1 Introduction

The history of electronics spans over more than a century. A key milestone in the history of electronics was the invention of the telephone in 1876 and patents for the device were filed independently by Elisha Gray and Alexander Graham Bell on 14 February that same year. Bell filed first, and thus the patent was granted to him. This timely, or untimely for Gray, coincidence has become a textbook example for teaching the importance of intellectual property law in engineering schools across the globe.

Years later, the first radio broadcast took place in 1910 and is credited to the De Forest Radio Laboratory, New York. Lee De Forest, inventor of the electron vacuum tube, arranged the world's first radio broadcast featuring legendary tenor Enrico Caruso along with other stars of the New York Metropolitan Opera to several receiving locations within the city. Experimental television broadcasts can be traced back to 1928, but practical TV sets and regular broadcasts date back to shortly after the Second World War.

During this initial phase of development, electronics was based on vacuum tubes and electromechanical devices. The first transistor was invented at Bell Labs by William Shockley, John Bardeen, and Walter Brattain in 1947 and they used a structure named a point-contact transistor. Two gold contacts acted as emitter and collector contacts on a piece of germanium. William Shockley made and patented the first bipolar junction transistor in the following year, 1948. It is worth noting that the point-contact transistor was independently invented by German physicists Herbert Mataré and Heinrich Welker of the Compagnie des Freins et Signaux, a Westinghouse subsidiary located in Paris [1].

The first patent for a metal-oxide-semiconductor field-effect transistor (MOSFET) was filed by Julius Edgar Lilienfeld in Canada and in the USA during 1925 and 1928, respectively [2,3]. The semiconductor material used in the patent was copper sulfide and the gate insulator was alumina. However, a working device was never successfully fabricated or published at that time. The first functional MOSFET was made by Dawon Kang and John Atalla in 1959 and patented later in 1963 [4]. The successful field-effect operation was enabled by the use of silicon and silicon dioxide for the metal-oxide-semiconductor (MOS) stack. Unlike other insulator–semiconductor structures of the time, the Si–SiO_2 interface could be formed without a large density of electrically active defects that would otherwise prevent the penetration of the electric field from the gate into the semiconductor. Even when defects were present, means of deactivating them by chemical and other means, known as passivation, were found.

Because of practical fabrication reasons, *p*-channel (pMOS) technology was developed first and relied on aluminum as the metal for the gate electrode. Later on, the advent

of ion implantation and the use of polysilicon (heavily doped polycrystalline silicon) as gate material made self-aligned *n*-channel (nMOS) transistors feasible [5]. In a 1963 paper presented at the IEEE International Solid-State Circuits Conference, C. T. Sah and Frank Wanlass showed that *p*-channel and *n*-channel MOS transistors could be integrated onto a single integrated circuit or "chip" forming a circuit configuration with complementary symmetry [6]. This technology had the great advantage of drawing close to zero power in standby mode. It was initially called COS-MOS (complementary symmetry metal-oxide-semiconductor) and has since been universally adopted by the semiconductor industry under the name complementary metal-oxide-semiconductor (CMOS).

Another great advantage of MOS transistors is that they, unlike bipolar transistors, have a planar, basically two-dimensional structure. MOS transistors occupy only a small portion of the volume of a silicon wafer on which they are manufactured. The devices are located at the top surface of the wafer and extend into the wafer to a depth of only a fraction of a micrometer. As a consequence, the MOSFET is scalable, and scaled it has been for the last 50 years, giving rise to the microelectronics revolution at the end of the twentieth century and through to the beginning of the twenty-first.

1.1 Moore's law

The MOSFET is the workhorse of the electronics industry. It is the building block of every microprocessor, every memory chip, and every telecommunications circuit. A modern microprocessor contains several billion MOSFETs and a 256 gigabyte micro secure digital (SD) memory card weighing less than a gram contains a staggering 1,000,000,000,000 or 10^{12} transistors, assuming 2 bits stored per transistor. This number is larger than the number of stars in our galaxy, as there is an estimated 200–400 billion stars in the Milky Way. Although it can be used for other purposes, the MOSFET is mainly used as a switch in logic circuits and a charge-storage device in memory chips. Each day the semiconductor industry produces more MOSFETs than the number of grains of rice that have been harvested by mankind since the dawn of time. That number, astronomical as it is, is dwarfed by the rate at which transistors are increasingly packed on a chip. The exponential growth of chip complexity and number of transistors per chip is known as Moore's law.

In 1965, Gordon Moore published what was to become a classic paper in which he predicted that the density of transistors on a chip would double every 18 months [7]. This prediction was based on data spanning only a few technology generations produced during the period from 1959 to 1965, during which the number of transistors per chip increased from a single transistor to less than a hundred transistors. Extrapolating from the available data, Gordon Moore predicted that there would be 64,000 transistors per chip in 1975, ten years after the publication of the article. Even though completely an empirical observation, Moore's law has proven to be remarkably accurate, not only until 1975 but continues at present and covers a period of over 50 years. Whether plotted in terms of transistors per chip or transistors per square millimeter (Figs. 1.1 and 1.2), the

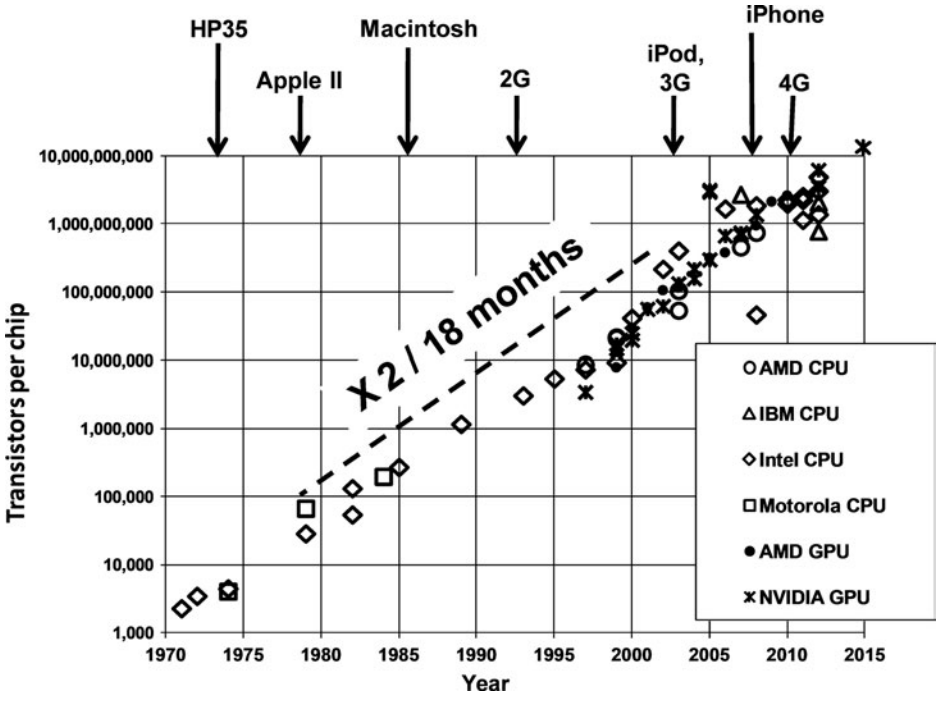

Figure 1.1 Evolution of the number of transistors per chip with time. Central processing units (CPU) or microprocessors and graphics processing units (GPU) or graphics processors from different vendors are shown. The top of the chart shows the date of introduction of some landmark products: HP-35 pocket calculator, Apple II and Macintosh computers, iPod, iPhone, and the introduction of second-, third-, and fourth-generation mobile phone networks (2G, 3G, 4G).

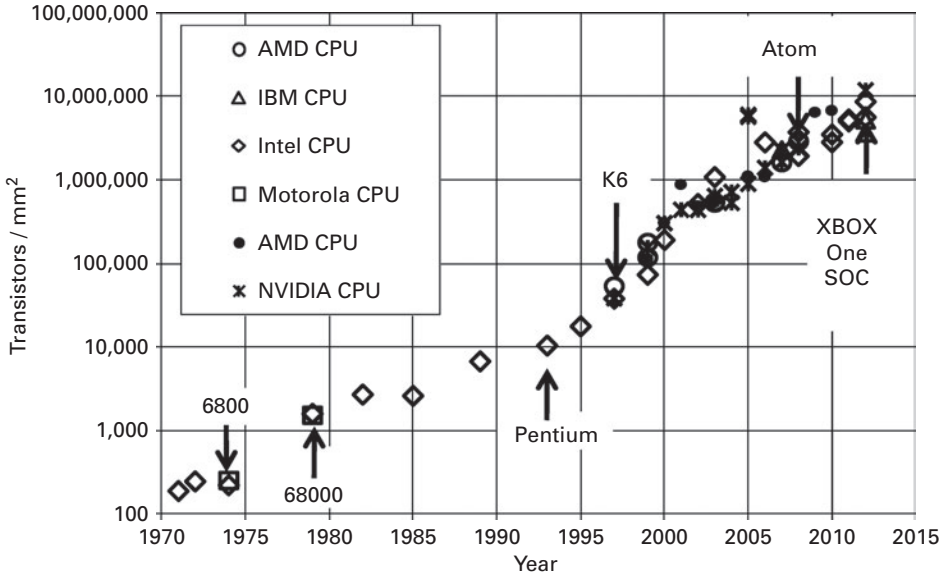

Figure 1.2 Evolution of the number of transistors per square millimeter with time. Microprocessors (CPU) and graphics processors (GPU) from different vendors are shown. Some landmark microprocessors are outlined for reference: Motorola's 6800 and 68000, Intel's Pentium and Atom, and AMD's K6 and XBOX One SOC (system on chip).

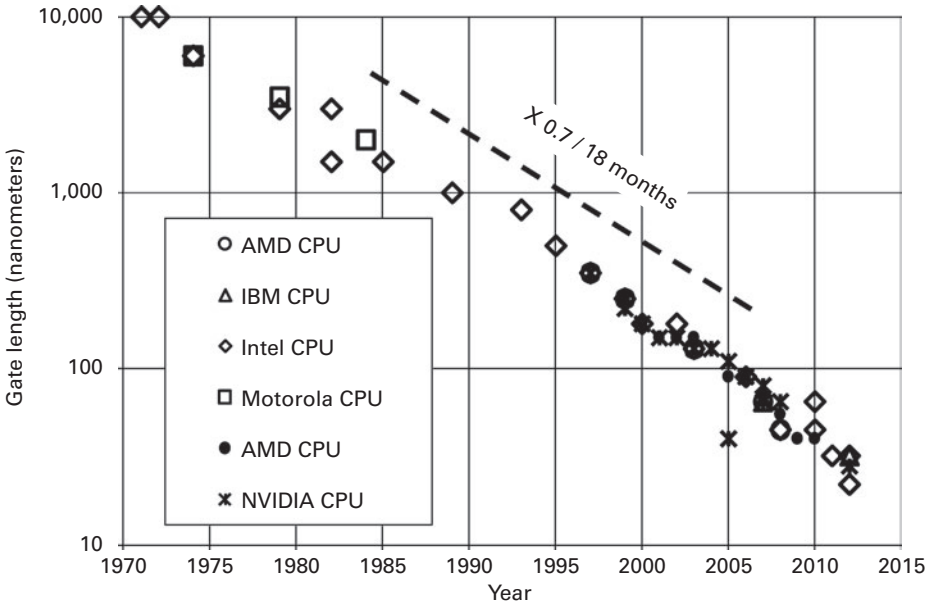

Figure 1.3 Evolution of the gate length with time. Gate length is the smallest printed feature in a MOS transistor, at least for traditional planar MOSFETs.

increases in the number of transistors and their density are spectacular. It is now part of popular legend that Bill Gates once joked that "If the car industry had kept up with technology like the computer industry has, we would all be driving 25-dollar cars that can run 1,000 miles to the gallon." He might have added that such a car would go around the world in a few seconds while carrying a million passengers.

It is quite obvious that reducing the size of transistors increases their density on a chip, which, for a constant chip size, increases the functionality of the circuits. There are other incentives for making the transistors smaller. Doubling the density of transistors on a chip implies reducing the linear dimensions, such as their length and width, by a scaling factor equal to $\sqrt{2}$. The gate length of MOS transistors has been steadily decreasing over the years, as shown in Fig. 1.3 where the data are plotted for the same circuits as for Figs. 1.1 and 1.2. One can clearly see that the linear dimensions of the patterns of a chip, such as the gate length, have been steadily decreasing by a factor of approximately $1/\sqrt{2} \cong 0.7$ every 18 months. Decreasing linear dimensions by 0.7 results in the surface area of the transistors halving every 18 months, in agreement with Moore's prediction.

1.2 The MOS transistor

The textbook example of a MOSFET is shown in Fig. 1.4. The device consists of a *p*-type semiconductor substrate in which two *n*-type regions have been formed. These *n*-type regions are called the "source" and the "drain." Typically the semiconductor

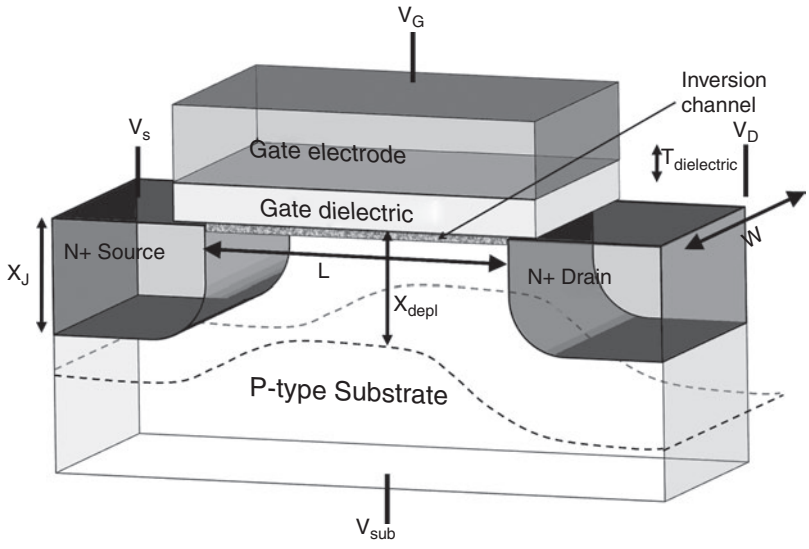

Figure 1.4 Schematic view of a classical bulk MOSFET.

material is silicon, although other semiconductors such as germanium (Ge), silicon germanium alloys (SiGe), indium arsenide (InAs), and indium gallium arsenide (InGaAs) can also be used. A thin layer of insulating material called the "gate dielectric" covers the region between the source and drain. For many years silicon dioxide (SiO_2) was the standard dielectric, but in recent years, silicon oxynitride (SiON) and stacks composed of insulators with high dielectric constant known as "high-κ dielectrics" have become common. An example of high-κ dielectric material is HfO_2 which has a dielectric constant approximately five times higher than SiO_2. The gate dielectric is formed by deposition and subsequently topped by a metal electrode called the "gate."

Under typical bias conditions, the source and the *p*-type substrate are grounded ($V_S = V_{sub} = 0$ V), and a positive voltage, V_D, is applied to the drain. Under these conditions, the drain *pn* junction is reverse biased and no current flows between the drain and the substrate. Since the bias across the source *pn* junction is zero, there is no current flowing from the substrate to the source either. As a result, there is no current flow between the source and the drain, and the transistor is turned OFF, playing the role of an open switch. If a positive voltage is applied to the gate, holes in the *p*-type substrate underneath the gate are pushed away from the surface and a region void of holes, called the "depletion region" forms beneath the gate. The depth of the depletion region, X_{depl}, increases with gate voltage up to a maximum value which depends on the *p*-type doping concentration. It is worth noting that the gate-induced depletion region merges with the source and drain junction depletion regions on the source side and drain side of the gate. If the gate voltage is further increased, further increments of gate-induced charge are not picked up by increasing the depletion depth, but rather by attracting electrons underneath the gate dielectric. Electrons literally "spill out" from the *n*-type source to form an electron-rich layer underneath the gate insulator called an "inversion channel." The term

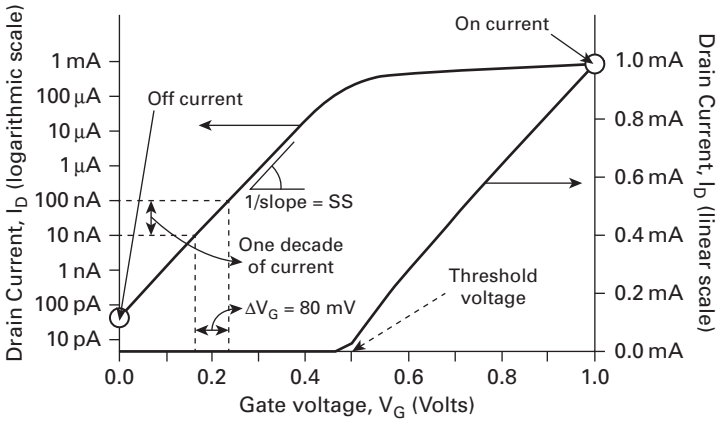

Figure 1.5 Drain current as a function of gate voltage in an MOS transistor at low drain bias. The two curves represent identical data, plotted using either a linear scale (right-hand *y* axis) or a logarithmic scale (left-hand *y* axis).

"inversion" is used because the top surface of the semiconductor, originally *p*-type (rich in holes), is now void of holes and rich in electrons, which technically makes it locally *n*-type. The silicon surface has thus been "inverted" from *p*-type to *n*-type. The inversion channel forms a continuous electron bridge between the source and drain and current can now flow between these two terminals. The transistor is considered to be in the ON state and behaves as a closed switch.

A perfect switch features zero current flow when it is open, zero resistance when it is closed, and is capable of switching sharply between the OFF state and the ON state. The MOSFET is unfortunately an imperfect switch; the OFF current is not zero and the ON-state resistance is finite. Furthermore, switching does not suddenly occur at a precise value of the gate voltage, but it takes place gradually, over a range of gate voltage values. Figure 1.5 illustrates how the drain current flowing through a MOSFET evolves as a function of gate voltage with a fixed positive drain voltage of 50 mV. In this example, the ON current is 1 mA and the OFF current is 50 pA. Looking at the current plotted on a linear scale, it appears there is no current below a given gate voltage, called the "threshold voltage" which is approximately equal to 0.5 V in the example shown in Fig. 1.5. If the drain voltage is low (typically 50 mV), the drain current basically increases linearly with the applied gate bias above threshold. The classical textbook expression for this current, called the "linear" or "non-saturation" current, is [8]

$$I_{D(\text{lin})} = \mu C_{\text{ox}} \frac{W}{L}\left[(V_G - V_{TH})V_D - \frac{1}{2}V_D\right], \tag{1.1}$$

where μ, C_{ox}, L, W, V_G, V_{TH}, and V_D are the carrier mobility in the channel ($\text{m}^2\ \text{V}^{-1}\ \text{s}^{-1}$), the gate capacitance (F m^{-2}), the gate length (m), the gate width (m), the gate voltage (V), the threshold voltage (V), and the drain voltage (V), respectively. The source and the substrate are assumed to be grounded.

For larger values of the drain voltage (when $V_D > V_G - V_{TH}$), the channel is pinched off near the drain due to the increase of the depletion region with increasing drain voltage and the drain current saturates (i.e. it no longer increases with increasing drain voltage V_D). In that case, the "saturation" drain current is given by

$$I_{Dsat} = \frac{1}{2}\mu C_{ox}\frac{W}{L}(V_G - V_{TH})^2. \tag{1.2}$$

Plotting the drain current on a logarithmic scale reveals that the drain current varies exponentially with gate voltage below threshold, and that the OFF current is not equal to zero. The rate of increase of current below threshold is characterized by a parameter called the "subthreshold slope," also called subthreshold swing (SS), defined by the relationship $SS = dV_G/d(\log(I_D))$ where the logarithm is chosen to be base 10. The subthreshold slope is expressed in units of millivolts per decade. A typical value for the subthreshold slope of a bulk MOSFET is 80 mV/dec, which means that an 80 mV increase of the gate voltage brings about a tenfold increase of drain current. Thus, in order to "switch" the current from its OFF value (50 pA) to the ON state ($I_D = 100$ μA at threshold), a gate voltage swing of 80 mV × log[100 μA/50 pA] = 0.5 V is required.

It can be shown that the subthreshold slope is equal to:

$$SS = n\frac{k_B T}{|q|}\ln(10) = n \times 59.6 \times \frac{T}{300\ \text{K}}\ \text{mV/dec}, \tag{1.3}$$

where k_B is Boltzmann's constant, T is the temperature, q is the charge of an electron (taken in absolute value, since the charge of an electron is negative by convention), ln(10) is the natural logarithm of 10, and n is the "body factor." The body factor represents the efficiency, or rather the inefficiency with which the gate voltage electrostatically controls the channel region. The body factor is proportional to the change in gate voltage with a change in channel potential (Φ_{CH}) and is expressed mathematically through the relationship $n = dV_G/d\Phi_{CH}$. In the best possible case, if the electrostatic coupling between the gate and the channel region is 100% effective, $n = 1$ and the subthreshold slope is equal to $[k_B T/|q|] \times ln(10) =$ 59.6 mV/dec at room temperature ($T = 300$ K = 26.85°C). In practice, the gate control of the channel region is not perfect due to the electrostatic coupling between the substrate through the depletion layer. As a result, n typically has a value between 1.2 and 1.5 in bulk MOSFETs, which results in subthreshold slope values ranging from 70 to 90 mV/dec. It is impossible, as can be shown from thermodynamics arguments, to reduce the subthreshold slope below 59.6 mV/dec at room temperature in classical MOSFETs; the best one can hope for is to approach that limit as closely as possible. The 59.6 mV/dec barrier can be breached using impact ionization effects [9,10], quantum tunneling effects [11,12,13], and with special ferroelectric gate materials [14], but none of these techniques have yet been proven to be reliable or reproducible enough for industrial applications. The lack of scalability for the subthreshold slope is a fundamental limit for the MOSFET and is sometimes referred to as the "Boltzmann tyranny" [15,16].

Table 1.1 Constant-electric field scaling rules for planar MOS transistors [17].

Parameter	Equation	Unit	Scaling factor
Physical dimensions: L, W, X_j, X_{depl}		m	γ^{-1}
Integration density	$\frac{1}{WL}$	m^{-2}	γ^2
Equivalent oxide thickness (EOT)	$t_{ox} = t_{dielectric}\frac{\varepsilon_{SiO_2}}{\varepsilon_{dielectric}}$	m	γ^{-1}
Dielectric capacitance	$C_{ox} = \frac{\varepsilon_{SiO_2}}{t_{ox}}$	F/m^2	γ
Gate capacitance	$C_G = WLC_{ox}$	F	γ^{-1}
Voltages V_{DS}, V_{GS}, V_{TH}	Electric field $E = V/L$ = constant	V/m	γ^{-1}
Drain current	$I_{Dsat} = \frac{1}{2}\frac{W}{L}\mu C_{ox}(V_{GS} - V_{TH})^2$	A	γ^{-1}
Power density	$\frac{V_{DS}I_{Dsat}}{WL}$	W/m^2	$\gamma^0 = 1$
Power consumption per transistor	$P = V_{DS}I_{Dsat}$	W	γ^{-2}
Intrinsic gate delay	$\tau = \frac{C_G V_{DS}}{I_{Dsat}}$	S	γ^{-1}
Power × delay product	$P \times \tau$	J	γ^{-3}

1.3 Classical scaling laws

In 1974, Robert Dennard and co-workers published a seminal paper in which they demonstrated the benefits of scaling [17]. Based on the assumption of maintaining a constant electric field inside the transistor, Dennard *et al.* demonstrated that scaling the device by a factor γ increases the switching speed by a factor γ, reduces the transistor power dissipation by a factor γ^2, and improves the power-delay product by a factor γ^3. It is worthwhile noting that this scaling law implies reducing the supply voltage by a factor γ, as well as reducing the threshold voltage by the same factor γ. The latter has not been achieved in subsequent technologies because of the impossibility of scaling the subthreshold slope to achieve values lower than 59.6 mV/decade because of fundamental thermodynamic reasons. Dennard's scaling law was more or less followed by the semiconductor industry for a duration of approximately 30 years, familiarly called the "happy scaling" period. These years are now over, and the improvement of performance due to scaling, at least in terms of microprocessor clock frequency, has reached saturation. This is caused by so-called "short-channel effects" that arise when the distance separating source from drain becomes very small. Short-channel effects increase as devices are scaled down in length, as will be described in the following. The classical scalling laws are shown in Table 1.1.

1.4 Short-channel effects

Short-channel effects result from the sharing of the electrical charges in the channel region between the gate on one hand, and the source and drain on the other hand. The source and

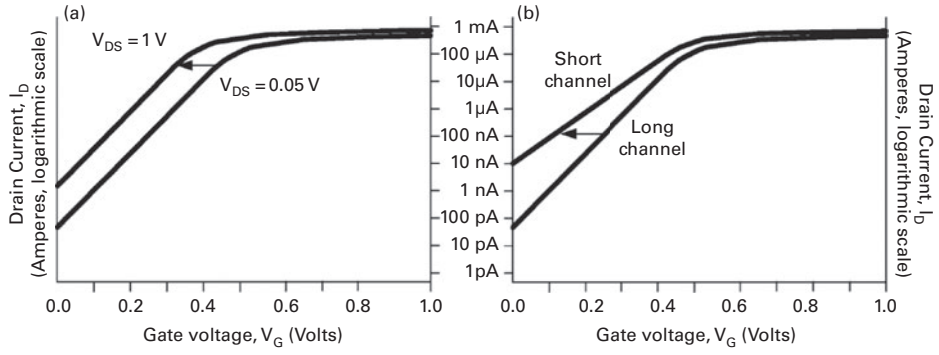

Figure 1.6 (a) The drain-induced barrier lowering (DIBL) effect decreases the threshold voltage when the drain voltage V_{DS} is increased, which typically occurs when the device needs to be turned OFF. (b) The subthreshold slope increases when channel length is decreased, which slows down the variation of current with gate voltage below threshold. Both effects increase the OFF current.

drain junctions create depletion regions that penetrate the channel region from both sides of the gate, thus shortening the effective channel length. These depletion regions carry with them electric fields that penetrate some distance into the channel region and "steal" some of the channel control from the gate. When the drain voltage is increased, this penetration is amplified. As a result, the potential in the channel region and the resulting concentration of electrons are no longer controlled solely by the gate electrode, but are also influenced by the distance between source and drain and by the voltage applied to the drain. The observable effects resulting from this loss of charge control by the gate are known as "drain-induced barrier lowering" (DIBL), which causes the threshold voltage to decrease as the drain voltage is increased, and a degradation (i.e. an increase) of the subthreshold slope results; see Fig. 1.6. The effects are additive and increase the leakage current of the transistors, which constitutes a serious impediment to further scaling of MOSFETs. The loss of switching speed caused by the DIBL effect is given by $\Delta f/f = -2\text{DIBL}/(V_{DD} - V_{TH})$, where f is the maximum operating frequency, V_{DD} is the supply voltage, and V_{TH} is the threshold voltage of the transistor. For example, in a circuit operating with a supply voltage of 0.9 V with transistors having a threshold voltage of 0.4 V, an increase of DIBL by 50 mV will slow down operating frequency by as much as 20% [18].

1.5 Technology boosters

Scaling down the size of transistors is not just a matter of being able to pattern smaller structures by improvement of lithography techniques. It also involves a constant striving to improve the performance of both the "intrinsic" transistors (i.e. the channel) and the "extrinsic" elements such as gate, source, and drain resistance. Reducing the dielectric constant of inter-layer dielectrics, and using low-resistivity metals such as copper, has also contributed to continuous improvement of the performance of integrated circuits. Aside from the reduction of device dimensions using ever more sophisticated lithography techniques, the performance of transistors has been enhanced by three main "technology boosters": the use of new materials, the use of strain, and the change of transistor architecture.

1.5.1 New materials

During the 1980s, only a handful of elements were used in silicon chip manufacturing: boron, phosphorus, arsenic, and antimony were used to dope silicon, oxygen, and nitrogen for growing or depositing insulators, and aluminum for making interconnections. A few elements, such as hydrogen, argon, chlorine, and fluorine, are, and continue to be, used during processing in the form of etching plasmas or oxidation-enhancing agents. Gold was usually used at the end of the process to form an ohmic contact to the back of the silicon wafer. Potassium was used in the form of KOH solutions, which can etch silicon in an anisotropic manner.

Later during the 1990s, a few more elements were added to the list, such as titanium, tungsten, cobalt, and nickel, which were used to form low-resistivity metal silicides. Tungsten was introduced to form vertical interconnects known as "plugs," and bromine started to be used in a plasma form to etch silicon.

The 2000s saw an explosion in the number of elements used in silicon processing: the rare earth metals, hafnium and lanthanum lanthanide are being used to form oxides with high dielectric constants (high-κ dielectrics), carbon and germanium are used to change the lattice parameter and induce mechanical stresses in silicon, fluorides of noble gases are used in excimer laser lithography, and a variety of metals are used to synthesize compounds that have desirable work functions or Schottky characteristics. Mercury, cadmium, and tellurium are used in HgCdTe infrared sensors.

The 2010s saw the beginning of the use of sulfur and selenium as surface passivation elements, as well as the use of tin, alloyed to Ge, for making high-mobility, low-band-gap devices. Virtually all elements of the periodic table are now being put to use in nanoelectronics manufacture, with the notable exception of alkaline metals, which create mobile charges in MOS oxides and, of course, radioactive elements; see Fig. 1.7.

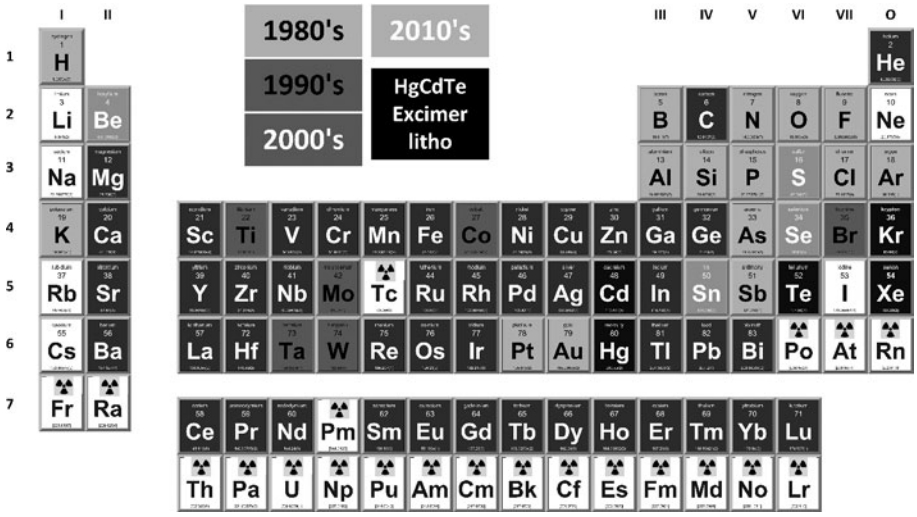

Figure 1.7 Elements used in semiconductor (silicon) industry. Radioactive elements are not used for obvious reasons.