Chapter 2
MOS Device and Interconnects Scaling Physics

Marc Van Rossum

2.1 Device Fundamentals

2.1.1 The MOSFET Transistor

2.1.1.1 Basic Device Physics

The metal-oxide-semiconductor field-effect transistor (MOSFET) is the most common active device in today's integrated circuits. Its basic structure consists of a doped silicon well, with at the opposite ends two highly doped contact regions (the source and drain junctions) allowing the current to pass close to the well surface (Fig. 2.1). In an n-type MOSFET, the well region is p-type doped and the source and drain are n+ doped, whereas the reverse polarity scheme applies for p-type devices. CMOS circuits contain both n-MOS and p-MOS transistors combined to form various logic gates. The transistor body is electrically isolated from the surrounding circuitry by a thick "field" oxide. A third electrode (the gate), to which the input signal is applied, is sitting on top of the well. It consists of an electrical contact layer (usually heavily doped polysilicon with a metallic top layer) separated from the silicon substrate by a thin insulator film made of thermally grown silicon dioxide. The substrate is thus capacitively coupled to the gate electrode, making the MOSFET a nearly ideal switch element due to the high isolation between input and output.

The output signal modulation takes place by varying the potential of the gate with respect to the substrate, which affects the charging of the MOS capacitor. In an n-MOSFET for instance, a negative gate voltage induces a positive (hole) charge accumulation region under the gate insulator. At positive gate voltages, holes are repelled into the substrate, creating a depletion region with fixed negative charges due to the ionized acceptor ions. At even more positive voltages, a negative charge

M.V. Rossum (✉)
IMEC, Kapeldreef 75, B-3001, Leuven, Belgium
e-mail: marc.vanrossum@imec.be

Y. Shacham-Diamand et al. (eds.), *Advanced Nanoscale ULSI Interconnects: Fundamentals and Applications*, DOI 10.1007/978-0-387-95868-2_2,

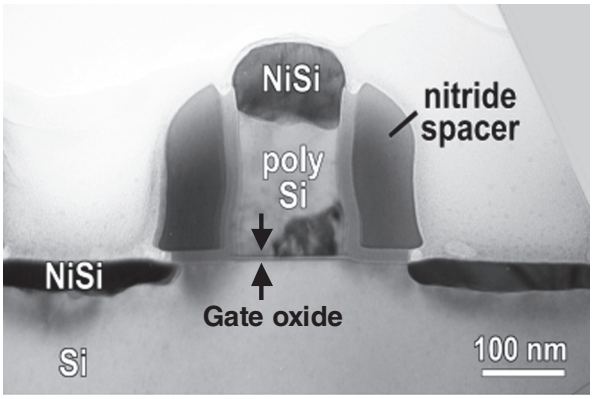

Fig. 2.1 The MOSFET transistor (IMEC)

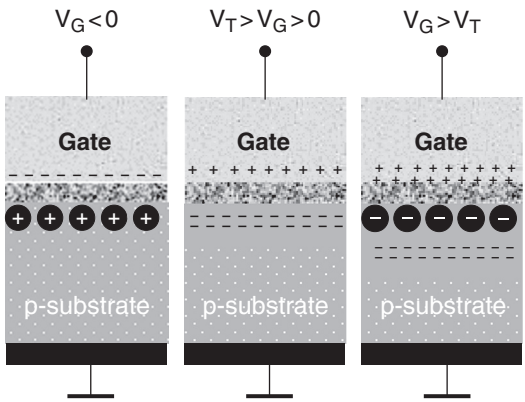

Fig. 2.2 Charge regimes of a MOS capacitor in n-MOSFET

inversion layer (i.e., with a negative charge imbalance) starts forming at the top of the depletion region by the minority carriers (electrons) that are attracted to the surface. The gate voltage that corresponds with the transition between the depletion and the inversion regime is called the threshold voltage V_T (Fig. 2.2).

According to the MOS capacitor model, the charge density Q_S induced into the substrate per unit area is linearly proportional to the applied gate voltage V_G:

$$Q_S = -(V_G - \psi_S)C_{ox} \tag{2.1}$$

where ψ_S is the band bending potential at the silicon surface and C_{ox} is the gate oxide capacitance expressed as

$$C_{ox} = \frac{\varepsilon}{t_{ox}} \tag{2.2}$$

with ε as the dielectric constant of the gate insulator and t_{ox} its physical thickness.

When a bias voltage is applied between source and drain (with the source usually kept at ground potential), a current is allowed to flow through the inversion layer. The threshold voltage therefore separates the "off-state" and "on-state" of the transistor. In a p-MOS structure, similar mechanisms take place with the proper reversion of polarities.

2.1.1.2 Technology

Today, the most advanced commercial transistors have a physical gate length of about 30 nm. Since the gate is the narrowest feature on any IC, its formation involves the most demanding steps of the front-end process flow. First, the SiO_2 gate insulator is grown by thermal oxidation of a clean silicon wafer in a controlled atmosphere. Subsequently, a blanket layer of polysilicon is deposited by chemical vapor deposition (CVD), after which the gate fingers are defined by lithography and patterned by dry etching. In order to achieve the right threshold voltage for the transistor, the polygate on the NMOS is n-doped whereas on PMOS it is p-doped. Doping is performed by ion implantation followed by annealing of the implantation damage. The source and drain electrodes are defined by local implantation of suitable doping species (n for p-well and p for n-well) very close to the surface, thereby forming shallow p–n or n–p junctions in the well, depending on the transistor type. The depth of the junctions scales with the other dimensions of the transistor (see Section 2.2), and in today's advanced devices it is often less than 100 nm, in which case they are referred as "ultra-shallow junctions." Fine-tuning of the junction profiles may require several implantation steps followed by annealing.

Fabrication of the electrical contacts to the source, gate, and drain involves specialized metallurgy. The contact material must exhibit low electrical resistance and be chemically compatible with silicon in order to avoid interface degradation over time. For many years, metal silicides have been used extensively on source and drain, first titanium disilicide ($TiSi_2$), later replaced by cobalt disilicide ($CoSi_2$), and more recently by nickel monosilicide (NiSi). The silicide layers are formed by solid-state reaction of a deposited metal film with the underlying silicon; therefore, it is important that the reaction should not consume too much silicon. In the same way, a polycrystalline silicide layer or polycide is formed on top of the polygate in order to reduce the gate series resistance (Fig. 2.3).

2.1.2 Current Regimes

MOSFET can operate in three distinct current regimes, depending on the gate bias and the source–drain voltage V_{DS} [1]. Following are simple expressions for the source–drain current as a function of V_{DS} (drain–source bias) and V_{GS} (gate–source bias) in a long-channel n-MOSFET:

The linear region: in this region the MOSFET behaves as a linear resistor with a resistance modulated by the gate voltage. According to Ohm's law, current

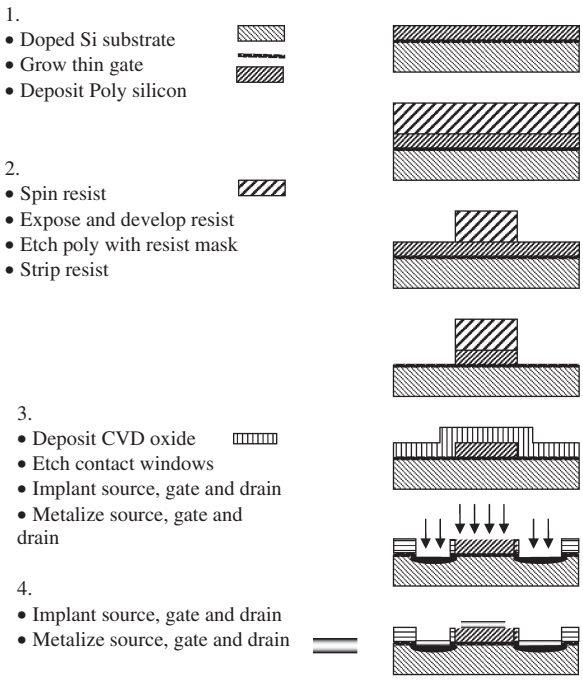

Fig. 2.3 Simplified process flow for MOSFET fabrication

modulation follows the variation of the channel resistance. The source–drain current is given by

$$I_{\mathrm{DS}} = \frac{\mu_n C_{\mathrm{ox}}}{2} \frac{W}{L} \left(2 \left(V_{\mathrm{GS}} - V_{\mathrm{T}} \right) V_{\mathrm{DS}} - V_{\mathrm{DS}}^2 \right) \tag{2.3}$$

where μ_n is the charge–carrier mobility, W is the gate width, L is the channel length, and V_{T} the threshold voltage. In analog circuits, devices often operate in this regime to take advantage of the linear amplification mode (Fig. 2.4a).

The saturated region: at fixed gate voltage, the channel width gradually narrows toward the drain with increasing source–drain voltage. Current saturation occurs when the channel nearly vanishes at the drain end ("channel pinch-off"). The saturation current depends on the gate bias but not on the source–drain bias; this behavior is usually referred to as the "long-channel characteristics." In this regime, I_{DS} is given by

$$I_{\mathrm{DS}} = \frac{\mu_n C_{\mathrm{ox}}}{2} \frac{W}{L} \left(V_{\mathrm{GS}} - V_{\mathrm{T}} \right)^2 \tag{2.4}$$

The MOSFET saturation current is usually written as I_{DSAT}. In digital circuits, the on-state of the device is normally set in the saturation region.

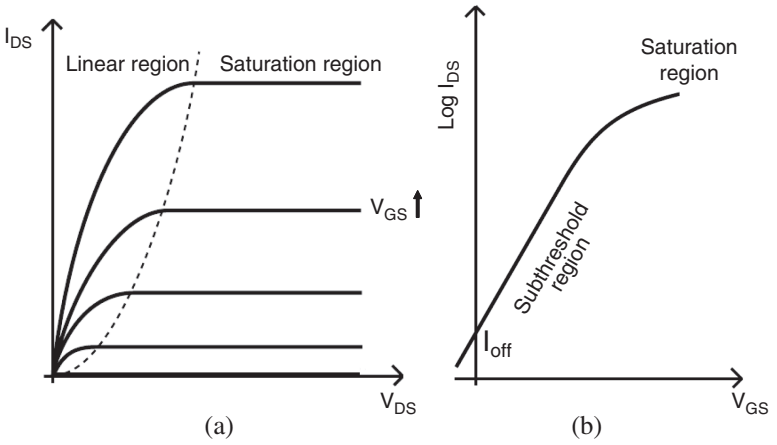

Fig. 2.4 MOSFET characteristics (**a**) linear and saturation regions; (**b**) subthreshold region

The subthreshold region: at fixed source–drain voltage, the inversion charge density decreases when the gate voltage approaches the threshold voltage. Below V_T, the inversion charge approaches zero asymptotically on a logarithmic scale. Therefore, the actual off-current reduces exponentially below the threshold voltage as

$$I_{DS} \propto \exp\left(\frac{V_{GS} - V_T}{\eta \cdot V_{Th}}\right) \tag{2.5}$$

where $V_{Th} = kT/q$ is the thermal voltage with k the Boltzmann constant, T the temperature, and q the electron charge. η is a nonideality factor which is approximately equal to $(1 + C_d/C_{ox})$ where C_d is the depletion layer capacitance at the onset of inversion:

$$C_d = \frac{\varepsilon_{Si}}{W_d} \tag{2.6}$$

with ε_{Si} the dielectric constant of the silicon substrate and W_d the depletion layer width. This is the so-called subthreshold regime (Fig. 2.4b), which can be characterized by the subthreshold slope S of I_{DS}, according to

$$S = \left(\frac{d(\log_{10} I_{DS})}{dV_{GS}}\right)^{-1} \cong 2.3 \cdot \frac{kT}{q} \cdot \eta = 2.3 \cdot \frac{kT}{q} \cdot \left(1 + \frac{C_d}{C_{ox}}\right) \tag{2.7}$$

Control of the subthreshold slope is an important aspect of the MOSFET scaling theory. It is important to notice that, to a first approximation, S does not depend on device dimensions. This creates extra complication for the scaling rules as will be shown below.

2.1.3 Mobility and Carrier Velocity

A simple estimate for carrier mobility μ in a semiconductor is given by the well-known Drude-type expression:

$$\mu = \frac{q\tau}{m^*} \tag{2.8}$$

where q is the electron charge, m^* the effective mass of the carrier (electron or hole), and τ the average relaxation time of the carriers, i.e., the average time lapse between successive collision events on the carrier's path. τ is a complex quantity including contributions from various scattering mechanisms, also taking into account the delocalization of the electron wavefunction.

Most scattering probabilities depend on the driving force experienced by the carriers and therefore on the electric field in the channel. The carrier mobility being thus electric field dependent will vary along the channel together with the effective field E_{eff} which follows from the applied source–drain bias. This complicated dependence can be approximated by a simple expression [2] for the effective mobility μ_{eff}, which holds for $E_{\text{eff}} < 5 \times 10^5$ V/cm:

$$\mu_{eff}\,(E) \approx 32500 \times \mathrm{E}_{eff}{}^{-1/3} \tag{2.9}$$

The effective mobility directly determines the field-dependent carrier velocity, and therefore the source–drain current, through the relationship

$$v_{\text{eff}} = \mu(E) \cdot E_{\text{eff}} \tag{2.10}$$

The decrease of the effective mobility with increasing field leads to the phenomenon of velocity saturation at high fields. This velocity saturation is caused by the increased scattering rate of highly energetic electrons, primarily caused by optical phonon emission. The overall proportionality between carrier velocity and electric field also changes with the device scale, since the effective field at constant bias increases at smaller channel dimensions. Moreover, the carrier mobility in the inversion layer is lower than in bulk material, because in this region the electron wavefunction penetrates into the gate oxide where higher scattering rates are experienced; high transverse electric fields at the channel surface – which typically result from device downscaling – shift the electron wavefunction even more into the oxide. The saturation velocity will ultimately depend on the balance between these effects.

The field-dependent mobility and the velocity saturation effect are some of the basic ingredients of the well-known drift-diffusion model, which computes the source–drain current under the assumption of a thermal equilibrium between the conducting electrons and the silicon lattice. However, this assumption no longer holds in very short gate devices, where high fields are present in the channel. In these devices, electrons will be driven to very high kinetic energies near the drain end of the channel, thereby effectively decoupling their energy from the lattice thermal bath. These "hot" carriers may acquire effective velocities that significantly

exceed the saturation velocity, which is about 1×10^7cm/s in planar MOSFETs. This effect is called velocity overshoot [3] and is at the origin of the increase in current drive and transconductance experimentally observed in nanoscale MOSFETs. At these scales, one may thus expect a stronger impact of the channel dimensions on the transistor switching speed. In fact, this effect is not as strong as one could expect from estimates of the maximal electron velocity, as obtained from Monte-Carlo simulations [4]. This is mainly because the velocity overshoot regime only affects a small fraction of the total path of the electrons, which remain at the velocity saturation threshold for most of their trajectory.

2.2 Digital Signal Propagation

2.2.1 Gate Delay

In digital data processing, bits represented by fixed voltage levels are shifted from one logic gate to the next following the rules of binary Boolean logic. An "input switching threshold" is the point at which an input signal to a logic gate first records the occurrence of a voltage transition. Input switching thresholds are usually specified as a percentage of the voltage differential between logic 0 and logic 1 (Fig. 2.5). The speed at which this voltage signal is processed by the CMOS device is associated with the latter's gate propagation delay. The gate delay (or propagation delay) is divided into two terms: the intrinsic gate delay and the (external) gate load delay. The intrinsic gate delay depends on the physical characteristics of the MOSFET transistors. The load delay includes the slowing effect of the load on the gate propagation delay. Therefore, the intrinsic gate delay equals the propagation delay under zero load condition. It can be defined as the time needed for the saturated transistor current I_{DSAT} at drain voltage V_{DS} to charge the gate capacitance C_G:

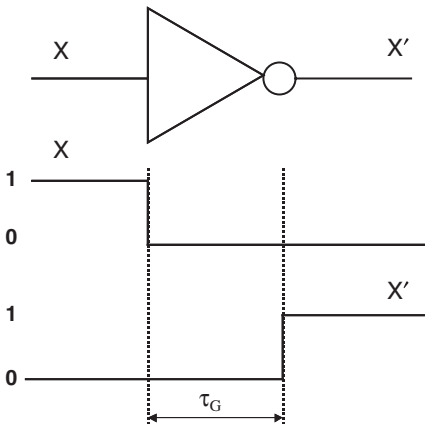

Fig. 2.5 Gate delay of logic inverter

$$\tau = C_G V_{DS} / I_{DSAT} \tag{2.11}$$

The *CV/I* characteristic time is an estimate of the intrinsic device switching speed, as it includes the intrinsic device capacitance, the voltage swing of the transistor, and its drive current. The dependence of τ on the device dimensions depends on the particular scaling rules being applied but until the latest generations the trend has been steadily downward (see Section 2.3). Moreover, there is also an intrinsic link between gate delay and gate length, through the scale-dependent carrier velocity (including possibly velocity overshoot) and the latter's influence on I_{DSAT}. This reinforces the tendency for τ to decrease with shrinking device sizes, as long as the downscaling has a positive influence on the drive current.

Since the *CV/I* figure does not take into account the external loads, it cannot provide a realistic estimate of the total propagation delay. From a practical point of view, the actual switching speed of a CMOS gate can be better derived from the inverter delay, which is defined as the time to propagate a digital signal through an inverter stage with a fan-out of one as, e.g., in a ring oscillator. This figure is correlated with the intrinsic gate delay, but will usually be an order of magnitude larger, because of the need to drive the next inverter stage.

2.2.2 Gate Delay Versus Interconnect Delay

In the past, the circuit delay was mainly determined by the gate delay figure. However, the latter has been continuously decreasing with device downscaling, due to the shrinking of the load capacitances. On the other hand, with shrinking circuit dimensions, the interconnect wire spacing (or pitch) becomes smaller, which must be compensated by the interconnect wire thickness in order to carry the required current for high-speed signal transmission. Overall, the interconnect resistance increases because of the smaller wire cross section, whereas the interconnect capacitance increases due to the combination of closer spacing and thicker wires. Since several layers of wiring are now required for interconnect and power distribution, this interconnect parasitic loading becomes the real limiting factor to speed.

Approximate expressions can be given for the latency of a single isolated interconnect that is *RC* limited within an ideal return path [5]

$$\tau\ (90\%) \cong r_{int}\, c_{int}\, L^2 + 2.3\, R_{tr}\, c_{int}\, L + 2.3\, C_L\, (r_{int}\, L + R_{tr}) \tag{2.12}$$

where r_{int} and c_{int} are the interconnect resistance and capacitance per unit length, R_{tr} is the source resistance, C_L is the load capacitance, and L is the interconnect length. Already at the 0.25 μm generation, the interconnect delay began to surpass the intrinsic gate delay. Figure 2.6 shows the gate delay with the corresponding interconnect delay for various CMOS nodes using aluminum interconnect technology. This rapid degradation has triggered the shift from Al wires to Cu technology around the 0.18 μm node.

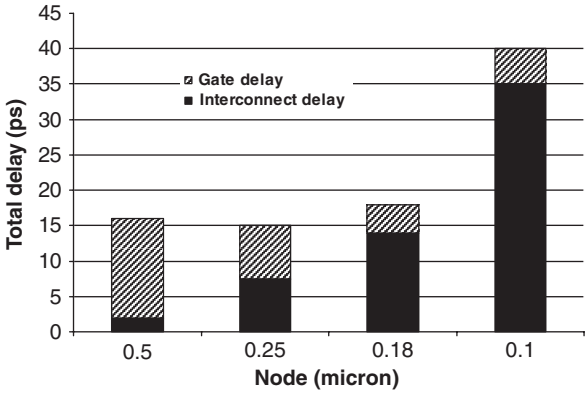

Fig. 2.6 Gate delay and interconnect delay (for Al wires) dependence on CMOS scaling

It is clear that, for state-of-the-art technologies, the gate delay is no longer the limiting factor for the circuit speed, and therefore the transistor switching speed can be traded for optimal overall performance against other device parameters such as the power dissipation. This is a very important consideration, since power, rather than speed, is becoming the main limiting factor for further miniaturization. As the technology proceeds into the nanometer era, the shift from device limited to interconnect limited design rules becomes a major trend, which is discussed at length in other chapters of this book.

2.2.3 Trends in CMOS Miniaturization

2.2.3.1 Moore's Law

Fast expansion of the semiconductor industry started very early after the invention of the first integrated circuits (1959–1960) and has since long been associated with Moore's law. Moore's original statement, issued in 1965, was modestly presented as an "educated guess" at the expected development of integrated circuits over the next 10 years. Or, to put it in his own words [6]: **"**With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip." Forty years later, unit cost is still falling with the number of components, and as long as this favorable trend persists, the "law" will remain firmly in place.

The mechanism underlying Moore's law can be understood using a simple model which we call "Moore's clock." Its two main features are found in any well-behaved watch, i.e., a spring and a pendulum. The spring provides the driving force that keeps the wheelwork running. In Moore's clock, this drive is essentially provided by the set of MOSFET scaling rules (see Section 2.2.8) which were first put forward some 6 years after Moore's initial paper, and which have shown almost the same remarkable endurance over time as the "law" itself (Fig. 2.7). With dimensional shrinking

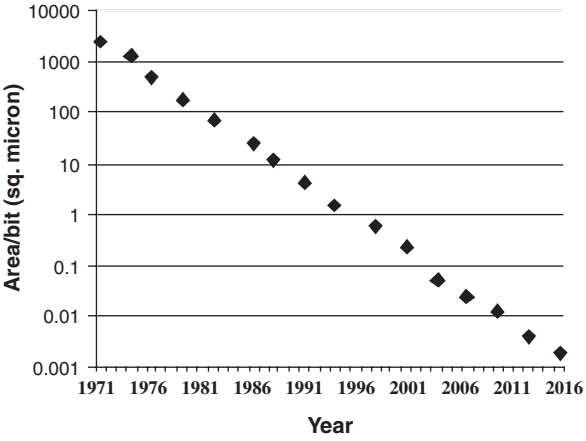

Fig. 2.7 Moore's law at work: scaling trend of DRAM cell area

now spanning more than two orders of magnitude, the persistence of scaling algorithms for CMOS ICs is a truly unique occurrence in the history of technology.

At this point, it should be realized that scaling rules apply to the definition of spatial dimensions, but do not define the pace of the miniaturization process. Therefore, Moore's clock also needs a pendulum to set the time with its periodic motion. In contrast to the spring, the pendulum is not only based on technical algorithms but also on business development factors. As such, it is closely linked with the microeconomic base cycle of the IC industry. The latter consists of four phases:

1. Downscaling of circuit components allows more functions to be integrated on the same area; therefore the average cost per function decreases.
2. Lower cost leads to market expansion and higher profit margins.
3. Profits are reinvested in R&D to prepare the next scale reduction.
4. In this way the cycle has repeated itself, on the average every 2–3 years, for almost four decades!

In spite of the apparent regularity of the pendulum, "setting the timescale" has always been the weak side of Moore's law. In reality, the speed of Moore's clock (Fig. 2.8) is not constant over time, but has gone through multiple stages [7]. In his 1965 article, Moore noted that the complexity of minimum cost semiconductor components had doubled every year since the first prototype IC (which did not contain MOSFETs but bipolar transistors) was produced in 1959. He then extrapolated the same trend until 1975, but at that time the cycle was already slowing down, as Moore himself later acknowledged. In the 1980s, Moore's law became stated as the doubling of number of transistors on a chip every 18 or 24 months. Later in the 1990s, it was widely associated with the claim that computing power at fixed cost is doubling every 18 months. In fact, none of these recent statements can be

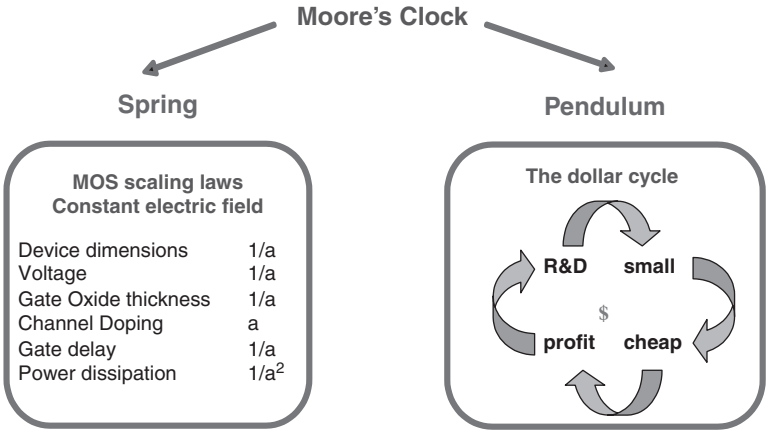

Fig. 2.8 Moore's clock

corroborated with accurate data. In its most general version, the law just points to regular doubling of "some measure of IC complexity."

2.2.3.2 Roadmaps

The National Technology Roadmap for Semiconductors (now the International Technology Roadmap for Semiconductors, or ITRS) was established in 1992 to codify the technological progress expected from Moore's law into a set of process targets and specifications, structured by the definition of future CMOS technology generations. The ITRS document is the product of a worldwide consensus building process in predicting the main trends of CMOS technology out to a 15 years horizon. The participation of experts representing the main actors in IC manufacturing and R&D ensures that the ITRS is a valid source of guidance for the global semiconductor industry.

The expected technological developments are timed with respect to the ITRS technology nodes, which are identified by the critical dimensions (e.g., the smallest half-pitch of contacted metal lines) of the circuits (90 nm, 65 nm...).[1] These numbers are "rounded off" figures derived from complex scaling formulas. Guidance for progress in the technological areas is provided by the definition of "grand challenges" to be met in moving to successive nodes.

Pressed by the champions of the semiconductor industry, the ITRS has regularly updated its forecasts of the CMOS scaling trends. At the end of the previous century, a phenomenon called "roadmap acceleration" was witnessed, by which the time window of each generation had gradually shortened toward a 2-year cycle (Fig. 2.9). For instance, the 1997 edition specified that the minimum device features of 100 nm

[1]The 2005 iteration of the ITRS roadmap has abandoned the simple concept of a unique node for all IC types, yet the technology generations are still labeled according to their critical feature sizes.

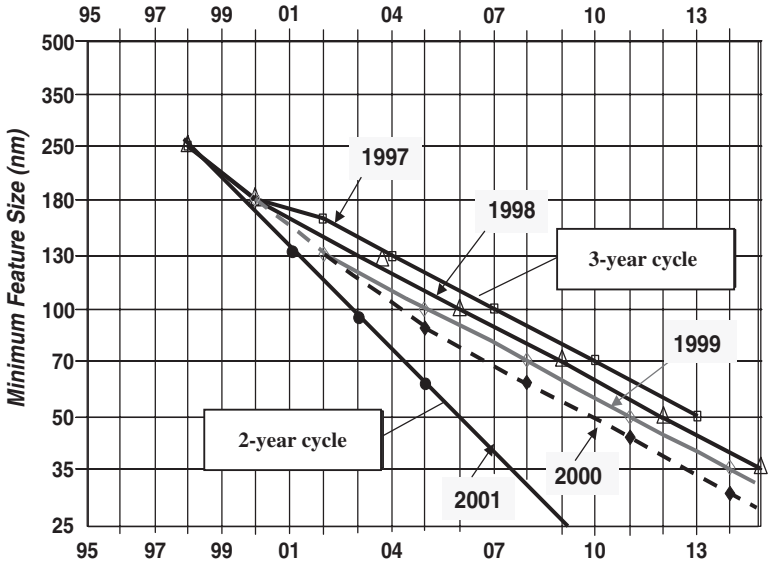

Fig. 2.9 Scaling acceleration of ITRS between 1997 and 2001 (IMEC)

would be reached in 2010. In fact, the 90 nm node was first delivered in 2003 and has a printed gate length (for high-performance devices) of 54 nm.

However, as a result of accumulating difficulties in the technological development of the latest generations, the ITRS has recently reversed this trend. Its most recent timescale (ITRS update 2005) assumes 3-year lapses between so-called "major" generations (see Table 2.1). In spite of this, some of the largest IC companies still attempt at "beating the roadmap" in an effort to secure a competitive advantage.

In spite of the recent slowdown, actual trends still clearly display the characteristic shape of an exponential growth rate. Exponential growth, however, also means that the fundamental limits of miniaturization are approaching rapidly. Many observers have therefore speculated about the "the end of Moore's law." The ITRS itself is putting a definite time horizon on each of its prognostics, but does not speculate on the endpoint position. In fact, up to this point the industry has been remarkably successful in keeping itself on schedule with the Roadmap timetables. In the last few years, however, the sustainability of the "Moore era" is being increasingly

Table 2.1 Near-term technology trend targets of ITRS 2005 (Source: ITRS).

Year of production	2006	2007	2008	2009	2010	2011	2012	2013
Microprocessor printed gate length (nm)	48	42	38	34	30	27	24	21
Microprocessor physical gate length (nm)	28	25	23	20	18	16	14	13

questioned among experts. In particular, there is a growing awareness within the ITRS groups that, within the next 10–15 years, most of the known technological capabilities will approach or have reached their limits. This threat was identified at the turn of the century and has since then been known as the "red brick wall."

2.2.3.3 Scaling Theory

For almost 40 years, shrinking the MOSFET transistor has been the dominant drive behind Moore's law. This process has been guided by the scaling laws, which in their original version were proposed as early as 1974 by Dennard et al. [8]. The basic idea of scaling is to reduce the dimensions of the MOSFET and associated interconnect wires, to produce a smaller transistor without degrading its performance. The original algorithm is based on the so-called "constant field" rule, whereby the applied voltages are scaled together with the geometrical dimensions of the device, such as to keep the internal electric fields constant. This can be achieved with a single scaling factor α, as can be seen in the second column of Table 2.2. This scaling algorithm has long been considered the most attractive, as it results in several simultaneous advantages:

1. The density of devices on the circuit increases by α^2, which is the basic claim of Moore's law.
2. The power dissipation per device scales like α^{-2}, which together with (1) results in a constant power dissipation density on the circuit.
3. The gate delay decreases by $1/\alpha$, due to the reduction of the device capacitance for a constant intrinsic transconductance.

The success of this model is largely due to the fact that, at least in the earlier stages, transistor performance as measured by its intrinsic gate delay would actually improve with scaling, whereas the potentially harmful high-field effects were avoided by decreasing supply voltages. However, the latter cannot be scaled down

Table 2.2 Scaling parameters for MOSFET transistors (see Refs. [8] and [10])

Physical parameter	Constant field scaling	Generalized scaling
Gate length	$1/\alpha$	$1/\alpha$
Gate width	$1/\alpha$	$1/\alpha$
Electric field	1	ε
Voltage	$1/\alpha$	ε/α
Gate oxide thickness	$1/\alpha$	$1/\alpha$
Channel doping	α	$\varepsilon\alpha$
Device area	$1/\alpha^2$	$1/\alpha^2$
Gate capacitance	$1/\alpha$	$1/\alpha$
Gate delay	$1/\alpha$	$1/\alpha$
Power dissipation	$1/\alpha^2$	ε^2/α^2

to arbitrary low levels. Indeed, at some point voltage reduction has to slow down due to following main reasons [9]:

- Reducing the threshold voltage increases the subthreshold slope of the MOSFET and the off current; this is mainly a consequence of the non-scaling of the thermal potential kT/q.
- As the power-supply voltage is reduced, the transistor performance degrades significantly at higher threshold voltages and also becomes more sensitive to tolerances in V_T.

For these reasons a switch was made in the 1980s from constant field scaling to a generalized scaling scheme [10], also shown in Table 2.2. In this new scheme, a second scaling constant ε is introduced to allow the electric field to scale independently of α. The supply voltage now scales with ε/α, the power dissipation with $(\varepsilon/\alpha)^2$, thus partially decoupling the electrical parameters from the dimensional scaling factor.

However, even this adjusted model has only a restricted validity range, as some limiting factors, generally known as short-channel effects, become stronger at smaller dimensions [11]:

1. Drain-induced barrier lowering (DIBL): the depletion barrier formed in the channel under the gate is lowered at higher source–drain voltages, which causes a degradation of the transconductance. This effect can be accompanied by the so-called punch-through that occurs when the depletion region surrounding the drain extends to the source.
2. Surface scattering occurs when electrons are accelerated toward the surface by the vertical component of the electric field. The scattering of the electrons by the surface potential causes a reduction in the mobility.
3. Hot electrons degradation, caused by electrons injected into the oxide at the $Si–SiO_2$ interface with high kinetic energy, can cause permanent damage to the gate insulator.
4. Velocity saturation has a stronger impact due to the upscaling of the electric fields with ε.

Other negative effects must also be taken into account. As a general consequence of physical scaling, bulk depletion charges are smaller than expected and the threshold voltage expression must be modified to account for this reduction. The scaling of physical dimensions is also limited in a practical sense by the discreteness of dopants, since present manufacturing techniques do not control the exact placement of dopant atoms. Consequently, since very small device volumes contain only a small number of dopants, large statistical variations become likely. In fact, the statistical distribution of dopants is only one of the sources of electrical variability that are likely to affect future circuits. Shift of device parameters also results from the increasing difficulty to control lithographic dimensions on a nanometer scale over the full circuit area. The first impact of the variability bottleneck can already be felt in today's circuit design, and the problem will likely get much worse for future

generations. Below some critical dimensions single devices can still be built, but large functional circuits may be difficult to design and manufacture with available techniques.

2.2.3.4 Scaling and Power Dissipation

The scaling algorithms also impact on the power consumption of the IC. There are two main sources of power dissipation in a CMOS device: dynamic (or active) switching power in the on state due to the charging and discharging of circuit capacitances, and static dissipation from leakage currents in the off state.

When CMOS devices switch, the output is either charged up to the transistor bias voltage or discharged down to ground. The power dissipated during switching is therefore proportional to the switching speed and to the capacitive load. The dynamic power dissipation arising from normal circuit operation is given by

$$P_{\text{on}} = C_{\text{EFF}} V_{\text{DS}}^2 f \tag{2.13}$$

where C_{EFF} is the effective output capacitance that is driven by the transistor and f the clock frequency of the circuit. Decreasing the clock frequency and/or the drain bias is therefore an efficient (although not always desirable) way to lower the dynamic power consumption.

The static power dissipation is taking place between switching events and is associated with source-to gate and source-to-drain leakage mechanisms. The source-to-gate leakage will be discussed in the next section. The source-to-drain leakage has two components: reverse-bias diode leakage on the transistor drains and subthreshold leakage through the channel when the transistor is turned off. Reverse-bias diode leakage must be tackled through process optimization, mainly by improving the quality of the junctions. Subthreshold current is a more complex issue. The subthreshold power dissipation formula is [12]

$$P_{\text{off}} = W_{\text{tot}} V_{\text{DS}} I_{\text{off}} = W_{\text{tot}} V_{\text{DS}} I_0 \exp\left(-\frac{qV_{\text{T}}}{mkT}\right) \tag{2.14}$$

where W_{tot} is the total device width, I_0 the extrapolated drain current per unit device width at threshold voltage, and m the so-called body effect coefficient. A simple expression for m is

$$m = 1 + \frac{3t_{\text{ox}}}{W_{\text{dm}}} \tag{2.15}$$

where t_{ox} is the gate insulator thickness and W_{dm} the bulk depletion layer width under the gate, which itself depends on the doping level of the channel.

The parameter that predominantly affects the P_{off} value is the threshold voltage, which must therefore remain above a critical value corresponding to the power tolerances set by the circuit design. In this respect, a distinction is usually made

between low power circuits, where power constraints are the main priority, and high-performance circuits allowing for more dissipation.

It can be deduced from previous formulas for P_{on} and P_{off} that both V_{DS} and V_T are important parameters in the setting of the overall power dissipation levels. A lowering of V_{DS} decreases the active power level, but at the same time it will have a deleterious effect on P_{off}, since V_T is limited by V_{DS} for efficient transistor operation [13]. Since lowering the threshold voltage leads to an exponential increase of the off current, it can only occur between narrow margins.

For very short MOSFETs, the gate-induced potential barrier between source and drain is so thin that direct source-to-drain tunneling becomes possible. Calculations of the source-to-drain tunneling current based on one-dimensional transport models have demonstrated the exponential dependence of the off state current on the depletion barrier width, which is a clear signature of direct tunneling phenomena. The effect of such tunneling first shows up in the degradation of the subthreshold slope of the device. Even if two-dimensional effects might worsen the picture somewhat, the general conclusion of these simulations is that source–drain tunneling should gradually become a major limiting effect for transistor operation below 10 nm gate lengths [14].

Below roughly 100 nm gate length, a major problem arises from the gate insulator, which in standard CMOS technology consists of a thin layer of thermally grown SiO_2. According to the scaling rules of Table 2.2, the thickness of this layer is reduced in the same proportion as the gate length. This is necessary to insure sufficient capacitive coupling ε/t_{ox} between the gate and the channel, and hence a good transconductance of the device. However, the direct quantum tunneling current from the channel to the gate electrode increases exponentially with decreasing oxide thickness (see Fig. 2.10) for a graphical estimate; accurate calculations are

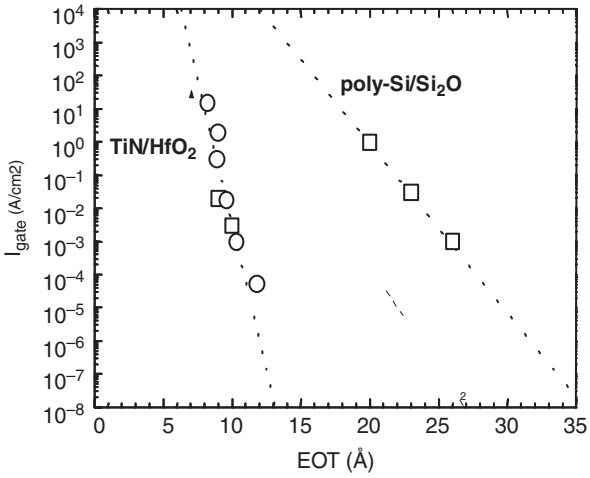

Fig. 2.10 Direct gate tunneling current density vs. effective oxide thickness for SiO_2 and for HfO_2-based gate stacks (M. Heyns, IMEC)

rather complex because they must take into account the band structure of the oxide). Although the tunneling current component is typically small as compared to the on current, it can become a sizable part of the off current in n-MOSFETs for gate oxide thicknesses below 2 nm, and will therefore contribute significantly to the static power dissipation of circuits below the 90 nm node. Assuming an acceptable gate leakage current of 10 A/cm^2, which at this time is a representative number for low operating power circuits, sets a practical lower limit of about 1.5 nm for the SiO_2 insulator thickness. Moreover, the direct tunneling current is only weakly dependent on the applied voltage and can therefore not be compensated by the scaling of V_{DS}.

The way out of this dilemma would be to replace SiO_2 by another insulator with a higher dielectric constant and therefore a lower equivalent SiO_2 thickness. This would suppress the direct tunneling current by increasing the gate insulator thickness, while keeping enough coupling between gate and channel to avoid short-channel effects. The challenge with high-*k* dielectrics is to find an insulator material resulting in an equivalent (SiO_2) oxide thickness (EOT) of at most 1 nm thick, and which would not degrade the operational properties of the transistor.

The key guidelines for selecting an alternative gate dielectric are

- Permittivity, band gap, and band alignment to silicon
- Thermodynamic stability; film morphology; interface quality
- Compatibility with other materials used in CMOS devices
- Process compatibility and reliability

Many dielectrics appear promising in some areas (see Table 2.3), but very few materials are capable of fulfilling all of these criteria. It must be kept in mind that replacing SiO_2 as the gate insulator will be a major milestone in the evolution of CMOS technology. Because it implies difficult changes in the fabrication process, the industry has opted for a gradual approach. The first alternative dielectrics to be introduced are based on silicon nitride or oxynitride, which are already well known and do not largely deviate from the standard technology. However, with shrinking

Table 2.3 Overview of high-*k* dielectrics (U. Berkeley)

High-*k* dielectric	*k* value
SiO_2	3.5
$S_{i3}N_4$	7
$Si_xN_yO_z$	4–7
Al_2O_3	9
Ta_2O_5	25
ZrO_2	25
HfO_2	40
TiO_2	50
$BaSrTiO_3$	300

gate dimensions a transition to more radical alternatives with higher k values is becoming more pressing.

As of today, the most promising candidates have been identified in the family of refractory metal oxides (mainly the Hf- or Zr-based ones) and their silicate compounds, such as $Hf_xSi_yO_z$, as well as their nitrided counterparts. The latter films are easier to etch than pure HfO_2. However, they must be deposited with sophisticated chemical vapor deposition (CVD) or atomic layer deposition (ALD) techniques. At this point, the main obstacle remains the poor quality of the high-k/silicon interface, resulting in gap states caused by Hf–Si bonds or oxygen vacancies. These local states lead to Fermi level pinning, V_T shifts, mobility degradation, and reliability problems. A common procedure nowadays in use is the intercalation of a thin SiO_2 interlayer between the high-k and the channel to improve the interface quality. Because of the difficult materials issues involved, introduction of high-k materials is likely to be pushed back to the 45 nm node, especially for high-performance circuits.

There are also problems to be solved with respect to the compatibility of high-k insulators with the gate electrode, which traditionally has been polysilicon ("poly"). Indeed, high-k materials and polysilicon gates are incompatible due to the above-mentioned Fermi-level pinning at the dielectric/poly interface. Therefore, many researchers believe that high-k layers will have to be used in conjunction with a metal gate or even two different metals for PMOS and NMOS devices for a better positioning of the respective threshold voltages. Moreover, there is evidence that metal gates by themselves offer some performance advantages, even with conventional dielectrics. One of the primary candidates is a metal gate made of NiSi, also known as "FUSI" (fully silicided gate). This approach can draw on the extensive knowledge of silicides processes, and especially of NiSi which is already in use for source and drain contacts.

2.3 New Device Structures and Materials

2.3.1 Strained-Silicon MOSFETs

As has been discussed above, many of the problems showing up in device miniaturization are related with the degradation of their transport properties. Strained silicon has recently been introduced in CMOS devices as a means to improve the carrier mobility in the channel, which should lead to shorter switching times. Evidence that transistors fabricated with strained-silicon channels were indeed faster accumulated during the 1990s, and was decisively demonstrated when the 90 nm node was reached. Therefore, strained-silicon channels have now become an integral part of the ITRS roadmap.

The first approach (so-called "global strain" and pioneered by IBM) for applying stress to the devices used a silicon germanium buffer layer between the substrate and the transistor channel (Fig. 2.11). $Si_xGe_{1-x}$is a near-ideal solid solution whose

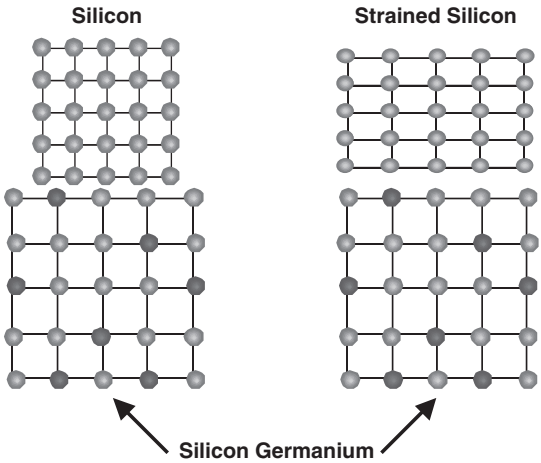

Fig. 2.11 Straining the silicon channel by growing it on a SiGe buffer layer

lattice constant follows Vegard's law to a good approximation. The lattice mismatch between Si and Si_xGe_{1-x} depends on the Ge content, but can easily reach 1% or more due to the 4% larger Ge lattice. When a thin silicon layer is grown epitaxially on top of a silicon germanium buffer, a pseudomorphic Si lattice results with a stretched in-plane lattice constant. The increase in lattice spacing produces biaxial strain in the silicon channel, which changes the shape of the energy bands both for electrons and holes. This deformation results in an increased mobility and channel drive current, which can be observed on n-type as well as on p-type devices. An alternative approach is the local strain method, developed by INTEL, which uses different processes for n- and p-MOSFETs. The n-channels are put under uniaxial tensile stress by depositing a thin silicon nitride film on the gate area, whereas the p-channels are compressed sideways (but also uniaxially) by growing local silicon germanium pockets under the source and drain areas (Fig. 2.12). Although the underlying solid-state mechanisms are basically the same as in the

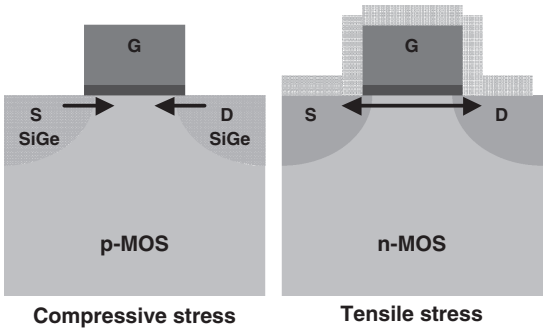

Fig. 2.12 Uniaxial strain applied to n-MOS and p-MOS devices

global scheme, the performance improvements at the circuit level tend to be better than with the global method, and therefore this local approach has now gained wider acceptance.

2.3.2 Silicon-On-Insulator (SOI)

At present, there is much interest in using Silicon-On-Insulator (SOI) wafers in advanced CMOS. SOI substrates consist of three layers: a thin surface layer of monocrystalline silicon on which the transistors are made, an underlying layer of amorphous SiO_2, and the bulk silicon support wafer underneath. The insulating silicon dioxide is referred to as the "buried oxide" or "BOX" and is typically a few thousand Ångström thick. There are several techniques for BOX fabrication, the most popular ones being buried oxygen implants and wafer bonding. The SOI wafer structure has several important advantages over bulk or epitaxial starting wafers. SOI wafers offer near-perfect transistor isolation (resulting in lower leakage currents and tighter transistor packing density), reduced parasitic drain capacitance (hence higher switching speeds and lower power consumption), and some process simplification relative to bulk or epitaxial silicon wafers. Due to these advantages, SOI wafers appear to be well suited for high-performance ICs requiring high-speed switches, high integration density, and low voltage/low power operation. Due to the reduced leakage levels, SOI should also be beneficial for battery operated systems. Moreover, SOI wafers offer an excellent platform for integrating RF and digital circuits on the same chip.

SOI substrates are used in two main application schemes; partially depleted (PD) and fully depleted (FD) SOI transistors, depending on the depth of the depletion layer with respect to the upper crystalline Si thickness. Both have specific advantages and drawbacks, but the general trend is toward the FD technology, because it allows higher circuit performance for a given power dissipation.

In the past, the main barriers to the widespread adoption of SOI wafers for mainstream CMOS fabrication have been the uncertain material quality and the higher cost of SOI wafers. The key materials quality issues are continuity and thickness uniformity of the BOX and defect density and thickness uniformity of the device-quality, single-crystal silicon layer. However, the wafers that are now commercially available are considered to be technically and economically ready for use in mainstream CMOS IC production.

2.3.3 Strained Silicon and SOI

The SOI MOSFET reduces the amount of current needed to switch a transistor, while the strained-silicon MOSFET increases the amount of current the transistor has available for switching. These improvements being complementary, it

would seem natural to combine both in an effort to achieve maximum performance. Moreover, there are other technical arguments for integrating strained silicon and SOI. Strained-silicon channels are usually adjacent to thick layers of SiGe, so the source and drain junctions of a bulk strained-silicon MOSFET will penetrate into the SiGe. Since the latter has a lower energy gap and higher dielectric constant than bulk Si, this leads to higher junction capacitances as well as higher junction leakage. However, when a strained-silicon channel is formed on an SOI structure, the increased junction capacitance and leakage associated with SiGe are restrained by the silicon-on-insulator structure, thereby improving the transistor performance. Strained silicon can even be deposited directly on SOI, without SiGe interlayer (SSOI process). After the strained-silicon SOI substrate has been formed, the rest of the fabrication process can continue as for a normal SOI circuit flow.

2.3.4 Germanium and III–V Channel Devices

With continuous downscaling, the degradation of transistor transport properties is likely to become more acute. Over the last decade, device dimensions have been shrunk by an order of magnitude, but drive current has only doubled. Since planar silicon may be unable to accommodate the rigorous current scaling requirements of sub-22 nm geometries, recent research has identified Ge as a potential alternative. The higher carrier mobility in Ge makes it a candidate for high-performance CMOS devices, which could easily be integrated into the existing silicon manufacturing infrastructure. Since high-*k* materials are under development to replace thermal oxides, the problem of the gate insulator on Ge is potentially solvable. Using similar arguments as for silicon, there is also the potential of using germanium-over-silicon–germanium and its combination with germanium-on-insulator at some point in the future.

However, much research is still needed to remove the possible showstoppers before applying Ge semiconductor material to advanced CMOS scaling. The main bottlenecks for a future Ge device technology are the passivation of interface states, reduction of diode leakage, and availability of high-quality germanium-on-insulator substrates. Although progress has been made on these three issues, specific problems with the n-type activation of Ge channels, as well as disappointing mobility date in n-type Ge MOSFETs, have cast some doubts on the future use of Ge for nMOS. Therefore, alternatives for Ge nMOS in advanced CMOS are presently investigated. It is well known that several III–V materials show large electron mobilities. Moreover, GaAs has a lattice parameter very close to that of Ge, which allows nearly defect-free epitaxial growth of GaAs on Ge. This opens the possibility of making high-performance CMOS with a Ge pMOS device and a GaAs nMOS device on the same substrate. In the future other III–V materials with even higher mobility than the one of GaAs could be investigated. One of the main challenges of this approach is to optimize the gate stack for MOS devices on Ge as well as on

III–V compounds. At this time, it is too early to tell if the Ge/III–V scheme will be able to reach the CMOS mainstream technology integration level.

2.3.5 Novel MOSFET Devices

In the coming years, the major scaling challenges at the device level will be

- Controlling leakage currents and short-channel effects
- Increasing the drive current while reducing the overall power supply
- Reducing the variability of the device operational parameters across the chip and from chip to chip

Although new materials are expected to play an important role with CMOS scaling now entering into the nanometer regime, it is also expected that radical changes in device geometries will be necessary to solve the bottlenecks just mentioned. The majority of the new materials has been reviewed in the preceding sections and includes gate stack (high-*k* dielectric and metal gate) materials, channel materials with improved carrier transport properties, as well as some new materials for the source/drain regions with reduced resistance and carrier injection properties. New transistor structures seek to improve the electrical behavior of the MOSFET and accommodate the integration needs of new materials. The combination of new structures and new materials enables novel device operating conditions that may provide better performance by overcoming the physical constraints of bulk planar CMOS. A starting point for this evolution could be provided by the double-gate MOSFET architecture. In this structure, a second gate and gate insulator are inserted at the device bottom between the channel and the substrate, thereby substantially improving the gate to channel coupling. The better gate control over the channel region steepens the subthreshold slope in the off state; moreover, it also increases the on-state current by providing a second current path along the channel bottom, thereby improving the I_{on}/I_{off} ratio. However, the price to be paid is a considerable complication in manufacturing processes. Indeed, the back gate must be self-aligned with the source and drain junctions as well as with the front gate, in order to avoid excessive parasitic capacitances. Furthermore, both gates must be connected via a low-resistance path to minimize the parasitic resistance. Because these steps are very difficult to optimize by standard lithographical means, attention has recently shifted to a more manufacturable version of the original DG-FET, called the FinFET. This device eases the process requirements by placing the silicon channel (the “fin”) perpendicularly on the substrate, thereby effectively creating three-dimensional device geometry. In this geometry, “top” and “bottom” become “front” and “back” gates, both of which can be easily accessed from the top of the wafer during processing (see Fig. 2.13).

The fin dimensions must be optimized to alleviate short-channel effects, which require the fin thickness to be no more than about a quarter of the gate length. In the

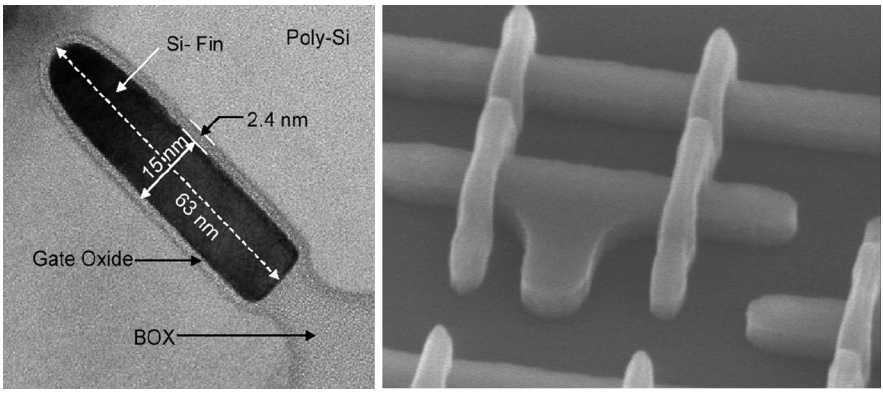

Fig. 2.13 FinFET structure and FinFET transistor top view (IMEC)

smallest devices, this ratio exceeds the present lithographic printing capabilities and thus requires special patterning techniques [15]. Another difficulty arises from the drive current requirements. In a planar MOSFET, the drive current can be increased just by making the device wider. In a FinFET, however, the effective width is limited by the height of the fin.

Increasing the drive current must be achieved by placing several fins in parallel and connecting them with bridge structures. This structure is known as the multi-gate FET or MuGFET. Among all available options, the multi-gate FET is considered as the most serious candidate due to the possibility of implementing the double-gated device concept with standard CMOS processing. Present development efforts focus on the modules specific for MuGFET topography such as fin and gate patterning, implementation of advanced gate stacks, ultra-shallow source and drain junctions, and mobility enhancement techniques.

As stated above, the device widths in the MuGFET architecture can be increased at a fixed lithographic scale by increasing the height of the silicon fins, thus providing more device area in a physical area than is possible to obtain with planar devices. While the MOSFET performance as measured by CV/I delay is not improved, since both C_G and I_{DSAT} increase in direct proportion to the fin height, interconnect contributions to delay may be decreased by allowing for closer placement of MOSFETs of the same drive capability and hence lower interconnect capacitance and resistance, [15,17] This is important since, as stated above, such interconnect delays already present major obstacles to scaling CMOS designs. Thus, one new direction (literally) for device scaling could become the vertical direction with respect to the wafer plane.

High parasitic resistance of source and drain regions are still obstacles on the way to reach high MuGFET performance. Selective epitaxial growth has been implemented into the MuGFET process flow to increase the fin width outside the spacers and lower the contact resistance. Successful implementation of this concept has already resulted in substantial drive current improvement for nMOS as well as pMOS devices [16].

References

1. Taur, Y. and Ning, T. H.: Fundamentals of Modern VLSI Devices. Cambridge: Cambridge University Press (1998)
2. Baccarani, G. and Wordeman, M. R.: Transconductance degradation in thin-oxide MOSFET's. IEEE Trans. El. Dev. **ED-30**, 1295 (1983)
3. Sai-Halasz, G. A.; Wordeman, M. R.; Kern, D. P.; Rishton, S.; and Ganin, E.: High transconductance and velocity overshoot in NMOS devices at the 0.1 mm-gate-length level. IEEE Electron Device Lett. **EDL-9**, 464 (1988)
4. Laux, S. E. and Fischetti, M. V.: Monte Carlo simulation of submicron Si n-MOSFETs at 77 and 300 K. IEEE Electron Device Lett. **9**, 467 (1988)
5. Meindl, J. D. et al.: Interconnect opportunities for gigascale integration. IBM J. Res. & Dev. **46**, 245 (2002)
6. Moore, G. E.: Cramming more components onto integrated circuits. Electronics **38(4)** (1965)
7. Tuomi, I.: The life and death of Moore's law, published on-line in First Monday, **7** (2002)
8. Dennard, R. H.; Gaensslen, F. H.; Yu, H. N.; Rideout, V. L.; Bassous, E.; and LeBlanc, A. R.: Design of ion-implanted MOSFETs with very small physical dimensions. IEEE J. Solid-State Circuits, **SC-9**, 256 (1974)
9. Davari, B.; Dennard, R. H.; and Shahidi, G. G.: CMOS scaling for high performance and low power-the next ten years. Proc. IEEE **83**, 595 (1995)
10. Baccarani, G.; Wordeman, M. R.; and Dennard, R. H.: Generalized scaling theory and its application to a 1/4 micrometer MOSFET design. IEEE Trans. Electron Devices **ED-31**, 452 (1984)
11. Frank, D. J.; Dennard, R. H.; Nowak, E.; Solomon, P. M.; Taur, Y.; and Wong, H.-S. Ph.: Device scaling limits of Si MOSFETs and their application dependencies. Proc. IEEE **89**, 259 (2001)
12. Taur, Y. and Ning, T. H.: Fundamentals of modern VLSI devices. Cambridge: Cambridge University Press, 271 (1998)
13. Gonzalez, R.; Gordon, B. M.; and Horowitz, M. A.: Supply and threshold voltage scaling for low power CMOS. IEEE J. Solid-State Circuits **32**, 1210 (1997)
14. Likharev, K. K.: Electronics below 10 nm. In: Nano and Giga, Challenges in Microelectronics. Amsterdam: Elsevier, 27 (2003)
15. Nowak, E. J.; Aller, I.; Ludwig, T.; Kim, K.; Joshi, R. V.; Ching-Te, C.; Bernstein, K.; and Puri, R.: Turning silicon on its edge [double gate CMOS/FinFET technology]. IEEE Circuits and Devices Magazine **20**, 20 (2004)
16. IMEC results (2005).
17. Nowak E. J.; Maintaining the benefits of CMOS scaling when scaling bogs down. IBM J. Res. & Dev. **46**, 169 (2002)

http://www.springer.com/978-0-387-95867-5

Advanced Nanoscale ULSI Interconnects: Fundamentals and Applications

Shacham-Diamand, Y.; Osaka, T.; Datta, M.; Ohba, T. (Eds.)

2009, XX, 552 p., Hardcover
ISBN: 978-0-387-95867-5