Chapter 5
Short-Channel Effects in MOSFETs

Abstract Short-channel effects are a series of phenomena that take place when the channel length of the MOSFET becomes approximately equal to the space charge regions of source and drain junctions with the substrate. They lead to a series of issues including polysilicon gate depletion effect, threshold voltage roll-off, drain-induced barrier lowering (DIBL), velocity saturation, reverse leakage current rise, mobility reduction, hot carrier effects, and similar other annoyances. Mitigation of the problem posed by polysilicon gate depletion effect via restoration of metal gate structure is presented. Threshold voltage reduction makes it difficult to turn the transistor off completely. By DIBL effect, electrostatic coupling between the source and drain makes the gate ineffective. Velocity saturation decreases the current drive. The leakage current increases the power dissipation. Enhanced surface scattering degrades the mobility of charge carriers affecting the output current. Apart from these factors, impact ionization and hot carrier effects seriously impair the MOSFET performance and cause the device to diverge in behavior from long-channel ones. Notable solutions are the gate oxide thickness cutback, use of high-κ dielectrics, strain engineering, etc. Nevertheless, the various effects mentioned severely downgrade the performance of planar CMOS transistors at process nodes <90 nm.

5.1 Meaning of “Short Channel”

Gate length L_g represents the physical length of the gate. Actual length of the channel L is obtained by subtracting the sum total lateral diffusions of the source and drain junctions from L_g. The length $L_g > L$ and L tracks L_g but the difference ($L_g - L$) cannot be quantified precisely. The channel length L is being continuously reduced to increase the operational speed and to accommodate more number of components per chip. At a certain channel length, the so-called short-channel effects arise. These effects are named so because they occur explicitly by virtue of the fact that the channel is short. A more concrete definition of short channel follows in the next paragraph.

V.K. Khanna, *Integrated Nanoelectronics*,
NanoScience and Technology, DOI 10.1007/978-81-322-3625-2_5

The decrease of channel length is associated with the enlargement of source and drain depletion regions and their incursion into the channel. This incursion may take place even without any bias. It may lead to trespassing of the depletion regions into such zones which usually fall under the jurisdiction of the gate. A MOSFET device is deemed to have a short channel when the channel length is reduced to a certain consented degree. This degree has been mutually decided to be attained when the channel length is of comparable magnitude to the depletion layer widths (x_{dD}, x_{dS}) surrounding the drain and source junctions. A short-channel effect is an effect which is produced only when the channel has become short, and which is not observed otherwise. As a consequence of this effect, a MOSFET of channel length $L \approx (x_{\mathrm{dD}}, x_{\mathrm{dS}})$, deviates in behavior from a long-channel MOSFET having $L \gg x_{\mathrm{dD}}, x_{\mathrm{dS}}$. Short-channel effects originate from a variety of reasons: (i) the production of high electric fields in the channel region. (ii) the two-dimensional potential distribution in this region. This distribution depends on the transverse field E_x controlled by the gate voltage as well as the bias applied to the back surface. Longitudinal field E_y due to the drain bias also plays its role. The two-dimensional potential degrades the threshold comportment of the device. It makes the threshold voltage dependent on the channel length and biasing voltages. Circuit designers will be scared if V_{Th} changes with device dimensions or biasing voltages.

5.2 Polysilicon Gate Depletion Effect

In the beginning stages of MOSFET development, the metal electrode of the gate was replaced by a heavily doped polysilicon electrode. But this polysilicon is a semiconductor, not a conductor like a metal. It can deplete when the electric fields become high. The finite depletion layer created inside the polysilicon becomes uncooperative with continued device scaling. The potential dropped across the polysilicon layer becomes a large fraction of the supply voltage. This mandates a further reduction of oxide thickness to maintain a constant electric field across the gate dielectric. But the oxide is already very thin and does not allow further thinning. The polysilicon depletion effect can also occur in long-channel devices but its probability increases at smaller dimensions because of the strong electric fields that are created. The obvious solution to polysilicon depletion effect is to increase the doping concentration of polysilicon as much as possible, but this is restricted practically to $\sim 1 \times 10^{20}\ \mathrm{cm}^{-3}$ for N-type and $\sim 1 \times 10^{19}\ \mathrm{cm}^{-3}$ for P-type polysilicon. Fully silicided (FUSI) polysilicon gates, e.g., cobalt silicide, nickel silicide, hafnium silicide, platinum silicide, and titanium silicide gates are the useful approach for highly scaled CMOS [1, 2]. This disruptive method was given up due to the problems faced in controlling the silicide phase to achieve low threshold voltages. The other alternative is to replace the polysilicon gates with metal gates of work functions 4.1 and 5 eV for N^+ and P^+ polysilicon, respectively [Nishi]. Aluminum gates were the bastion of MOS integrated circuits until the introduction of polysilicon gate-based self-aligned process. A comeback of metal gates is mandated to avoid depletion of the gate conductor.

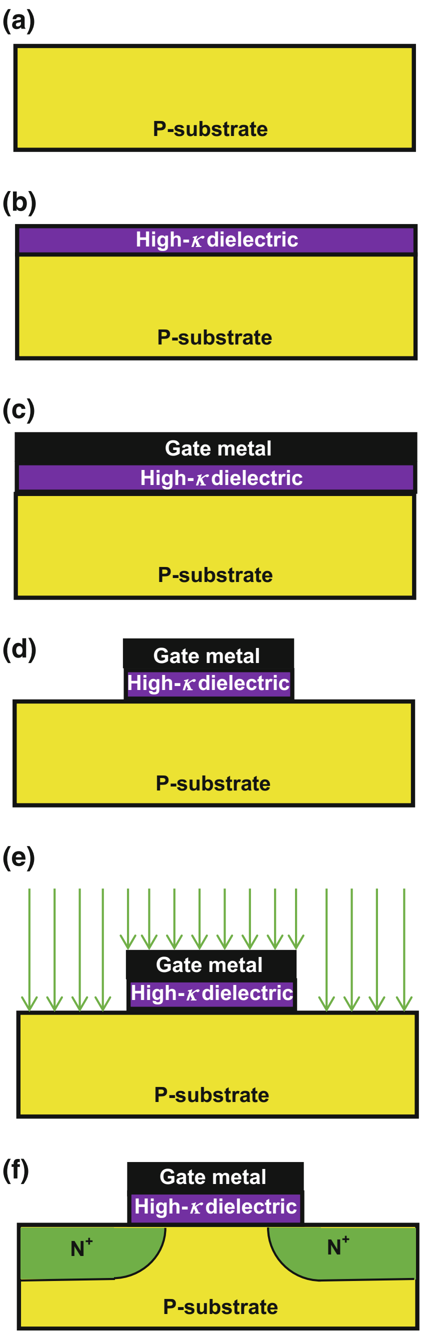
Advantages offered by metal gates are: (i) Low gate resistance and hence smaller gate RC (resistance–capacitance) delay. (ii) Absence of boron penetration from the polysilicon gate into the channel across the thin gate oxide. (iii) Appropriate work function adjustment: 4.1–4.4 eV for N-channel MOSFET and 4.8–5.1 eV for P-channel MOSFET for bulk silicon MOSFET and partially-depleted SOI-MOSFET. A spectrum of work function values covering the silicon band gap is available for fully-depleted SOI-MOSFET and FINFET devices. (iv) Diminution in electrical thickness of gate insulator.

5.3 Gate-First or Gate-Last Fabrication Flow

This issue is connected with that of high-κ dielectrics to be discussed in the next chapter. Actually, the gate consists of a series of layers; hence it will be expedient to talk about metal gate/high-κ or MG/HK stack. Two different fabrication flows are possible to integrate metal gates with the process [3]: (i) Gate-first approach: Traditional CMOS process is followed (Fig. 5.1). Hence, this kind of gate is called metal inserted-polysilicon gate (MIPG). The polysilicon gate is deposited at a very early stage in the process. It serves as a mask for the subsequent source/drain implantation. The lattice damages created in the implantation step are annealed out at a high temperature. Deposition of the metal gate in the beginning stage exposes it to high-temperature steps. Thermal annealing can destroy the integrity of MG/HK stack, causing thermal instability and threshold voltage shifts, besides contaminating the front-end equipment, specially the furnaces. These complications have obstructed the use of gate-first method for high-performance CMOS. (ii) Gate-last approach: This is designed to circumvent the obstacles faced in gate-first method. Here a dummy/sacrificial gate is made for masking the implantation (Fig. 5.2). After source/drain junctions have been formed and no further thermal cycling is required, the dummy gate is etched away. The real gate stack is then constructed. By chemical mechanical polishing (CMP), the gate metal is thinned down to desired thickness. Looking at the above process flow, this type of gate is known as a replacement metal gate (RMG). Table 5.1 presents a comparison between the two processes [4].

Essentially, the gate-first and gate-last approaches differ in the relative sequence or order of gate metal deposition and thermal annealing steps. In the gate-first approach, the metal gate is deposited prior to carrying out high-temperature activation annealing. In the gate-last approach, the annealing steps are carried out at an early stage in the process, and metal gate deposition is done afterwards.

Fig. 5.1 Gate metal first process. **a** Starting silicon wafer. **b** Deposition of high-κ dielectric film. **c** Gate metal deposition. **d** Photolithography and etching of metal and dielectric films. **e** N^+ source/drain implant. **f** High-temperature annealing



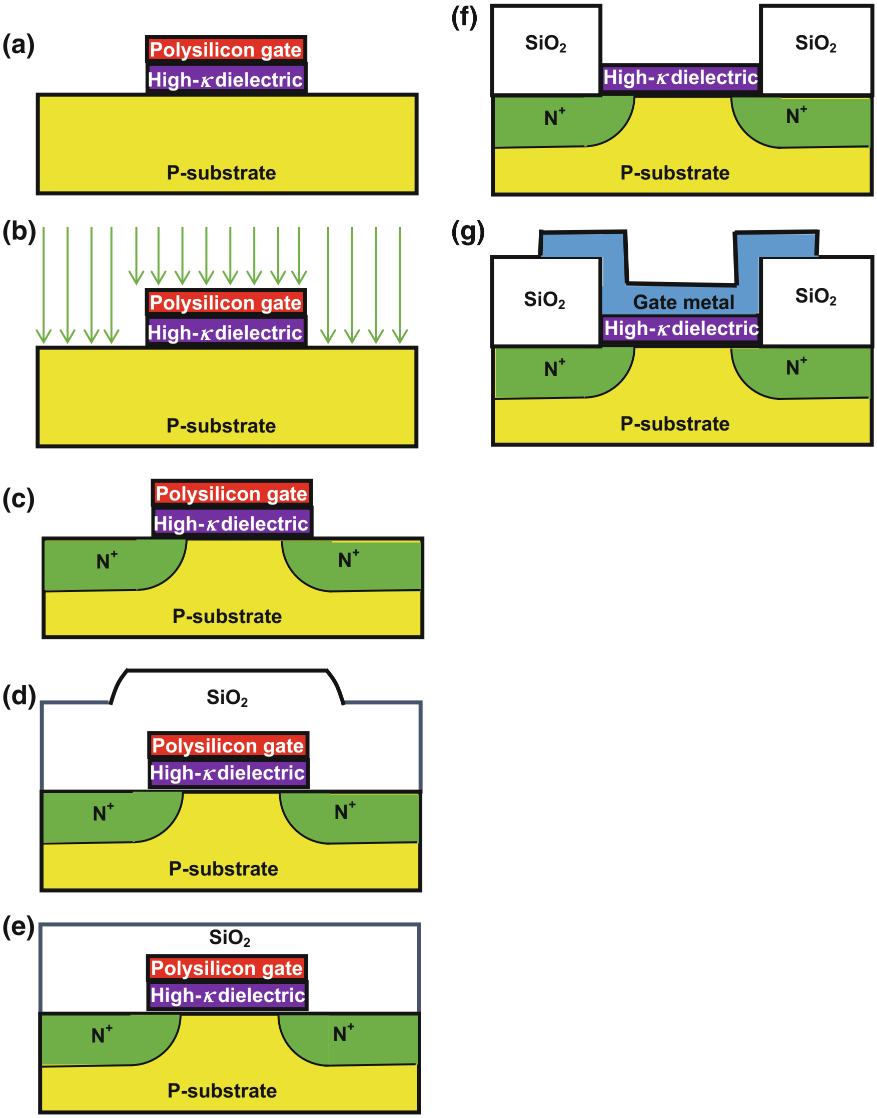


Fig. 5.2 Gate metal last process. **a** High-κ dielectric and polysilicon gate deposition. **b** N^+ source/drain implant. **c** High-temperature annealing. **d** Silicon dioxide coverage. **e** Chemical mechanical polishing. **f** Removal of dummy gate. **g** Replacement gate formation

Table 5.1 Comparison between gate-first and gate-last process flows

Feature	Gate-first	Gate-last
Gate insulator	First	First or last
Gate metal	First	Last
Process sequence	Conventional	Revised
Process complexity	Less	More
Thermal cycles given to gate metal	Large	Small
Reliability	Less because high-κ material has undergone temperature cycling	More
Applications	Low power DRAM where threshold voltage requirements are relaxed	High-performance applications

5.4 Threshold Voltage Roll-off and Drain-Induced Barrier Lowering (DIBL)

With the decrease in channel length of a MOSFET, the bulk charge terminating on the gate electrode decreases. The decrease in charge leads to a reduction of the threshold voltage. To explain this reduction, let us reiterate that the formation of an inversion layer underneath the gate dielectric is preceded by the depletion of this region up to a depth W_d. In depleting this region, the source and drain are copartners of the gate. Although the gate is responsible for a major fraction of this depletion, the source and drain junctions too contribute to depletion. Small portions of the depletion layer charge are balanced by the charges in the source and drain regions. Effectively, less gate charge is necessary for depletion than would be required if source and drain did not partake of the responsibility. This means that the threshold voltage is reduced by an amount ΔV_{Th} due to source and drain effects. For a long-channel MOSFET, ΔV_{Th} is negligibly small but for a short-channel device, it becomes conspicuous. Also, in the same wafer, any two transistors having different channel lengths will differ in V_{Th}. This is true even in the same die. The threshold voltage reduction due to the decreased channel length represents V_{Th} roll-off.

Drain-induced barrier lowering (DIBL) is the drain voltage-induced decrease in threshold voltage in a short-channel MOSFET at high drain voltages. It arises from electrostatic coupling between the drain and the source. In consequence to this coupling, the potential barrier of the source-to-channel junction is depressed. It is this depression of the potential barrier to current flow at the source under the influence of the drain voltage that is responsible for the reduction in V_{Th}.

DIBL occurs in a short-channel MOSFET due to the relatively pronounced charge sharing effect between the channel depletion region and source/drain depletion regions as compared to long-channel device case. If the barrier between the source and channel is decreased, electrons are more freely injected into the channel region. Consequent upon the injection of electrons, the transistor requires less gate voltage to deplete the substrate beneath the gate dielectric. Hence, V_{Th} is

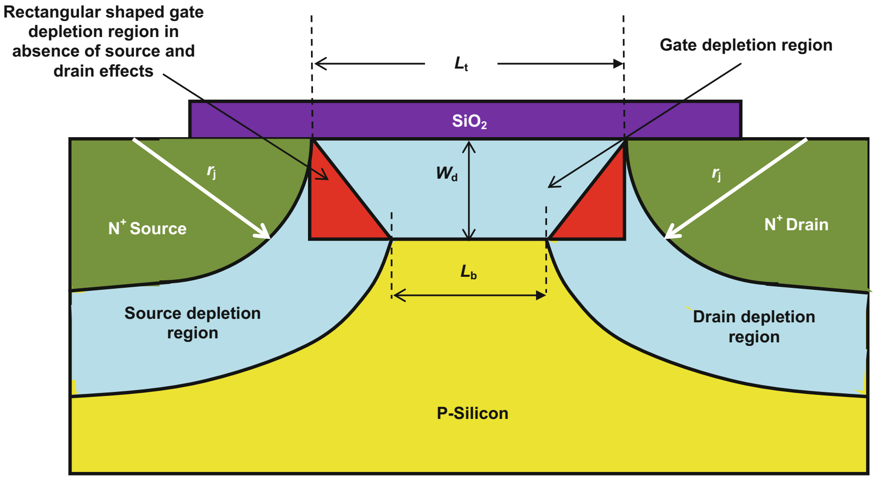


Fig. 5.3 Analysis of drain-induced barrier lowering

decreased and the transistor is switched on prematurely. Therefore, the threshold voltage is brought down and the gate has inferior control over the channel current. In a classical long-channel field-effect transistor, the barrier at the source end is situated far away from the drain contact. So, it is electrostatically shielded from the drain by the combination of the substrate and gate voltages. For this reason, the threshold voltage is independent of drain voltage.

Because of source and drain effects, the depletion region arising from the gate becomes trapezoidal in shape in place of the rectangular shape if these effects were absent (Fig. 5.3).

Let L_{t}, L_b denote the lengths of top and bottom parallel sides of the trapezium. Then the factor r by which the depletion charge is decreased from its value for the rectangle shape is given by

$$r = 1 - (L_{\mathrm{t}} + L_b)/(2L_{\mathrm{t}}) \tag{5.1}$$

If r_j is the radius of curvature of source/drain diffusion, it can be shown that

$$L_b = L_{\mathrm{t}} - 2r_j\left(\sqrt{1 + (2W_d)/r_j} - 1\right) \tag{5.2}$$

where W_d is the depletion region width underneath the gate. For a first order analysis, the change in threshold voltage with respect to a long-channel device is

$$\Delta V_{\mathrm{Th}} \equiv |V_{\mathrm{Th}}| - \left|V_{\mathrm{Th(Longchannel)}}\right| = -\left(\frac{qN_A W_d}{C_{\mathrm{ox}}}\right)\left(\frac{r_j}{L_{\mathrm{t}}}\right)\left(\sqrt{1 + (2W_d)/r_j} - 1\right) \tag{5.3}$$

where N_A is the doping concentration of the P-substrate and C_{ox} is gate oxide capacitance per unit area.

The change ΔV_{Th} is decreased by lowering N_A; by increasing C_{ox} with a thinner gate oxide; and by using a smaller r_j. For smaller r_j, the source and drain junctions can be made shallower but parasitic resistances R_{source}, R_{drain} of these regions will increase

$$R_{\mathrm{source}}, R_{\mathrm{drain}} \propto \rho / \left(W r_j\right) \tag{5.4}$$

where ρ is the resistivity of source/drain region and W is the channel width. For r_j reduction, shallow extensions of the source/drain regions are formed. The accompanying increase in R_{source}, R_{drain} is considerably low. For providing shallow extensions, a dielectric spacer is included in the MOSFET structure.

At very high V_{DS} values, the depletion regions of the source and drain touch each other. When they mingle together, a high current flows from source to drain. It is irrepressible by the gate. This phenomenon is called punch-through. If L is the channel length, the punch-through voltage V_{PT} is given by

$$V_{\mathrm{PT}} = qN_A L^2 / (2\varepsilon_0 \varepsilon_{\mathrm{Si}}) \tag{5.5}$$

showing that the punch-through voltage decreases as L becomes smaller.

5.5 Velocity Saturation

As the gate length is decreased to smaller values, the longitudinal electric field E_x between the source and drain increases and becomes larger. The behavior of the carrier velocity v_d varies according to E_x. At low values of E_x, the carrier velocity v_d is proportional to E_x. But at higher values of E_x, the proportionality relationship is grossly violated. When E_x exceeds 3×10^4 V/cm for electrons and 10^5 V/cm for holes, the carrier velocity saturates at a value $v_{\mathrm{sat}} = 10^7$ cm/s for electrons and $v_{\mathrm{sat}} = 6 \times 10^6$ cm/s for holes. This saturation is explained by the increased scattering rate at the high electric fields. The saturation current I_{DS} is no longer a quadratic function of V_{GS}. It increases linearly with $(V_{\mathrm{GS}} - V_{\mathrm{Th}})$:

$$I_{\mathrm{DSsat}} = v_{\mathrm{sat}} W C_{\mathrm{ox}} (V_{\mathrm{DS}} - V_{\mathrm{Th}}) \tag{5.6}$$

It is independent of channel length. It is lower than that for a long-channel MOSFET, resulting in a substantial decline in current drive. The short-channel MOSFET saturates at a lower V_{DS} value.

5.6 Carrier Mobility Degradation

5.6.1 Horizontal Field Effect

Mobility decreases as the carrier velocity saturates and becomes constant. Caughey-Thomas analytical equations for dependence of electron mobility $\mu_n(E)$ and hole mobility $\mu_{\mathrm{p}}(E)$ on electric field E parallel to the current flow are [5]

$$\mu_n(E) = \mu_{n0}\left\{1 + \left(\frac{\mu_{n0}E}{v_{\mathrm{satn}}}\right)^{\beta_n}\right\}^{-1/\beta_n} \tag{5.7}$$

$$\mu_{\mathrm{p}}(E) = \mu_{\mathrm{p}0}\left\{1 + \left(\frac{\mu_{\mathrm{p}0}E}{v_{\mathrm{satp}}}\right)^{\beta_{\mathrm{p}}}\right\}^{-1/\beta_{\mathrm{p}}} \tag{5.8}$$

$$v_{\mathrm{satn}} = v_{\mathrm{satp}} = \frac{2.4 \times 10^7}{1 + 0.8\exp(T/600)}\mathrm{cm/s} \tag{5.9}$$

where μ_{n0}, $\mu_{\mathrm{p}0}$ are low-field electron and hole mobilities, respectively (μ_{n0} = 1375 cm^2/V s, $\mu_{\mathrm{p}0}$ = 487 cm^2/V s); v_{satn}, v_{satp} are the corresponding saturation velocities; β_n, β_{p} are user-defined unit-less parameters determined experimentally (β_n = 2, β_{p} = 1); and T is the lattice temperature. These are empirical equations which are derived by the above authors from published experimental data.

5.6.2 Vertical Field Effect

Near the surface, enhanced scattering of carriers takes place due to surface acoustic phonons and surface roughness. Since the carrier transport is confined in the constricted MOSFET inversion layer near the silicon–silicon dioxide interface, the carriers experience great difficulty in moving parallel to the interface. Surface mobility is thereby decreased to $\leq$ half the bulk mobility. The mobility degradation by surface scattering is taken into account by writing the reciprocal of mobility as the sum of reciprocals of three terms according to Matthiesen's rule [6]:

$$1/\mu = 1/\mu_b + 1/\mu_{\mathrm{ac}} + 1/\mu_{\mathrm{sr}} \tag{5.10}$$

where μ_b is the carrier nobility in the bulk, μ_{ac} is the mobility limited by surface acoustic phonons, and μ_{sr} is the mobility limited by surface scattering. The μ_b term is obtained from the Klaassen model [7]. Lombardi model gives the μ_{ac} term [8]. Surface roughness scattering is treated in [8, 9].

5.7 Impact Ionization

In a short-channel NMOS transistor, the electrons accelerated to high velocities by the large longitudinal electric field bombard silicon atoms, liberating electrons from their outermost shells. The electrons thus released also acquire high velocities, taking part in further collisions, and generating electron–hole pairs. The drain attracts the ejected electrons while the holes move to the P-substrate. This multiplicative process produces an avalanche of free carriers. The N^+-source-P-substrate-N^+-drain acts like an NPN transistor. If the aforesaid holes are collected by the source, and the hole current produces a voltage drop in the substrate, the source-substrate junction will be forward-biased. Then electron injection will start from the source to the substrate. These electrons can move toward the drain, creating electron–hole pairs and aggravating the situation.

5.8 Hot Carrier Effects

The decreasing feature size of MOSFET devices is accompanied by an increase of the electric field in their channel regions. Hot carriers are charged particles, either electrons or holes. They include particles which have acquired very high kinetic energies upon acceleration by the large electric field prevalent across the channels of MOSFETs. These carriers have higher energies than those of carriers normally found in semiconductor devices. Due to their high energies, hot carriers may migrate into and roam around the unwelcome areas of the devices. Such areas are the gate dielectric and substrate of a transistor. They cause shifts in threshold voltage. Device transconductance is also degraded.

Hot carrier injection is more serious in N-channel MOSFETs than P-channel devices because of the higher mobility of electrons. Consequently, electrons acquire higher energies and become hotter than holes. Further, the energy barrier is lower for electrons than that for holes.

5.8.1 Substrate Hot Electron (SHE) Injection

When exceedingly high positive or negative positive or negative voltages are applied to the MOSFET substrate or body, SHE injection is triggered. Then the substrate field impels charge carriers of one type in the substrate toward the Si–SiO_2 interface. They gain high kinetic energy and are hurled into SiO_2.

5.8.2 Channel Hot Electron (CHE) Injection

When both V_{GS} and V_{DS} are very high, some electrons are driven toward the gate oxide.

5.8.3 Drain Avalanche Hot Carrier (DAHC) Injection

When $V_{DS} > V_{GS}$, the acceleration of charge carriers in the channel region causes impact ionization from atomic-level collisions near the drain. Electron–hole pairs are thus produced and further carrier multiplication ensues. The generated electron–hole pairs gain enough energy to break the barrier at Si–SiO_2 interface and penetrate into the oxide.

5.8.4 Charge Generation Inside SiO_2

(a) Negative charge buildup by trapping of hot electrons in the oxide near the drain. (b) Positive charge accumulation by injection of holes into the oxide. (c) Interface state generation at Si–SiO_2 interface. This happens because some Si–H, Si–Si and Si–O bonds need less energy to break. Taking advantage of this situation, any electron with energy >2 eV can release H_2 and create interface states.

5.9 Random Dopant Fluctuations (RDF)

These arise from statistical irregularities in dopant concentrations, which become pronounced when the MOSFET area decreases. As a consequence, threshold voltages of neighboring devices may differ.

5.10 Overcoming Short-Channel Effects in Classical MOSFETs

5.10.1 Avoiding DIBL Effect

DIBL effect is reduced by decreasing the gate oxide thickness. The thickness reduction makes the gate more effective in controlling the channel region.

5.10.2 Reducing Gate Leakage Current

Thermal silicon dioxide is amorphous in structure with dielectric constant 3.9. It is a three-dimensional network of tetrahedral cells. Atomic diameter of silicon is 0.236 nm, that of oxygen is 0.13 nm, average silicon–oxygen bond length = 0.162 nm, oxygen ion–oxygen ion distance = 0.262 nm, silicon–silicon bond distance = 0.31 nm [10]. Thermal SiO_2 has been the preferred MOSFET gate insulator. It was constantly scaled down in thickness up to the 130 nm technology node at the rate of $0.7x$ per MOSFET generation. But after reaching this node, the scaling became slower in pace particularly in sub-100 nm range. At 90 nm and 65 nm, it was considerably sluggish due to the resulting high leakage current. The ultimate limit at which bulk oxide could be used was that required in 70 nm technology node = 0.7 nm. It is about two atomic layers in thickness.

Now, gate oxide thickness = 1.2 nm is being used in MOSFET manufacturing. So, the present oxide thickness is hardly a single atomic layer thick [11]. The oxide thickness at which direct tunneling starts is 3 nm. With reduction of gate oxide thickness, direct quantum-mechanical tunneling of electrons from the gate across the gate oxide to the underlying silicon causes an increase in the gate leakage current. The gate leakage current density = 100 Acm^{-2} at 1 V for 1.2 nm thick oxide.

After reaching the 70 nm node, a desperate need was felt to use a high dielectric constant insulating material above the silicon dioxide layer to subdue the leakage current to ignorable proportions. The problem created by ultra-thin oxide has been tackled by using gate dielectric materials of high dielectric constant k such as zirconium oxide (ZrO_2) with $k = 25$, hafnium oxide (HfO_2) with $k = 30$, etc. For a high k dielectric, an equivalent oxide thickness (EOT) is defined as

$$\text{EOT} = t_{\text{ox}} = t_k \times (3.9/k) \tag{5.11}$$

where t_{ox} is thickness of oxide film and t_k that of the gate dielectric. Because $t_k \gg t_{\text{ox}}$, the gate leakage current is significantly reduced by using such a dielectric.

Beyond 70 nm node, requirement of high-κ materials was essential. They were necessary despite the fact that many of these materials showed poor thermal stability and interface quality with silicon dioxide. Most of these materials are oxides of transition metals. They are deposited on silicon dioxide instead of being thermally grown. The methods commonly used are based on physical and chemical vapor deposition; sol-gel process is also used. Among physical vapor deposition (PVD) methods, thermal evaporation and reactive sputtering stand out prominently. Chemical vapor deposition methods include atomic layer deposition (ALD), plasma-enhanced atomic layer deposition (PEALD), plasma-enhanced chemical vapor deposition (PECVD), metal-organic chemical vapor deposition (MOCVD), and molecular beam epitaxy (MBE). Surface preparation prior to deposition plays a critical role in obtaining desired film quality and adhesion. Post deposition thermal treatments such as sintering or annealing too are decisive factors. Besides oxides, silicate films, notably $ZrSiO_x$ and $HfSiO_x$ are also used.

5.10.3 Strain Engineering for Enhancing Carrier Mobility

One could replace silicon (electron mobility $\leq 1400\,\text{cm}^2\text{V}^{-1}\ \text{s}^{-1}$) with high-mobility semiconductors GaAs (electron mobility $\leq 8500\,\text{cm}^2\text{V}^{-1}\ \text{s}^{-1}$) or InP (electron mobility $\leq 5400\,\text{cm}^2\text{V}^{-1}\ \text{s}^{-1}$). But for these semiconductors, the level of technological maturity for large-scale production of ICs is much lower.

Use of strained silicon increases mobility. In strained silicon, the silicon atoms are pulled apart from their normal positions in the lattice, increasing their interatomic distance by a small amount $\sim 1\%$. Strain engineering is a strategy employed in silicon IC manufacturing to increase the carrier mobility. By virtue of the increase in spacing between the atoms than for regular silicon, the electronic band structure of silicon is modified in such a manner that effective mass of charge carriers in silicon is reduced. Lower is the effective mass higher is the mobility. Also, carriers are diverted to regions of lower effective mass. Effectively, better transport properties are achieved [12].

Strain engineering is done in one of the two ways, either globally or locally. Uniaxial global strain is introduced by bonding the wafer to a substrate with cylindrical surface. Biaxial global strain is generated by epitaxial growth of a thin strained silicon layer on a thick relaxed SiGe virtual substrate.

Epitaxy is the process of growing a single crystal film on a crystalline substrate. In this growth process, the substrate serves as a template according to which the deposited atoms arrange themselves. Without this template, the growth will be either polycrystalline or amorphous. Since Si and Ge have identical crystal structure, a Si overlayer can be grown on a Ge substrate. But Ge atoms are larger in size than Si atoms. Moreover, interatomic spacing is 4.2% greater in Ge than in Si. Due to this reason, direct deposition of Si atoms on Ge lattice results in a high defect concentration in the form of dislocations. A favorable condition is provided by the fact that lattice constant of SiGe alloy defining the unit cell of the lattice has a value between the lattice constants of Si and Ge. So if a Si overlayer is deposited on a SiGe substrate, the silicon atoms attempt to be coincident with SiGe atoms below. As a result, the Si atoms are placed at larger distances than they would be in a normal Si lattice. So, the Si layer grown by epitaxy is stretched slightly as if like a rubber diaphragm from the normal Si lattice. This is equivalent to generation of a strain in the Si lattice similar to the strain in the extended diaphragm. But the crux of the problem is that it is difficult to obtain a SiGe template of high quality on which the epitaxial Si layer could be grown. A way to get out of this situation has been found out (Figs. 5.4 and 5.5). One starts with Si substrate instead of SiGe substrate. A buffer layer is grown over this substrate. This buffer layer contains Si and Ge. To begin with, the concentration of Ge is zero at the bottom of the buffer layer. As thickness increases, the concentration of Ge is slowly raised. The final

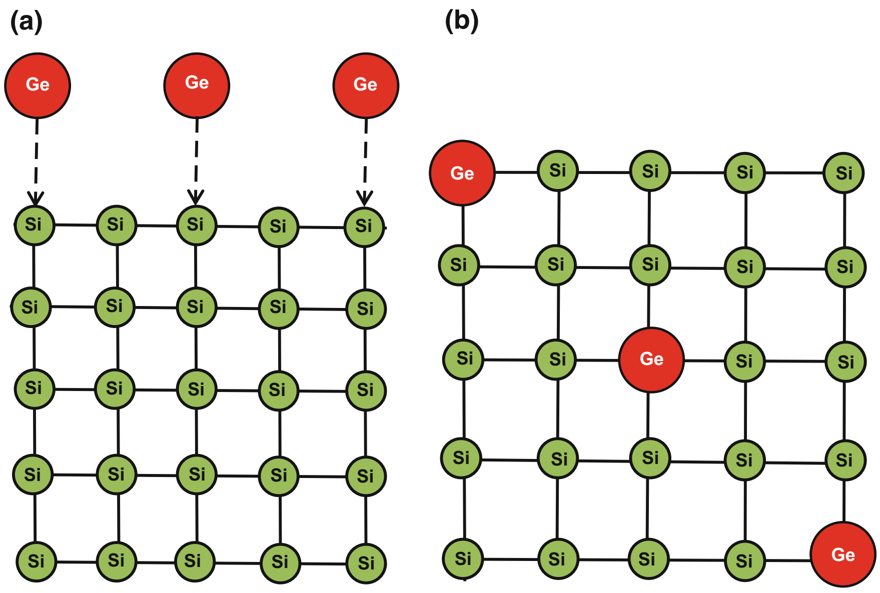


Fig. 5.4 First step in strained silicon process: **a** adding Ge atoms to surface layers of silicon crystal to produce **b** $Si_{1-x}Ge_x$ layer

concentration of Ge at the top of the buffer layer is 20%. After the buffer layer has been grown, a template layer of uniform SiGe concentration is grown. As this template layer maintains the concentration of the buffer layer, it is not strained. It is therefore a relaxed layer. Over this layer, the Si capping layer is grown. This capping layer made of Si only follows the structural arrangement of SiGe template layer below. Obviously, it is a strained layer since the lattice constant of Si differs from that of SiGe. But the lattice constant does not differ to the extent to be able to cause dislocations. Hence, it is said to be optimally strained.

In the 90 and 65 nm modes, process-induced stress generation was extensively applied: compressive stress for PMOS (Fig. 5.6) and tensile stress for NMOS (Fig. 5.7). Local compressive stress is created by selectively growing a thin epitaxial film of SiGe in the source/drain regions of P-channel MOSFET.

In case of N-channel MOSFET, growth of SiC film develops local tensile stress. Other local stress generation techniques include the formation of shallow trench isolation or a stressed silicon nitride capping layer [13].

Use of silicon nitride as a strain-inducing capping layer (Fig. 5.8) offers the flexibility of controlling the type of strain (compressive or tensile) as well as the degree of strain by proper choice of deposition conditions, mainly temperature. This approach is known as dual-stress liner method [14]. In a CMOS process, after the self-aligned silicide step, a tensile silicon nitride layer is deposited over the complete wafer. Photolithography is performed to selectively remove the silicon nitride from regions where P-channel MOSFET is located. Thereafter, a compressive

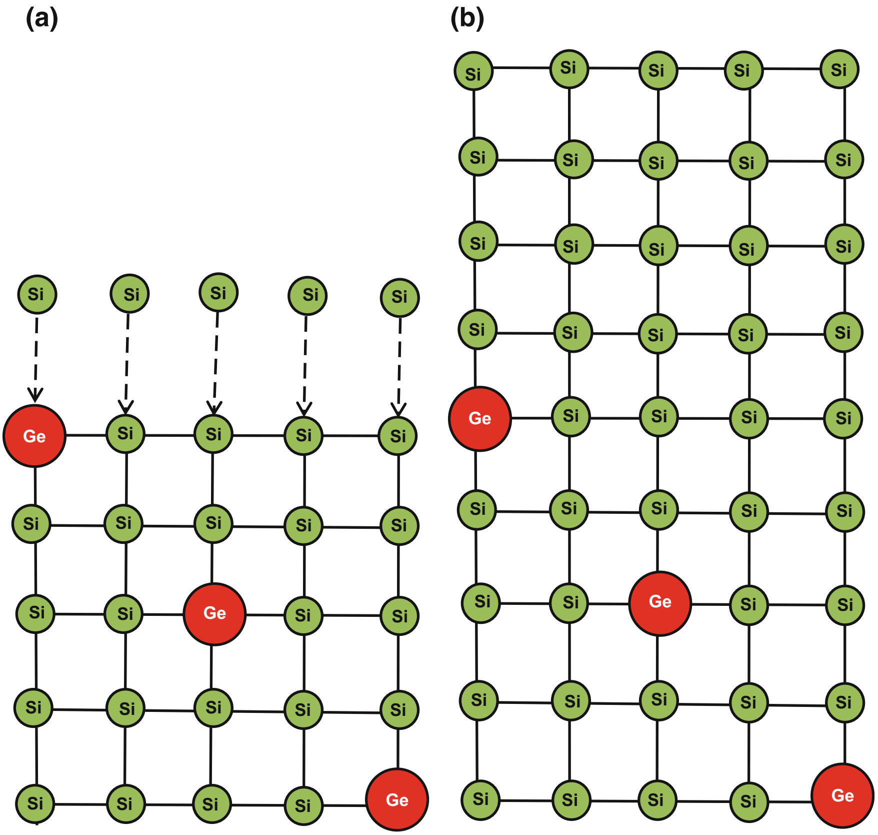


Fig. 5.5 Second step in strained silicon process: **a** epitaxial growth of silicon on $Si_{1-x}Ge_x$ layer, and **b** resulting strained silicon

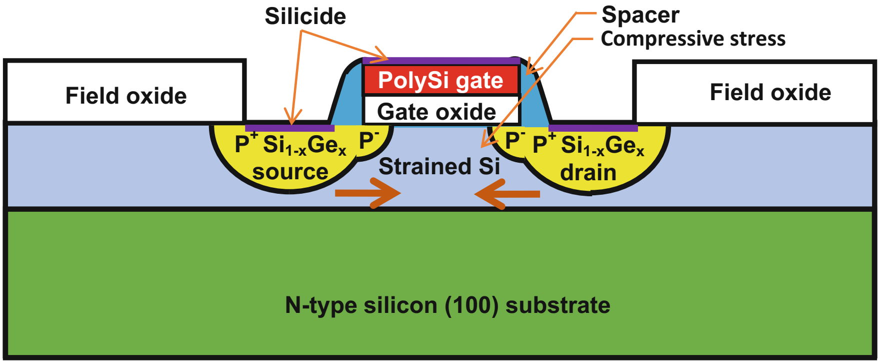


Fig. 5.6 Compressive strain induction in the silicon channel region of a P-channel MOSFET by $Si_{1\text{-}x}Ge_x$ source/drain regions

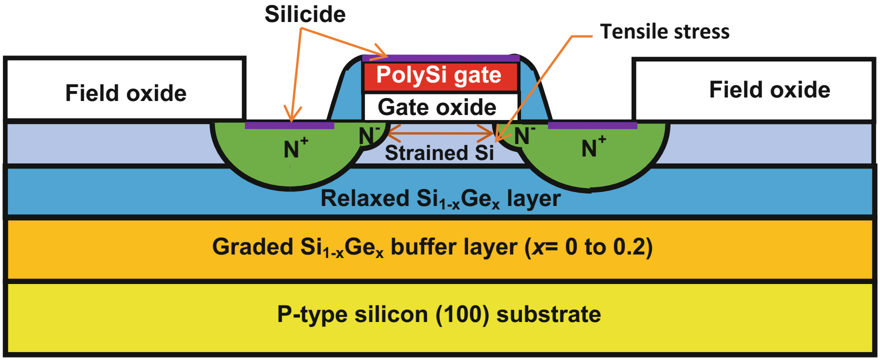


Fig. 5.7 Tensile strain generation in the silicon channel region of an N-channel MOSFET; the channel region is grown epitaxially on a $Si_{1-x}Ge_x$ layer over a P-silicon substrate

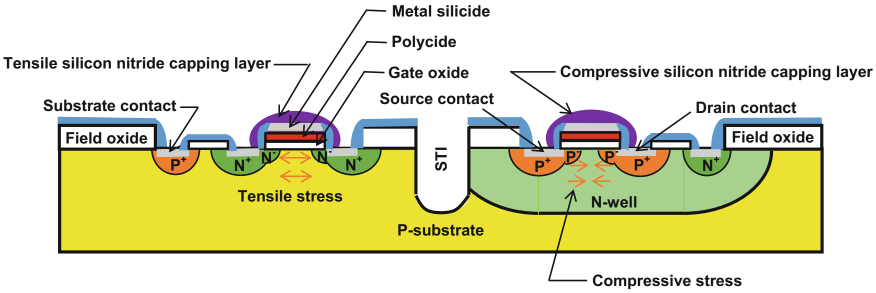


Fig. 5.8 Strained silicon produced by two types of silicon nitride stress liners formed under different deposition conditions

silicon nitride layer is deposited over the full wafer and etched from N-channel MOSFET regions. A significant increase in drive current was reported by simultaneous action of two different strain layers in the CMOS process flow without any SiGe layer.

5.10.4 *Minimization of Hot Carrier Effects*

Different techniques have been proposed for reduction of hot carrier effects, such as

(i) **Gate Oxide Thickness Reduction** By using a thinner oxide, the point of peak of electron injection is shifted to a greater extent toward the drain region. Hence, the stretch of damaged region overlying the channel is smaller in size.

(ii) **Lightly Doped Drain (LDD) Structure** This structure consists of two doping concentrations, one high and another low. Light doping, shallow implant is done in a region abutting the channel, and therefore establishing contact with it. Small overlap of the gate with source or drain regions produces minimum overlap capacitance. Further, the reduced carrier concentration at the drain edge decreases the field between drain and channel regions. The field thus lowered brings down the amount of carrier injection into the oxide, impact ionization and other related effects. The heavy doping, deep implant is done after depositing silicon nitride sidewall spacers on both sides of the polysilicon gate. It covers most of the source and drain areas, creating low series resistance with the channel. Thus the combination of two implants not only provides small access resistance and overlap capacitance but also minimizes hot carrier injection. Figure 5.9 shows the main steps of this process.

Elaborating the mechanism of the LDD structure for hot carrier inhibition, it essentially performs a kind of drain engineering. In this drain engineering, the peak of the lateral electric field located near the edge of the drain is moderated and weakened by modification of doping profile through low dose implantation in the concerned regions. As the lightly doped regions created by low dose implantation look like extensions of actual heavily doped source and drain, they are referred to as source/drain extensions.

(iii) **Double-Diffused MOSFET Structure** It has deeper N-phosphorous profile than N^+As profile. The outcome is that the path of maximum current is located away from the position of the peak field. This helps to reduce the impact ionization, and thereby hot carrier generation.

(iv) **Incorporating Si_3N_4 as the Gate Oxide** The Si–N bonds require more energy to break than Si–H bonds.

(v) **Deuterium Post-metal Annealing** Post-metallization annealing in hydrogen ambient at a low temperature can reduce Si–SiO_2 interface traps.

5.10.5 *Preventing Punch-Through*

It can be avoided by increasing the channel doping. The higher doping shortens the source and drain depletion regions. However, higher channel doping decreases the mobility by enhanced scattering. Therefore, it will solve the punch-through problem at the expense of mobility reduction and consequent decrease in on-state current.

Halo or pocket implants (Fig. 5.10) are implants used to suppress punch-through between source and drain through the substrate. Hence, they are often called punch-through suppression implants. To achieve this objective without mobility degradation, the dopants are placed a little below the channel adjoining the source and drain regions. In this way, they can accurately adjust the background doping concentration of the well in the intended location. At the same time, carrier mobility is not decreased.

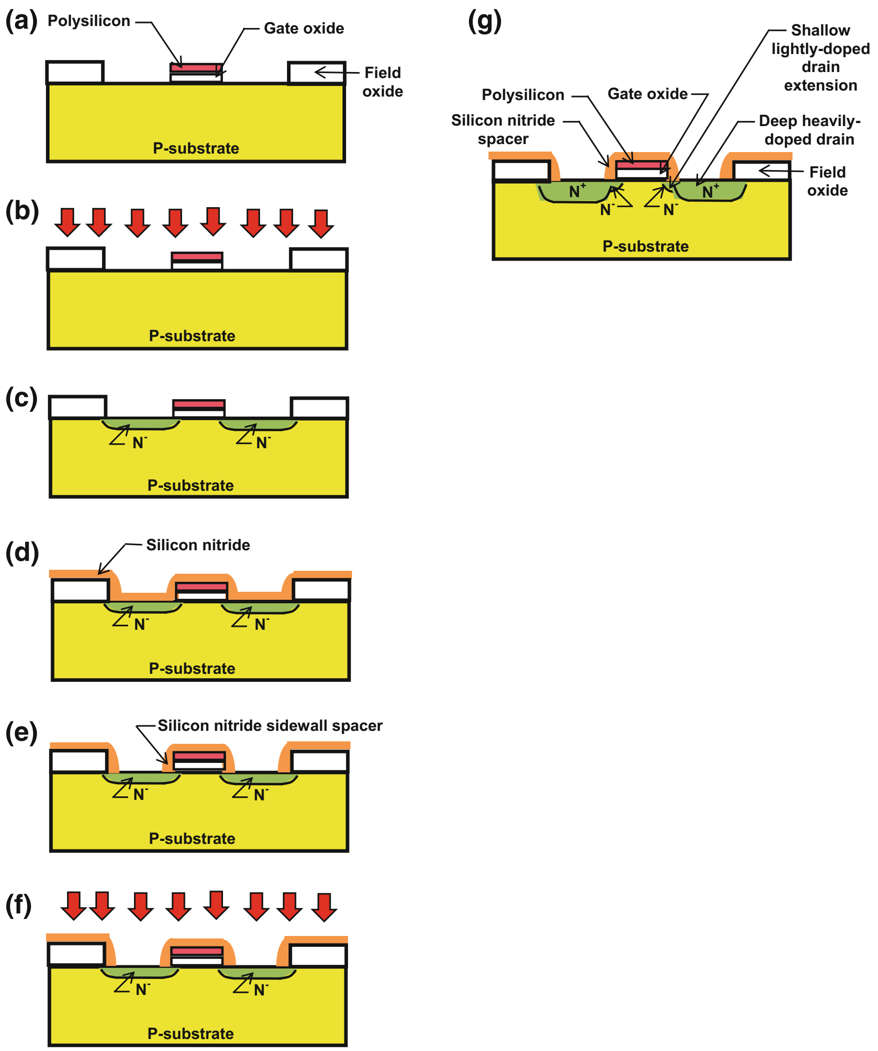


Fig. 5.9 Self-aligned process for fabrication of lightly doped source/drain structure. **a** Gate oxide and polysilicon etching. **b** N-source/drain extension implant. **c** High-temperature annealing. **d** Nitridation for sidewall spacer formation. **e** Nitride etching. **f** N^+ source/drain implant. **g** High-temperature annealing

Another strategy for avoidance of punch-through employs the super steep retrograde well (SSRW) and a thin intrinsic region in the channel region of the MOSFET. “Retrograde” means backwards. In a retrograde well, the doping concentration is lowest in the region near the gate insulator. It increases as one moves far away from the gate insulator and the channel region deep into the P-substrate. By doing so, the mobility in low-concentration zone close to the gate insulator is

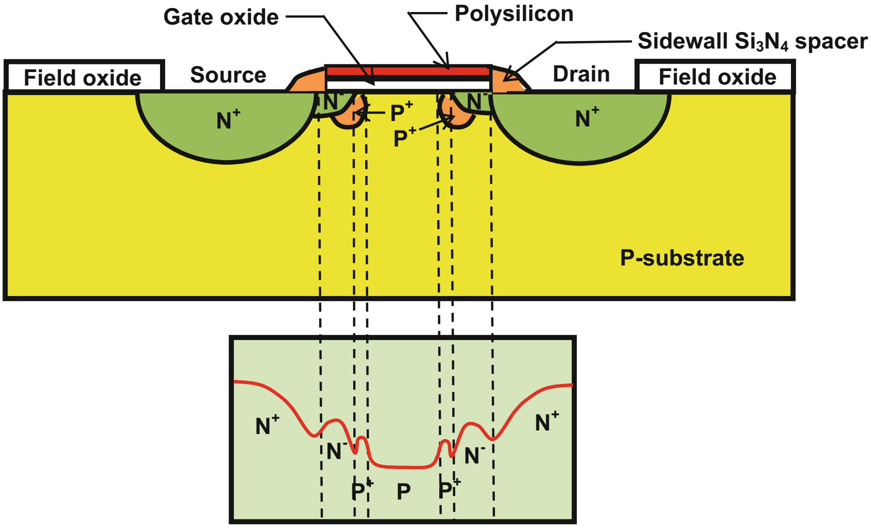


Fig. 5.10 Cross-sectional diagram of N-channel MOSFET with lightly doped drain structure and halo implant. Doping profile at the surface of the device is shown below the diagram

preserved at a high value. This happens because ionized impurity scattering in the zone is reduced. Hence, on-state current is increased. Also, the high-concentration zone buried deep into the substrate counteracts the substrate punch-through. The off-state current is minimized because the depletion depths on both sides of the channel are smaller due to the high doping. As a result, space charge generation current and thereby off-state current are diminished.

5.10.6 Innovative Structures Superseding Classical MOSFET

Evidently, short-channel effects are a manifestation of loss of control over the channel by the gate electrode due to interfering electric fields. The channel should be under tighter control of the gate. Then only such disturbances can be barred from exerting their influence.

5.11 Discussion and Conclusions

With strained silicon using SiGe, bulk MOSFET crossed the 100 nm signpost, reaching 90 nm technology node in 2003. Second generation SiGe led to 65 nm node devices in 2005. Metal gate and high-κ dielectric with gate-last approach provided 45 nm node in 2007 and 32 nm node in 2009 in second generation devices. Gate length is 70% less than the technology node, e.g., for 90 nm node, the gate length is <63 nm, … for 32 nm node, it is <22.4 nm. An important observation is that in all these efforts, either by using SiGe or other strain-inducing layers, or by reverting to metal gates with high-κ dielectrics, enhancements in properties of materials helped us a great deal in treading the Moore's curve. However, no major structural breakthroughs were necessary.

Review Exercises

5.1 When is a MOSFET channel said to be: (i) long, and (ii) short? Can a short channel be avoided during miniaturization of MOSFET?

5.2 What is polysilicon depletion effect? Up to what extent doping elevation can obviate this effect? Can it be avoided by using metal gates?

5.3 Do source and drain junctions play any role in depleting the MOSFET substrate? In what way does this contribution of source and drain junctions change in a short-channel device? What is threshold voltage roll-off?

5.4 Why is the threshold voltage of a long-channel MOSFET independent of drain voltage but this is not so in a short-channel device?

5.5 Explain drain-induced barrier lowering in a MOSFET. What is the effect on threshold voltage of the device?

5.6 How does the threshold voltage reduction in a short-channel MOSFET differ for source/substrate and drain/substrate junctions, which are: (i) deep, and (ii) shallow?

5.7 At high drain voltages, the source and drain junctions may touch each other. What is this phenomenon called? How is the voltage at which this phenomenon occurs related with the channel length of MOSFET?

5.8 What is velocity saturation? What is its effect on the current drive of a MOSFET?

5.9 Write the Caughey-Thomas equations for the dependence of mobility on electric field.

5.10 How are the contributions of bulk and surface effects in mobility of carriers in a MOSFET inversion layer expressed by Matthiesen's rule?

5.11 How is impact ionization produced in a MOSFET at high drain voltages? How the parasitic bipolar transistor makes the situation worse?

5.12 What is a hot carrier? Describe the different hot carrier effects in a MOSFET.

5.13 How is leakage through gate oxide avoided by using a high-κ dielectric? What is equivalent oxide thickness?

5.14 Write the equation relating the equivalent oxide thickness with thickness of high-k dielectric.

5.15 How does straining the silicon lattice increase the carrier mobility? Briefly explain the dual-stress liner approach.

5.16 How do the following structures help in avoiding hot carrier effects? (i) light-doped drain, and (ii) double-diffused MOSFET?

References

1. Maszara WP (2005) Fully silicided metal gates for high-performance CMOS technology: A Review. J Electrochem Soc 152(7):G550–G555
2. Kittl JA, Lauwers A, Mv Dal et al (2006) Ni, Pt and Yb based fully silicided (FUSI) gates for scaled CMOS technologies. ECS Trans 3(2):233–246
3. Moyer B (2011) Gate first vs last: a summary of the issue now that things should have settled down. Electronic Engineering Journal © 2003–2015 techfocus media, Inc. http://www.eejournal.com/archives/articles/20111114-gate/. Accessed 8 Oct 2015
4. Hoffmann TY (2015) Integrating high-κ/metal gates: gate-first or gate-last? Solid State Technology ©2015 Extension Media, http://electroiq.com/blog/2010/03/integrating-high-k/. Accessed 8 Oct 2015
5. Caughey DM, Thomas RE (1967) Carrier mobilities in silicon empirically related to doping and field. Proc IEEE 55:2192–2193
6. Darwish MN, Lentz JL, Pinto MR et al (1997) An improved electron and hole mobility model for general purpose device simulation. IEEE Trans Electron Devices 44(9):1529–1538
7. Klaassen DBM (1992) A unified mobility model for device simulation—Part I: model equations and concentration dependence. Solid State Electron 35(7):953–959
8. Lombardi C, Manzini S, Saporito A et al (1988) A physically based mobility model for numerical simulation of nonplanar devices. IEEE Trans Comput Aided Des 7:1164–1171
9. Hartstein A, Ning TH, Fowler AB (1976) Electron scattering in silicon inversion layers by oxide and surface roughness. Surf Sci 58:178–181
10. Silicon Dioxide Properties. http://www.iue.tuwien.ac.at/phd/filipovic/node26.html. Accessed 8 Oct 2015
11. Misra D, Iwai H, Wong H (2005) High-κ gate dielectrics. Electrochem Soc Interface Summer 2005:30–34
12. Intel: Strained Transistors. http://www.intel.com/pressroom/kits/advancedtech/doodle/ref_strain/strain.htm. Accessed 30 Aug 2015
13. Ungersboeck E, Sverdlov V, Kosina H et al (2006) Strain engineering for CMOS devices. In: 8th international conference on solid-state and integrated circuit technology (ICSICT'06), Shanghai, Oct 2006, pp 124–127
14. Yang HS, Malik R, Narasimha S et al (2004) Dual Stress liner for high performance sub-45 nm gate length SO1 CMOS manufacturing. In: IEEE international electron devices meeting, IEDM technical digest, 13–15 Dec 2004, pp 1075–1077