14.7\ \mu$m and $w = 89\ \mu$m; 2—V_t for a short and wide channel with $d = 3.3\ \mu$m, $w = 89\ \mu$m.

behavior of V_t as a function of both the length and width of the channel.

7. A decrease in the length of the channel of a metal-insulator-semiconductor transistor without a corresponding optimization of the other parameters sharply increases the current just before the threshold and its dependence on the voltage between the drain and the source.[44,45] The minimum channel length at which the current just before the threshold increases sharply is given by the empirical relation[44]

$$d_{\mathrm{m}} = c\,(X_{\mathrm{j}} a w_{\mathrm{l}}^2)^{1/3}, \qquad (13)$$

where c is a constant ($c = 0.41$ Å$^{-1/3}$), a is the thickness of the insulator, and w_1 is the total width of the depletion regions of the drain and the source, calculated from the model of an infinite, planar, sharp p-n junction.[33] It can be seen from this example that substantial progress toward miniaturization of devices can be achieved by optimizing the parameters. From (13) we find $d_{\mathrm{m}} \sim K^{-4/3}$, in comparison with the behavior $d_{\mathrm{m}} \sim K^{-1}$ predicted by scaling.

c) Restrictions on the dimensions of elements which are imposed by strong electric fields

1. A serious restriction on the length of the channel in a field-effect transistor results from a combination of requirements: 1) The field in the oxide of the field-effect transistor which is required for inversion of the channel must be below the breakdown field of the oxide, and 2) the space-charge regions of the p-n junctions of the drain and the source must not touch (puncture must be avoided).[46,47]

Let us estimate the corresponding limitations. We assume that puncture does not occur if the length of the channel is greater than three times the lengths of the p-n junctions of the drain and the source ($d > 3L$). Assuming that the barrier height of the p-n junctions of the drain and the source is of the order of $2e\Psi_{\mathrm{B}}$, we find

$$d \geqslant 3\sqrt{\frac{\varepsilon_{\mathrm{s}}\Psi_{\mathrm{B}}}{\pi e N_{\mathrm{a}}}}. \qquad (14)$$

We must therefore increase N_{a} in order to reduce d, but as N_{a} is increased there is an increase in the field in the insulator which is required for the formation of the n-type channel,

$$E_1 = \frac{\varepsilon_{\mathrm{s}}}{\varepsilon_1} E_{\mathrm{sn}} = \frac{4}{\varepsilon_1}\sqrt{\varepsilon_{\mathrm{s}}\pi \cdot e\psi_{\mathrm{B}} N_{\mathrm{a}}},$$

and this field must remain below the breakdown field in the oxide, $E_{\mathrm{c}} = 6\cdot10^6$ V/cm.

Setting $E_{\mathrm{i}} = mE_{\mathrm{c}}$, where $m \approx 1/4$, we find the maximum impurity density N_{m}; substituting the result,

$$N_{\mathrm{m}} = \frac{\varepsilon_1^2 m^2 E_{\mathrm{c}}^2}{16\pi\varepsilon_{\mathrm{s}}\psi_{\mathrm{B}}}, \tag{15}$$

into (14), we find the minimum length of the space-charge region of the p-n junction and thus the minimum channel length:

$$d_{\mathrm{m}} \approx \frac{12\varepsilon_{\mathrm{s}}\psi_{\mathrm{B}}}{\varepsilon_1 m E_{\mathrm{c}}}. \tag{16}$$

Substituting into (15) data for a silicon field-effect transistor, $\varepsilon_{\mathrm{s}}/\varepsilon_1 \approx 3$ and $\Psi_{\mathrm{B}} \approx 0.5$ V, we find

$$N_{\mathrm{m}} = 3\cdot 10^{17}\,\mathrm{cm}^{-3};\quad d_{\mathrm{m}} = 0.2\ \mu\mathrm{m}.$$

A more rigorous analysis[46,47] yields similar estimates. For normal operation of a field-effect transistor, the minimum insulator thickness a must exceed the electron tunneling length, which is 50 Å for silicon field-effect transistors.

A more severe limitation on a is imposed by the possibility of breakdown of the insulator[47]:

$$a > \frac{V_{\mathrm{g}}}{mE_{\mathrm{c}}}.$$

Assuming a working gate voltage $V_{\mathrm{g}} = 2$ V, we find $a > 130$ Å.

2. Electron heating leads to a limitation on the dimensions of the channel in a field-effect transistor. The effects of the heating intensify with decreasing channel length because the minimum voltage between the drain and source remains constant ($V_{\mathrm{Dm}} \approx 10kT/e$). In the case of energy relaxation by phonons, the electron gas is not heated if the electron drift velocity satisfies $v_{\mathrm{d}} < u$, where u is the sound velocity. This inequality can be rewritten as

$$\frac{\mu V_{\mathrm{D}}}{\mathrm{d}} = \frac{10\mu kT}{ed} < u,$$

or

$$d \gtrsim \frac{10\mu kT}{eu} \simeq d_{\mathrm{m}}. \tag{17}$$

If $d < d_{\mathrm{m}}$, there is accordingly a heating of the electron gas. This heating can lead to an avalanche breakdown, to the appearance of leakage currents across the p-n junctions, to a saturation of the drift velocity, and to an injection of hot electrons from the channel into the oxide, which will in turn destabilize the threshold voltage.

Let us quote some numerical values for the minimum channel length at which the electrons are not heated. Substituting $\mu = 10^2$ cm^2/(V·s), $T = 300$ K, and $u = 5\cdot 10^5$ cm/s into (17), we find $d_{\mathrm{m}} = 0.5\ \mu$m; with $\mu = 10^4$ cm^2/(V·s) and $T = 100$ K we would instead have $d_{\mathrm{m}} = 20\ \mu$m.

3. A limitation on the cross-sectional area of a channel arises from the saturation of the drift velocity.[20] A decrease in the dimensions of an element leads to an increase in the electric fields since V_{Dm} remains constant, so that the carrier drift velocity also increases. If the fields are sufficiently strong, the carrier drift velocity will reach saturation (if breakdown has not occurred previously), and this saturation will limit the current density.

If it is necessary to achieve a steep control characteristic and thus a high conductivity, a limitation is imposed on the minimum cross-sectional area of the channel of a field-effect transistor (or of the base of a bipolar transistor).

In a silicon field-effect transistor the drain current is $J_{\mathrm{D}} \leqslant ev_{\mathrm{sat}} n_{\mathrm{M}} w$, where v_{sat} is the saturation velocity, w is the channel width, and $n_{\mathrm{M}} \approx 10^{12}$ cm^{-2} is the maximum surface density of carriers, which is limited by the dielectric strength of SiO$_2$. It follows that in order to achieve a drain current $J_{\mathrm{D}} \gtrsim 1\ \mu$A at $v_{\mathrm{sat}} = 10^7$ cm/s we need a cross-sectional width of the channel satisfying $w > w_{\mathrm{m}} = 10^{-2}\ \mu$m.

4. To conclude this subsection we consider yet another substantial limitation on the dimensions of elements, but a limitation which is not due to the effects of strong electric fields. It arises from shot noise. If this noise is to be low, the number of carriers in the channel of a field-effect transistor (or in the base of a bipolar transistor) must be sufficiently high:

$$n_{\mathrm{M}} w d \gg 1.$$

Assuming $w \approx d$ and $n_{\mathrm{M}} = 10^{12}$, we then find

$$d_{\mathrm{m}} \approx 10 n_{\mathrm{M}}^{-1/2} \approx 10^{-5}\ \mathrm{cm}.$$

The same condition could be found from the requirement that the charge which flows through the channel of the field-effect transistor during the switching time must be much greater than the charge of one electron.

d) Limitations of the dimensions of memory elements

The limitations which we have discussed for field-effect transistors also apply to memory elements. In addition, memory elements have their own specific limitations.[25]

1. The charge on a memory element which represents information can be preserved for an adequately long time, not disappearing because of a tunneling transport, if the region in which the charge is captured is separated from other regions or elements by a sufficiently wide and high potential barrier. The minimum width of the barrier and thus the minimum size of the memory element are

$$d_{\mathrm{m}} \approx \frac{10}{\hbar}\, 2m\Delta\varepsilon,$$

where m is the mass of an electron, and $\Delta\varepsilon \geqslant 100kT$ is the barrier height. Assuming $m = 10^{-27}$ and $T = 300$ K, we find $d_{\mathrm{m}} \approx 0.01\ \mu$m.

2. A source of errors which proves important in practice in the operation of very large-scale memory integrated circuits is the inversion of bits by radiation.[48-50] The source of the radiation could be, for example, impurities of radioactive elements in the device. For example, the α particles emitted by uranium or thorium atoms have an energy of 9 ± 4 MeV. When such particles enter silicon, they generate more than $N_0 \approx 10^6$ electron-hole pairs in a volume $\rho^2 b \approx 1 \times 1 \times 10\,\mu\mathrm{m}^3$, where b is the mean free path of the α particle, and ρ is the diameter of the cylinder in which the pairs are produced. In a memory element with dimension d, the number of particles produced is therefore ($\rho < d < b$)

$$N_{\mathrm{g}} = N_0 \left(\frac{d}{b}\right).$$

The separation of electrons and holes in the internal or external electric fields leads to the appearance of a captured charge which may be perceived in sufficiently small elements as an amount of charge Q_i which is deliberately produced and which represents information.

If the radiation-induce charge Q_{g} is not to invert the memory state, this charge must be small in comparison with the information charge Q_i.

The maximum value of this charge can be found from the condition that the field in the barrier must be well below the breakdown field E_c:

$$Q_g = eN_g < Q_1 \approx \beta \frac{\varepsilon_1}{4\pi} E_c d^2,$$

where d is the size of the gate, and $\beta \ll 1$. With $E_c = 10^6$ V/cm, $\varepsilon_i = 4$, $\beta = 0.1$ and $b = 10^{-3}$ cm, we find

$$d_m \approx \frac{4\pi e N_g}{\beta \varepsilon_i E_c b} \approx 0.6 \ \mu\text{m} .$$

4. LIMITATIONS ON THE INTEGRATION OF ELEMENTS

In the planar technology the maximum degree of integration which can be achieved without causing interactions between elements with linear dimensions d, i.e., the maximum number of elements on a chip of area S, is

$$N_M = \frac{S}{d^2} .$$

The degree of integration N which is actually achievable is much lower than N_M. Limitations on integration arise from the interaction of elements with each other and are particularly important as the dimensions of the elements decrease. These limitations may, in particular, rule out any further miniaturization of elements.

I. The most serious physical limitation on the degree of integration, the working speed, and the minimum element size d_m is imposed by the heating of the chip.[25] We denote by N the number of elements on a chip of area S (the density of elements is $\bar{N} = N/S$), and we denote by ν the average frequency at which each of the elements is used, i.e., the cycling frequency. Upon each switching of an element, accompanied by a voltage change $\Delta V = n \cdot kT/e$ (we have already noted that this numerical factor is $n \geqslant 10$), an energy $C\Delta V^2$ is dissipated in the interconnections and in the elements themselves, where C is the capacitance of the element to which the voltage ΔV is applied. Accordingly, the heat dissipated per unit time in the integrated circuit is

$$N\nu \frac{C\Delta V^2}{2} .$$

If heat balance is to be preserved, this heat must be lower than the maximum amount of heat which can be removed from a chip of area S per unit time, $Q_M S$, where Q_M is the maximum amount of heat which can be removed from a unit area. We can thus write[25]

$$N\nu \frac{C\Delta V^2}{2} \ll Q_M S .$$

From this condition we find a restriction on the working speed:

$$N\nu < \frac{2Q_M S}{C\Delta V^2}, \tag{18}$$

The capacitance of the element is $C = \varepsilon d^2/4\pi a \approx \varepsilon^* d$, where a is the thickness of the dielectric, and the quantity $\varepsilon^* = \varepsilon d/4\pi a$ can be assumed constant.

We rewrite (18) in the following form ($\bar{N}\nu = N\nu/S$ is the working speed of a unit area of the chip):

$$\bar{N}\nu < \frac{2Q_M}{\varepsilon^* d n^2 (kT/e)^2} . \tag{19}$$

It can be seen from (18) and (19) that the speed is limited to a level which depends only on the size of the elements d, the heat removal Q_M, and the temperature T. Let us estimate the speed for elements with $d = 1\ \mu$m at $T = 300$ K with air and liquid cooling, $Q_M = 4 \times 10^6$ erg/(cm^2·s) and $Q_M = 2 \cdot 10^8$ erg/(cm^2·s), respectively. From (19) we find that the speed is less than 10^{18} Hz·element and less than 10^{20} Hz·element in the two cases, respectively.

Under certain assumptions, the limitation on the speed leads to a lower limit on the dimensions of the elements. The basic purpose of the miniaturization is to increase the number of elements and thus reduce the distance between them. The main purpose of increasing the speed of the elements is to increase the cycling frequency for the operation of the elements. Let us assume that the average distance between elements, a_e, is proportional to their dimensions d ($a = pd$) and that the cycling frequency ν is proportional to the speed of an element, as represented by τ: $\nu = (m\tau)^{-1}$, where τ is the carrier transit time through a channel, and m is the maximum fraction of the elements which operate simultaneously in a very large-scale integrated circuit on a chip. From (19) we then find a restriction on d $\left(\tau = \dfrac{d^2}{\mu V_D}\right.$ for $\mu V_D/d < \nu_{sat}$ and $\tau = d/v_{sat}$ for $v_{sat} \geqslant \mu V_D/d \Big)$:

$$d > \begin{cases} \left(\dfrac{\varepsilon^* \mu \Delta V^3}{2mp^2 Q_M}\right)^{1/3} & \text{for} \quad \dfrac{\mu V}{d} < v_{sat}, \\ \left(\dfrac{\varepsilon^* \Delta V^2 v_{sat}}{2mp^2 Q_M}\right)^{1/2} & \text{for} \quad \dfrac{\mu V}{d} > v_{sat}. \end{cases} \tag{20}$$

In (20), $v_{sat} = 10^7$ cm/s is the saturation carrier velocity.

Setting $\varepsilon^* = 10$, $m = 10^2$, $p = 10$, $\Delta V = 0.3$ eV ($\Delta V \gtrsim 10kT/e$), $T = 300$ K, $Q_M = 2 \cdot 10^8$ egr/(cm^2·s) (liquid cooling), and $\mu = 3 \times 10^4$ esu, we find $d \geqslant 10^{-5}$ cm and $a > 10^{-4}$ cm.

Since the cycling frequency of the operation of memory elements is considerably lower, this limitation imposed by the heating of the chip is considerably less stringent. With $m = 10^4$–10^5, for example, we find $d > 10^{-6}$ cm.

II. We conclude with a list of some other possible factors which could limit the degree of integration:

a) the parasitic coupling between elements, which could destabilize the operation of very large-scale integrated circuits and which could result from the injection of (hot) carriers, tunneling, etc., capacitive coupling, ballistic phonons, plasmons, and other elementary excitations;

b) the overheating of interconnections and their malfunction;

c) a parasitic coupling between connections.

5. CONCLUSION

Let us summarize the conclusions which emerge from this analysis of the basic physical limitations on the miniaturization of the elements of very large-scale integrated circuits.

The limitations in lithography which result from the scattering of the beam in the resist and in the semiconductor set a minimum width $\approx 0.01\ \mu$m on the technological line. This limitation will not become crucial for the development of microelectronics over the next 10 yr.

Fluctuations in the dopant density, the surface diffusion of the dopant, and the effects due to electrical puncture and breakdown of the p-n junction and the insulator restrict the dimensions of bipolar transistors and field-effect transistors with p-n junctions to values greater than 0.1 μm. Heterojunctions presently look promising for fabricating devices with smaller dimensions.

The Joule heating due to a limited heat removal leads to a limitation on the working speed of elements ($\approx 10^{20}$ Hz·element) and to a limitation on the dimensions of elements (≈ 0.1 μm). Since memory elements are used less frequently, this limitation will be less important in their case than for logic elements. This limitation can be relaxed by working at lower temperatures.

On the other hand, the dimensions of memory elements (≈ 0.1–1 μm) are limited more severely by the radiation-induced inversion of bits. This limitation can be overcome by, for example, shielding against radiation.

Finally, since new conduction mechanisms—the quasi-ballistic and tunneling transport of charge carriers—arise in structures with dimensions less than 0.1 μm, there is the problem of developing new types of ultraminiature elements: ballistic and tunnel transistors.

This physical analysis is ultimately governed by the requirements of the modern technology of microelectronics. It is our deep conviction that the problems of physical limitations in microelectronics are of special interest to physicists, more and more of whom are involved in the development of microelectronics in one way or another. This interest will unavoidably intensify in the future, since technological developments will lead to decreases in the sizes of elements to the point that they approach a number of the microscopic scale lengths. It is thus obvious that further developments in microelectronics will require a constant analysis of the operating principles of elements and the use of progressively more fundamental physical approaches. The clearest evidence of this circumstance even now is the intimate relationship between the problems of the physics of a two-dimensional electron gas at a solid interface and the problems of miniaturization.

NOTATION

N—number of elements on a chip
$\bar{N}$—density of elements
τ—switching time
ν—cycling frequency
Speed—working speed
P—power dissipated during a switching
$\Pi = P\tau$—quality index of an element
I—intensity of the electron beam
$N_{\rm exp}$—number of electrons which reach the target
q—cost of operating the exposing apparatus for a unit time
$\bar{n}$—average impurity density
$N_0 = \bar{n}d^3$—average amount of impurities
$\varepsilon_{\rm M} = \dfrac{W_{\rm M} - N_0}{N_0}$—maximum permissible relative amount of impurities
P—probability
l—mean free path
L—thickness of space-charge region
$V_{\rm g}$—gate voltage
$V_{\rm t}$—threshold gate voltage
$V_{\rm D}$—source-drain voltage
d—scale dimension of the active region of an element
ΔX—bluring of the illuminated region
λ—wavelength
ε—electron energy
R—radius
B—brightness
S—area of emitting surface
$N_{\rm a}$—acceptor density
$\Psi_{\rm B}$—difference between the positions of the Fermi levels of the semiconductor of the substrate and of the intrinsic semiconductor
$\varepsilon_i, \varepsilon_s$—dielectric constants of the insulator and the semiconductor
w—channel width
a—thickness of insulator
X_j—width of the p-n junction of the drain and source
$v_{\rm sat}$—saturation velocity
$\Delta\varepsilon$—barrier height
m—electron mass
$\hbar$—Planck's constant
ρ—radius of the cylinder in which electrons are produced
α—α-particle
$Q_{\rm M}$—maximum heat which can be removed from a unit surface area

[1] P. H. Abelson and A. L. Hammond, Science **195**, 1087 (1977).
[2] J. R. Pierce, Science **195**, 1092 (1977).
[3] J. G. Linvill and C. L. Hogan, Science **195**, 1107 (1977).
[4] R. N. Noyce, Science **195**, 1102 (1977).
[5] P. V. Nesterov, Zarub. radioélektron. No. 12, 3 (1980).
[6] J. R. Barker, in: International Conference on New Trends in Integrated Circuits, 7–10 April 1981, Paris, p. 15.
[7] K. A. Valiev, Mikroélektronika **9**, No. 6, 483 (1980).
[8] G. E. Moore, in: Proceedings of the International Electron Devices Meeting, New York, 1975, p. 11.
[9] G. E. Moore, IEEE Spectrum **16**, No. 4, 30 (1979).
[10] L. W. Sumney, IEEE Spectrum **17**, No. 4, 24 (1980).
[11] R. Connely, Élektronika No. 10, 7 (1981).
[12] W. Larry, Élektronika No. 10, 71 (1981).
[13] V. F. Dorfman, Mikrometallurgiya v mikroélektronike (Micrometallurgy in Microelectronics), Metallurgiya, Moscow, 1978.
[14] A. N. Broers, IEEE Trans. Electron Devices **ED-28**, 1268 (1981).
[15] A. N. Broers *et al.*, Appl. Phys. Lett. **29**, 596 (1976).
[16] R. E. Howard, Solid State Tech. 127 (1980).
[17] N. G. Panish and A. J. Cho, IEEE Spectrum **17**, No. 4, 18 (1980).
[18] V. N. Derkach and M. S. Kukharchik, Mikroélektronika **9**, 498 (1980).
[19] S. Leisegang, Electron Microscopy [Russ. Transl. IL, Moscow, (1960)].
[20] T. Sugano, Suppl. Jpn. J. Appl. Phys. **15**, 329 (1976).
[21] G. R. Brewer, IEEE Spectrum **8**, No. 1, 23 (1971).
[22] I. N. Rubtsov, in: Fotolitografiya i optika (Photography in Optics), Sov. radio, Moscow; Tekhnika, Berlin, 1975, p. 341.
[23] J. T. Wallmark, in: Proceedings of the International Summer School, Szegeel, July 1–6, 1979; Lect. Notes Phys. 122 (1980).
[24] A. Yoshikawa *et al.*, Jpn. J. Appl. Phys. **20**, No. 2, L-81 (1981).
[25] R. W. Keyes, Science **195**, 1230 (1977).
[26] Elliot *et al.*, IEEE Trans. Electron Devices **ED-26**, 469 (1979).
[27] Yu. D. Chistyakov, Mikroélektronika **9**, 541 (1980).
[28] M. S. Shur and L. F. Eastman, IEEE Trans. Electron Devices **ED-26**, 1677 (1979).

[29]V. I. Ryzhiĭ and V. A. Fedirko, Pis'ma Zh. Tekh. Fiz. 7, 18, 1121 (1981) [Sov. Tech. Phys. Lett. 7, 8, 480 (1981)].

[30]A. A. Sukhanov, V. B. Sandomirskiĭ, and Y. Ya. Tkach, Fiz. Tekh. Poluprovodn. **17**, 2156 (1983) [Sov. Phys. Semicond. **17**, 1378 (1983)].

[31]K. Hess, IEEE Trans. Electron Devices **ED-28**, 937 (1981).

[32]B. T. Bosch, Proc. IEEE **67**, (1979).

[33]S. M. Zi, Fizika poluprovodnikovykh priborov (Physics of Semiconductor Devices), Énergiya, 1973.

[34]R. H. Dennard *et al.*, IEEE J. Solid State Circuits **SC-9**, 250 (1974).

[35]F. M. Klassen, Solid State Electron. **21**, 56 (1978).

[36]J. L. Prince, in: Springer Series in Electrophysics (ed. D. F. Barbez), Vol. 5, Springer-Verlag, N. Y., 1980, p. 4.

[37]E. Sun *et al.*, International Electronics Devices Meeting: Technical Digest, Washington, 1978, p. 478.

[38]T. Toyabe *et al.*, IEEE Trans. Electron Devices **ED-25**, 825 (1978).

[39]H. C. Poon *et al.*, in: International Electron. Device Meeting, Technical Digest, Tokyo, 1973, p. 156.

[40]L. D. Yau, Solid State Electron. **17**, 1059 (1974).

[41]Dennard *et al.*, IEEE Trans. Electron. Devices **ED-26**, 325 (1979).

[42]G. Merkel, in: Processing and Device Modeling for Integrated Circuits Design, Nordhoff, Leiden, Netherlands, 1977, p. 705.

[43]P. P. Wang, IEEE Trans. Electron Devices **ED-25**, 779 (1978).

[44]J. R. Brews *et al.*, IEEE Electron Devices Lett. **EDL-1**, 2 (1980).

[45]G. W. Taylor, IEEE Trans. Electron Devices **ED-25**, 337 (1976).

[46]B. Hoenesen and C. A. Mead, Solid State Electron. **15**, 819 (1972).

[47]B. Hoenesen and C. A. Mead, Solid State Electron. **15**, 891 (1972).

[48]J. T. Wallmark and S. M. Marcus, Proc. IRE **50**, 286 (1962).

[49]T. H. May and M. H. Woods, Proceedings of the Sixteenth IEEE Reliability Plugs Symposium, 1978, p. 33.

[50]J. F. Ziegler and W. A. Lanford, in: Digest of Technical Papers of the International Solid State Circuits Conference, February, 1980, New York, p. 70.

Translated by Dave Parsons