SHORT CHANNEL EFFECTS IN MOSFETs

In this section, we look at several short-channel effects of the MOSFET. These include mobility degradation, velocity saturation, drain-induced barrier lowering (DIBL). We also look at transistor scaling theory.

1.1 Drain-Induced Barrier Lowering (DIBL)

For long-channel devices, the source-channel potential barrier is determined primarily by the voltage applied to the gate. Increasing the gate voltage causes the barrier height to reduce, resulting in injection of electrons from the source into the channel. At a particular gate voltage – the threshold voltage – the barrier has reduced sufficiently to allow a significant amount of injection, and flow of current, to take place.

In a short-channel device, the drain junction is now quite close to the source junction. As a consequence, the potential at the source-channel region is determined not only by the gate voltage, but also the drain voltage. The drain voltage can cause a lowering of the barrier at the source end of the channel, causing current to flow for a lower vale of gate voltage. This effect is called *Drain-Induced Barrier Lowering* or DIBL (pronounced "dibble").

The magnitude of DIBL obviously depends on the channel length L, but also on the doping NA. Higher doping's would mean that electric field from the drain is more effectively screened, resulting in lower DIBL. Another way to interpret DIBL is that the potential lowering is due to merging of the edges of the depletion regions emanating from the source and drain junctions. Higher doping's reduce the depletion widths, and consequently delay their merging, reducing the DIBL effect. Thus, increasing the bulk doping is a good strategy to minimize DIBL.

1.1.1 Effect on threshold voltage

One consequence of DIBL is an apparent reduction of the threshold voltage. This is clear physically, because with DIBL present, a smaller bias at the gate (threshold voltage) is reduce the barrier sufficiently to allow current to flow. since larger drain voltages result in increased barrier lowering, the threshold voltage keeps decreasing with increasing VDS, that is, VT is now a function of VDS, which was not the case for long channel devices.

An approximate empirical relationship used for ΔV_T due to DIBL is

$$\Delta V_T^{DIBL} = -\sigma V_{DS}\,, \tag{6.9}$$

where σ is called the DIBL factor, and is usually expressed in units of mV/V. Typical values of σ for a modern transistor fall in the range of 10 -100 mV/V. The DIBL factor is, of course, a strong function of L, and is often modeled empirically as

$$\sigma = \sigma_0 \exp(-L/L_0)\,, \tag{6.10}$$

where L_0 is a characteristic length. Note that Eq. (6.9) implies that the effect of DIBL is only present for $V_{DS} \neq 0$, whereas in fact the barrier and threshold voltage are lowered

1.1.2 Effect on Subthreshold Behaviour

The coupling of the drain field to the source-channel junction affects how current flows in the subthreshold region for a short-channel device. As drain voltage increases, the barrier reduces due to

DIBL, and a larger drain current flows. This behaviour is very different from what is seen for the long-channel device, where the subthreshold current is independent of drain voltage.

Parameters affecting sub-threshold regime

Recalling the equation which describes the drain current in the subthreshold regime

$$I_D = I_{pf} \exp\left[\frac{\beta(V_{GS} - V_T)}{\eta}\right],$$

where I_{pf} is the pre-exponential factor and $\eta = [1+ (C_D /C_{ox})]$. The first point to note is that I_D is larger for the short-channel case than the long-channel case because V_T is less for the former because of both charge-sharing and DIBL. Secondly, the V_{DS} dependence of I_D in the case of the short-channel device arises because V_T is now dependent on V_{DS} through Eq. (6.9). However, the effect of DIBL not only translates into a change of V_T, but also, in addition, changes the subthreshold slope. The reason for this is that due to proximity of the drain, the gate is less in control of current flow compared to the long-channel case. The gate has ceded some of the control to the drain. As a consequence, changing the gate voltage does not affect the current flow as strongly as it does for the long-channel device, and the subthreshold fall-off is not as tight. In other words, the

1.2 Velocity Saturation

As the MOS device is scaled down, electric fields in the device become large. This creates a significant problem. It is true that the voltages used have also scaled down from 5 V to about 1 V in an attempt to keep the field within limits, but with the corresponding channel lengths decreasing from 5 μm to 70 nm, and gate oxide thickness from 100 nm to 1.5 nm, it is clear that the electric fields, both horizontal and vertical, have increased significantly in short-channel MOSFETs. There are several consequences of the high electric fields. One of the most important is velocity saturation experienced by the carriers as they move along the channel in the presence of a high lateral electric field.

It is well-known that the drift velocity of carriers in a semiconductor does not continue to increase linearly with electric field at higher fields. On the contrary, the velocity, plotted as a function of field, starts to level off, and finally saturates at a value v_{sat}. For electrons and holes in silicon, the value of v_{sat} is about $1\text{x}10^7$ cm/sec at room temperature. Several models have been used for drift velocity v as a function of electric field E. One of the most common is

$$v = \frac{\mu_n E_y}{\left\{1 + \left(E_y / E_c\right)^n\right\}^{1/n}}. \qquad (6.20)$$

In the above equation, μ_n is the low-field electron mobility, E_c ($= v_{sat}/\mu_n$) is the critical field determining velocity saturation, and E_y is the electric field in the direction of motion of the electrons (the lateral y direction in case of the MOSFET). Eq. (6.20) correctly yields $v = \mu_n E_y$ for low fields, and $v = v_{sat}$ for high fields. The value of n typically lies between 1 and 2, and is chosen so as to match experimental data most closely.

For very small values of L, the value of V_{Dsat} is almost zero, from Eq. (6.36). This is understandable, since it says that the velocity is saturated in the channel even for small drain voltages. The drain current I_D is then always saturated, and is given, from Eq. (6.35) approximately by

$$I_{Dsat} = WC_{ox}v_{sat}[V_{GS} - V_T] . \quad (6.37)$$

The drain current is independent of L, that is, no matter how small L is made, the current does not increase beyond a point as velocity saturation finally limits the current. This implies that decreasing L will nor improve performance (higher currents and speeds) in future technologies as it has done so dramatically in the past (see Fig. 6.9). Further, note that drain current is linearly, not quadratically, dependent on $(V_{GS} - V_T)$ in saturation for the ultra-short channel transistor.

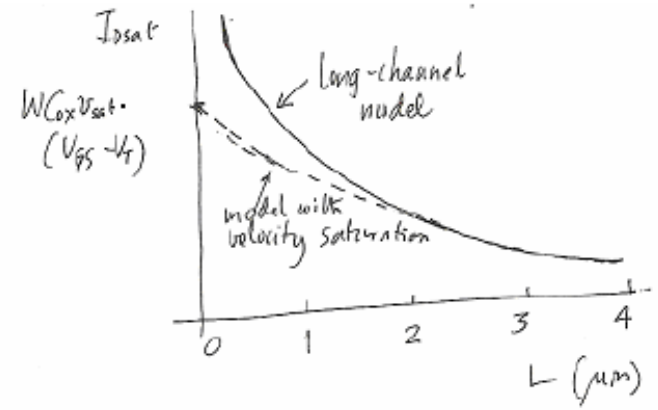


Fig. 6.9 Plot of I_{Dsat} versus channel length L. Note that, unlike the long-channel model which predicts ever-increasing I_{Dsat} as L reduces, the model with velocity saturation shows only a finite increase.

1.3 Mobility Degradation

Velocity saturation, which we explored in the previous section, is equivalent to a reduction of mobility at high electric fields in the direction of motion of the carriers. In addition to this mobility degradation, there also exists a reduction of mobility due to transverse electric fields, that is, due to the field emanating from the gate. This is a "short-channel" effect only in as much the vertical fields increase significantly as the oxide thickness scales down together with channel lengths.

Physically, the reason that the transverse fields affect lateral current flow is as follows. Carriers in a MOSFET flow in the thin inversion layer near the surface of the silicon. They undergo frequent scattering events with the surface (particularly when it is microscopically "rough"), as well as additional coulombic scattering due to interface states and charges in the insulator close to the interface. These effects conspire to make the so-called "inversion-layer" or surface mobility less than the bulk semiconductor mobility (typically about half). Now as the device scales down, the increased vertical fields pull the carriers towards the surface more strongly,

where they undergo more frequent scattering. This increased scattering reduces the mobility of carriers further, and the larger the transverse field the lower is the mobility.

It has been found experimentally that the mobility for electrons as well as holes in silicon can be written as

$$\mu = \frac{\mu_0}{1 + (\theta E_{eff})^{\nu}}, \tag{6.38}$$

where E_{eff} is an effective transverse electric field described below, μ_0 is the mobility without the transverse field effect, and ν is a factor between 1 and 2. The above form of the equation was first proposed by Sabnis and Clemens [6.7], and this or similar forms have been observed by many researchers since. Note that the above form of the equation has a physical origin, and can be derived using Mathiessen's rule. Let μ_0 be the mobility of channel carriers for low transverse fields, and let μ_T be the mobility due to extra scattering caused by the vertical field. Assuming that the latter scattering increases with effective field as E_{eff}^{ν}, we can write

$$\frac{1}{\mu} = \frac{1}{\mu_0} + \frac{1}{\mu_T} = \frac{1}{\mu_0} + KE_{eff}^{\nu}, \tag{6.39}$$

which gives the desired equation for μ.

A plot of I_D versus V_{GS} (on a linear scale) for small values of V_{DS} (the so-called transfer characteristic), which is often used to find the threshold voltage is now affected by the mobility degradation. Instead of a straight line with intercept V_T and slope μ_n, we now see a curve gently drooping downwards for increasing V_{GS} due to mobility degradation (Fig. 6.11). This is very commonly seen in measurements, and the variation of slope with V_{GS} can be used to estimate μ_{eff} as a function of $(V_{GS} - V_T)$. The measurement of threshold voltage is usually done by finding the intercept of the tangent at the point of highest slope (under the assumption that the device is here sufficiently into the strong inversion region, but not so strong as to entail significant mobility reduction). Note that as described in Section 5.6, series resistance also degrades the transfer curve, and both these effects must be taken into account in general.

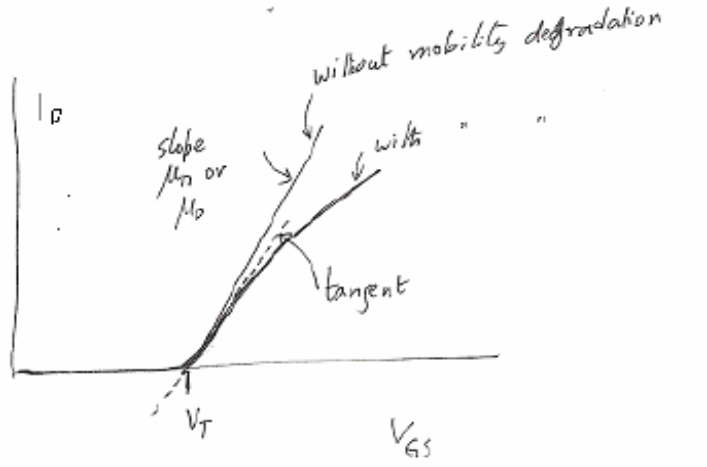


Fig. 6.11 Degradation of the transfer characteristic due to mobilitydegradation.

1.4 Hot Carrier Effects and Impact Ionization

As channel lengths reduce, the lateral electric field increases, if applied voltages remain the same. This causes carriers flowing along the channel to gain energy and become "hot". The hot carriers can cause impact ionization, which produces extra hole electron pairs. This produces extra drain current, and also a substrate current consisting of the ionization-generated holes which flow towards the substrate, the most negative point in the transistor. The hot carriers may also gain sufficient energy to surmount the potential barrier at the silicon-insulator interface, and get injected into the insulator. These effects are shown schematically in Fig. 6.14.

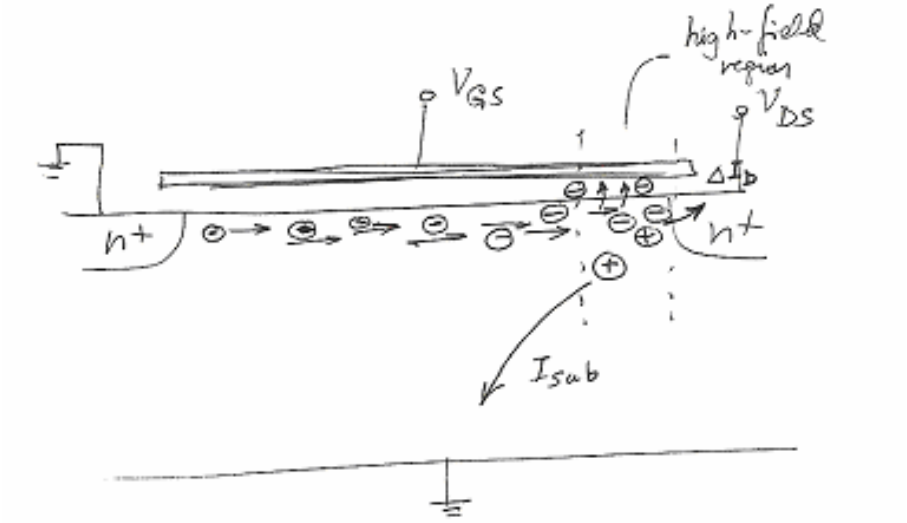


Fig. 6.14 Schematic figure showing effects of impact ionization (a) increase in drain current ΔI_D (b)increase in substrate current I_{sub} and (c) injection of hot carriers into the oxide.

The effects of hot carriers and impact ionization on MOSFET characteristics and performance are discussed in the sections below.

1.4.1 Increase in Output Conductance

The electrons generated by impact ionization contribute to extra drain current. In saturation, as the voltage V_{DS} increases, and field in part of the MOSFET increases. This increased field causes impact ionization, and results in increased drain current as well as substrate current

1.4.2 Hot Carrier Effects

Due to the high electric field in the channel, electrons, which constitute the drain current in n-channel transistors, can get hot. If some of these electrons get sufficiently heated up – more than the 3.1 eV barrier Φ_{ox} between silicon and silicon dioxide conduction bands – and also have their momentum directed (through an elastic collision) towards the interface, then these electrons can get injected into the insulator.

Furthermore, some of the holes created by impact ionization may also get heated up and be injected into insulator (though this process is more difficult for holes than for electrons, given the larger 4.9 eV barrier for holes at the Si/SiO_2 interface). The holes and electrons flowing into the insulator cause several problems, including electron and hole trapping, interface state generation, and generation of bulk and "border" traps in the insulator. These phenomena,

shown schematically in Fig. 6.17, are collectively called “hot carrier effects”, and pose an important reliability issue for MOSFETs.

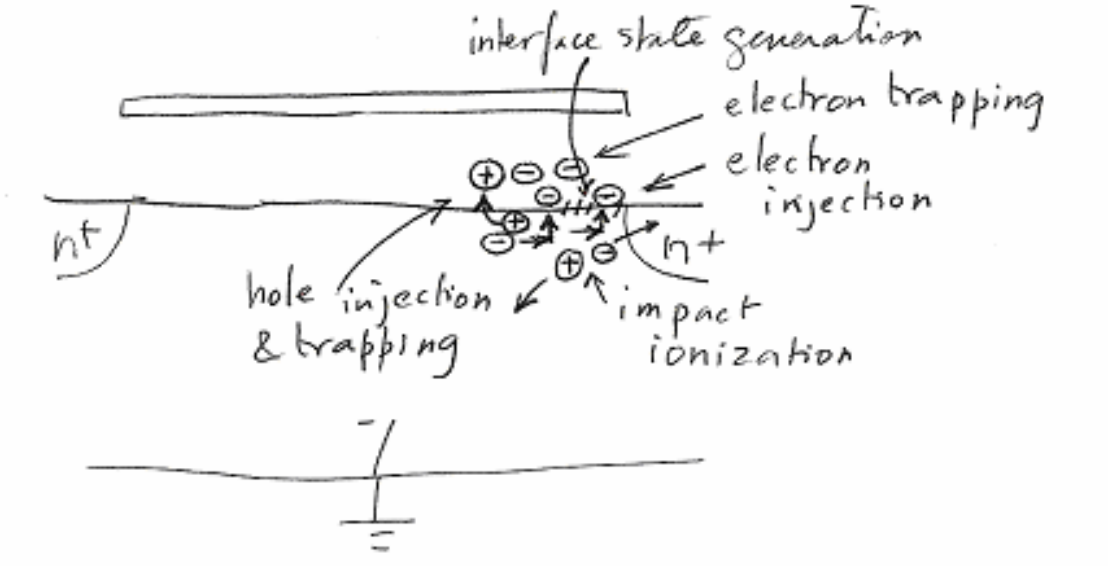


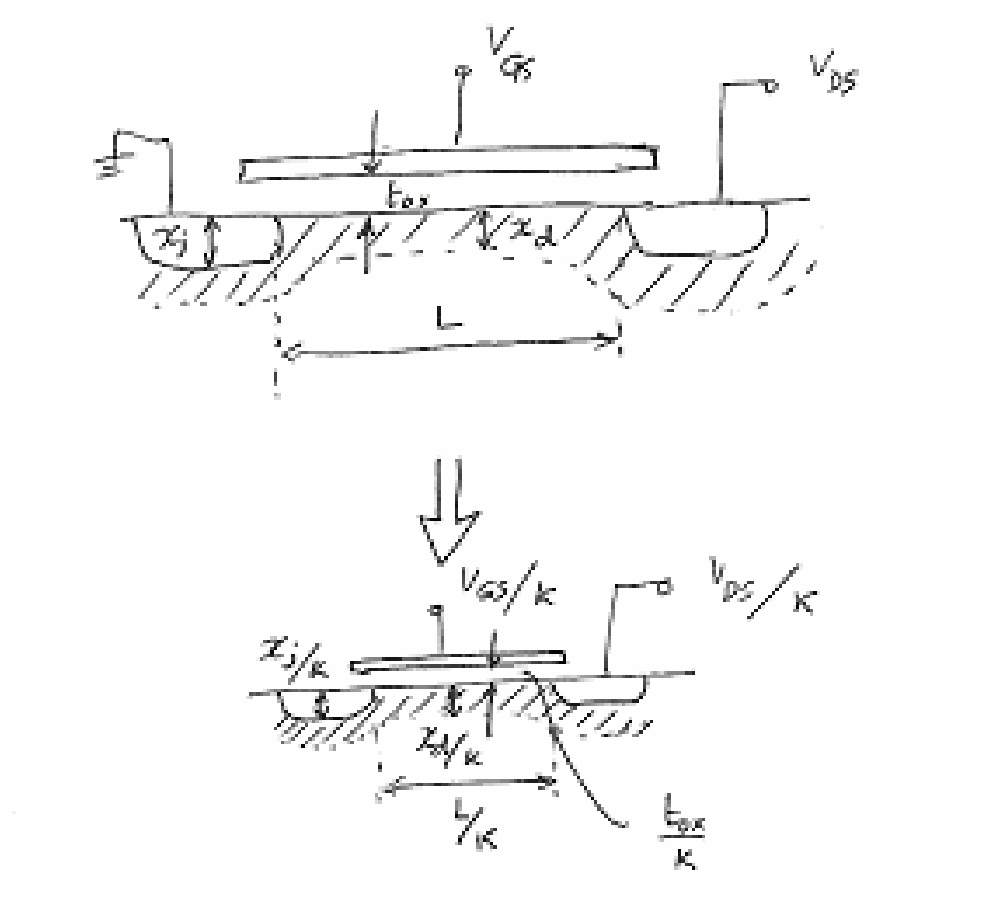
Fig. 6.17 Schematic representation of hot carrier effects in a MOSFET, showing impact ionization, injection of holes and electrons into the oxide, and subsequent trapping, and interface state generation.

Some of the deleterious effects caused by hot carrier effects on MOSFET performance include: threshold voltage shift, decrease in transconductance, excess leakage currents, instabilities and excess noise.

Transistor Scaling

As transistor sizes were scaled down from the tens of microns in the 1970’s to submicron dimensions in the 1990’s, a generic “scaling theory” was developed that helped the device designers to shrink the devices. The scaling theory laid down broad guidelines of how the various parameters in a scaled transistor should change, and what the consequences would be. Moving into the sub-100 nm regime, the general scaling theory is less useful than in previous years; however, it is still very instructive to look at it. Constant-Field Scaling

The basis of constant-field scaling, proposed by Dennard [6.21] in 1974, was that as the device dimensions scaled down, the electric field in the device should remain constant, so as not to aggravate reliability problems. Fig. 6.24 shows Dennard’s scaling, in which all physical dimensions of the device – W, L, t_{ox} and x_j – are scaled by a factor $(1/\kappa)$, where $\kappa > 1$. In order to keep electric fields constant, all voltages are also scaled by a factor $(1/\kappa)$.



Dennard scaling of the MOSFET – all dimensions and voltages are scaled by a factor (1/K).

It is desirable to ensure that depletion width x_d also scales down by a factor (1/κ), else the depletion regions will start to envelop the whole device. Using the equation for depletion width,

$$x_d = \sqrt{\frac{2\varepsilon_s(\psi_{bi}+V_X)}{qN_A}}, \qquad (6.64)$$

where V_X is the potential at any point along the channel, and assuming that V_X, which scales as $(1/\kappa)$, is much greater than the built-in voltage ψ_{bi}, we can see that in order for x_d to scale as $(1/\kappa)$, N_A should scale as κ. This tells us that the doping should increase, which will also reduce short-channel effects like DIBL and punch-through.

Obviously, the major advantage of scaling is that the complexity of the circuit (number of transistors on a chip) increases, in this case by a factor κ^2. Some of the other effects of constant-field scaling on the on drive current, capacitance, as well as delay time ($\sim CV/I$), dynamic power dissipation, power-delay product, and "functional throughput" (speed × complexity) are shown in Table 6.1. It can be seen that many useful advantages accrue out of this type of scaling: for example, speed improves by a factor κ, power-delay product for a transistor improves by a factor $(1/\kappa^3)$, and the functional throughput, which measures how powerful a chip made with such a technology is, improves as κ^3, all while keeping the (dynamic) chip power constant.

	Parameter	Scaling Factor
Dimensions	Physical dimensions (W, L, t_{ox}, x_j)	$(1/\kappa)$
	Depletion width (x_d)	$(1/\kappa)$
Transistor Electrical Parameters	Doping (N_A)	κ
	Electric field (E)	1
	Voltage (V)	$(1/\kappa)$
Transistor/Chip Performance	On drive current (I_{ON})	$(1/\kappa)$
	Capacitance (C)	$(1/\kappa)$
	Delay time (τ)	$(1/\kappa)$
	Power dissipation ($P = VI$)	$(1/\kappa^2)$
	Chip complexity (# of transistors)	κ^2
	Functional throughput (Complexity × speed)	κ^3
	Power density	1

Some of the problems which the constant-field scaling suffers from are (1) threshold voltage does not scale well, that is, it is not possible to scale down V_T easily, (2) the subthreshold current does not scale as (1/κ), and becomes larger in relation to the on drive current, (3) subthreshold slope (determined by kT/q) does not scale, and (4) it is difficult to scale voltages down as drastically as dimensions, since circuit voltages would soon reach the order of (kT/q).

In fact, constant field scaling is somewhat conservative since it is possible to allow the electric field to increase somewhat. This has particularly been true as the reliability of integrated circuits has improved over the years. Further, there has always been resistance to reducing the operating voltage too rapidly, partly because of circuit legacy requirements.