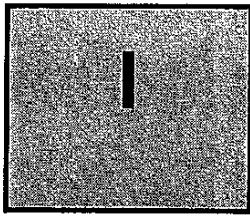


PART

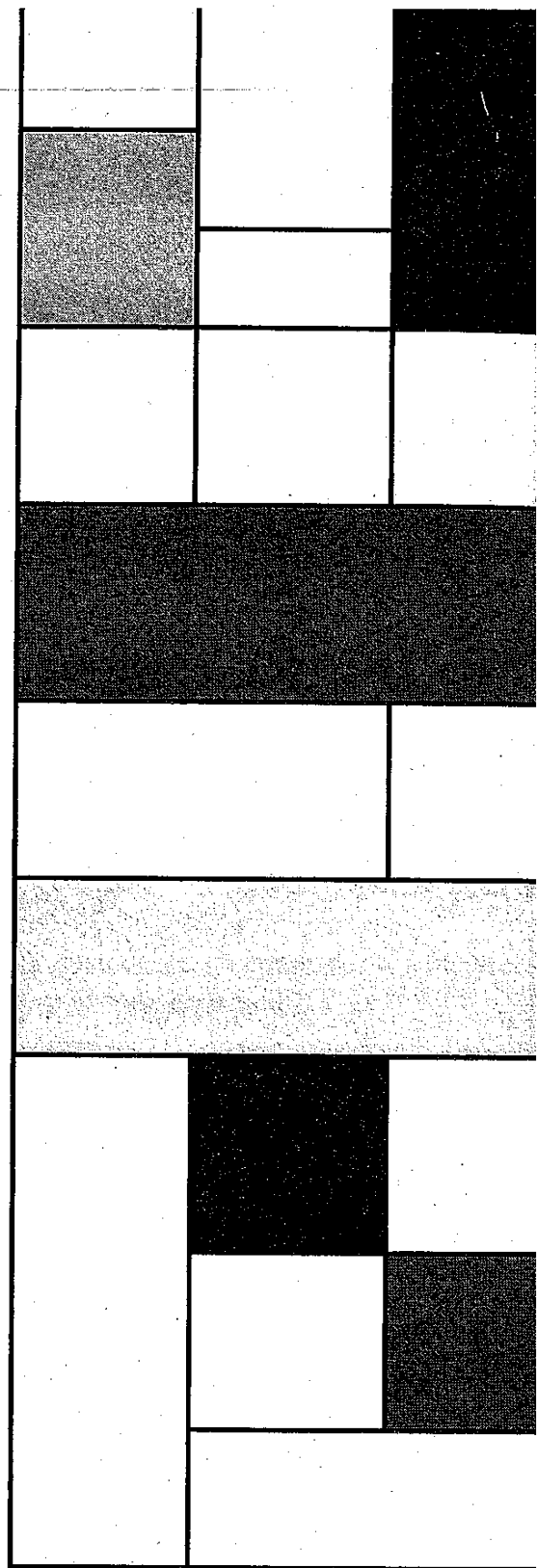


# A CIRCUIT PERSPECTIVE

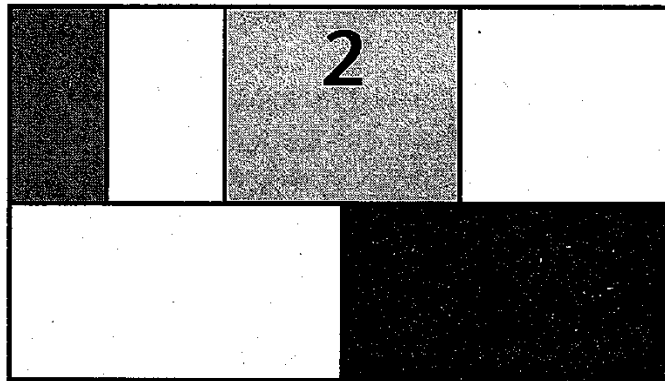
*Philosophy is written  
in this grand book  
—I mean the universe—  
which stands continually open  
to our gaze,  
but it cannot be understood  
unless one first learns  
to comprehend the language  
and interpret the characters  
in which it is written.*

*It is written in the language of  
mathematics, and its characters are  
triangles, circles and other geometrical  
figures, without which  
it is humanly impossible  
to understand a single word of it.*

*Galileo Galilei  
Il Saggiatore, 1623*



# CHAPTER



# THE DEVICES

*Qualitative understanding of MOS and bipolar device operation*

■  
*Simple device models for manual analysis*

■  
*Detailed device models for SPICE*

■  
*Impact of process variations*

- 2.1 Introduction
- 2.2 The Diode
  - 2.2.1 A First Glance at the Device
  - 2.2.2 Static Behavior
  - 2.2.3 Dynamic, or Transient, Behavior
  - 2.2.4 The Actual Diode—Secondary Effects
  - 2.2.5 The SPICE Diode Model
- 2.3 The MOS(FET) Transistor
  - 2.3.1 A First Glance at the Device
  - 2.3.2 Static Behavior
  - 2.3.3 Dynamic Behavior
  - 2.3.4 The Actual MOS Transistor—Secondary Effects
  - 2.3.5 SPICE Models for the MOS Transistor
- 2.4 The Bipolar Transistor
  - 2.4.1 A First Glance at the Device
  - 2.4.2 Static Behavior
  - 2.4.3 Dynamic Behavior
  - 2.4.4 The Actual Bipolar Transistor—Secondary Effects
  - 2.4.5 SPICE Models for the Bipolar Transistor
- 2.5 A Word on Process Variations
- 2.6 Perspective: Future Device Developments

## 2.1 Introduction

It is a well-known premise in engineering that the conception of a complex construction without a prior understanding of the underlying building blocks is a sure road to failure. This surely holds for digital circuit design as well. The basic building blocks in this engineering domain are the silicon semiconductor devices, more specifically the diodes, and the MOS and bipolar transistors.

Giving the reader the necessary *knowledge and understanding of these devices* is the prime motivation for this chapter. It is not our intention to present an in-depth discussion (we assume that the reader has some prior familiarity with electronic devices). The goal is rather to refresh the memory, to introduce some notational conventions, and to highlight a number of properties and parameters that are particularly important in the design of digital gates. We further identify the fundamental differences between bipolar and MOS transistors that helps to explain the differences in the topology of digital circuits manufactured in those technologies.

Another important function of this chapter is the introduction of the *device models*. Taking all the physical aspects of each device into account when designing complex digital circuits leads to an unnecessary complexity that quickly becomes intractable. Such an approach is similar to considering the molecular structure of concrete when constructing a bridge. To deal with this issue, an abstraction of the device behavior called a *model* is typically employed. A range of models can be conceived for each device presenting a trade-off between accuracy and complexity. A simple first-order model is useful for manual analysis. It has limited accuracy but helps us to understand the operation of the circuit and its dominant parameters. When more accurate results are needed, complex, second- or higher-order models are employed in conjunction with computer-aided simulation. In this chapter, we present both first-order models for manual analysis as well as higher-order models for simulation for each device of interest.

Designers tend to take the device parameters offered in the models for granted. They should be aware, however, that these are only nominal values, and that the actual parameter values vary with operating temperature, over manufacturing runs, or even over a single wafer. To highlight this issue, a short discussion on *process variations* and their impact is included in the chapter.

Since this text focuses on the *design aspect* of digital integrated circuits, a mere presentation of an analytical model of a device is not sufficient. Turning a conceived circuit into an actual implementation also requires a knowledge of the manufacturing process and its constraints. The interface between the design and processing world, is captured as a set of *design rules* that act as prescriptions for preparing the masks used in the fabrication process of integrated circuits. The design rules for a representative IC process are introduced in Appendix A to this chapter. A detailed description of IC fabrication processes is beyond the scope of this textbook.

## 2.2 The Diode

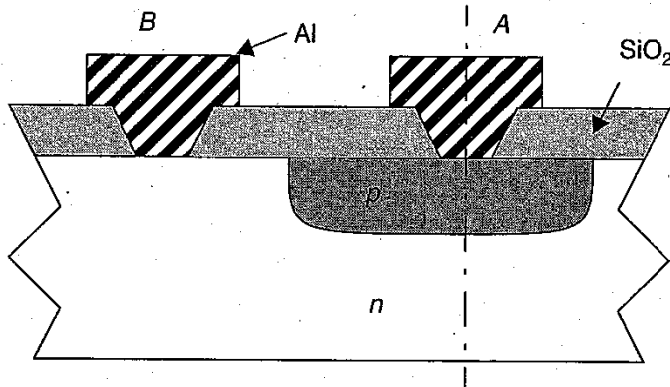
Although diodes rarely occur directly in the schematic diagrams of present-day digital gates, they are still omnipresent. For instance, each MOS transistor implicitly contains a

number of reverse-biased diodes. Diodes are used to protect the input devices of an IC against static charges. Also, a number of bipolar gates use diodes as a means to adjust voltage levels. Therefore, a brief review of the basic properties and device equations of the diode is appropriate.

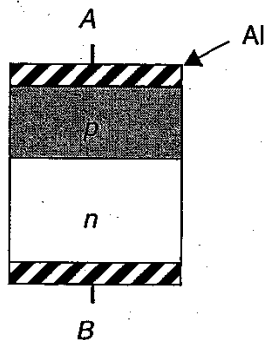
### 2.2.1 A First Glance at the Device

The *pn*-junction diode is the simplest of the semiconductor devices. Figure 2.1a shows a cross-section of a typical *pn*-junction. It consists of two homogeneous regions of *p*- and *n*-type material, separated by a region of transition from one type of doping to another, which is assumed thin. Such a device is called a *step* or *abrupt junction*. The *p*-type material is doped with *acceptor* impurities (such as boron), which results in the presence of holes as the dominant or majority carriers. Similarly, the doping of silicon with *donor* impurities (such as phosphorus or arsenic) creates an *n*-type material, where electrons are the majority carriers. Aluminum contacts provide access to the *p*- and *n*-terminals of the device. The circuit symbol of the diode, as used in schematic diagrams, is introduced in Figure 2.1c.

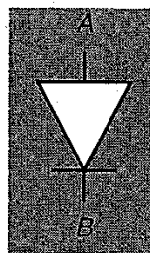
To understand the behavior of the *pn*-junction diode, we often resort to a one-dimensional simplification of the device (Figure 2.1b). Bringing the *p*- and *n*-type materials together causes a large concentration gradient at the boundary. The electron concentration changes from a high value in the *n*-type material to a very small value in the *p*-type



(a) Cross-section of *pn*-junction in an IC process



(b) One-dimensional representation



(c) Diode symbol

**Figure 2.1** Abrupt *pn*-junction diode and its schematic symbol.

material. The reverse is true for the hole concentration. This gradient causes electrons to *diffuse* from  $n$  to  $p$  and holes to diffuse from  $p$  to  $n$ . When the holes leave the  $p$ -type material, they leave behind immobile acceptor ions, which are negatively charged. Consequently, the  $p$ -type material is negatively charged in the vicinity of the  $pn$ -boundary. Similarly, a positive charge builds up on the  $n$ -side of the boundary as the diffusing electrons leave behind the positively charged donor ions. The region at the junction, where the majority carriers have been removed, leaving the fixed acceptor and donor ions, is called the *depletion* or *space-charge region*. The charges create an electric field across the boundary, directed from the  $n$  to the  $p$ -region. This field counteracts the diffusion of holes and electrons, as it causes electrons to *drift* from  $p$  to  $n$  and holes to drift from  $n$  to  $p$ . Under equilibrium, the depletion charge sets up an electric field such that the drift currents are equal and opposite to the diffusion currents, resulting in a zero net flow.

The above analysis is summarized in Figure 2.2 that plots the current directions, the charge density, the electrical field, and the electrostatic field of the abrupt  $pn$ -junction under zero-bias conditions. In the device shown, the  $p$  material is more heavily doped than the  $n$ , or  $N_A > N_D$ , with  $N_A$  and  $N_D$  the acceptor and donor concentrations, respectively. Hence, the charge concentration in the depletion region is higher on the  $p$ -side of the junction. Figure 2.2 also shows that under zero bias, there exists a voltage  $\phi_0$  across the junction, called the *built-in potential*. This potential has the value

$$\phi_0 = \phi_T \ln \left[ \frac{N_A N_D}{n_i^2} \right] \quad (2.1)$$

where  $\phi_T$  is the *thermal voltage*

$$\phi_T = \frac{kT}{q} = 26 \text{ mV at } 300 \text{ K} \quad (2.2)$$

The quantity  $n_i$  is the intrinsic carrier concentration in a pure sample of the semiconductor and equals approximately  $1.5 \times 10^{10} \text{ cm}^{-3}$  at 300 K for silicon.

---

### Example 2.1 Built-in Voltage of $pn$ -junction

An abrupt junction has doping densities of  $N_A = 10^{15} \text{ atoms/cm}^3$ , and  $N_D = 10^{16} \text{ atoms/cm}^3$ . Calculate the built-in potential at 300 K.

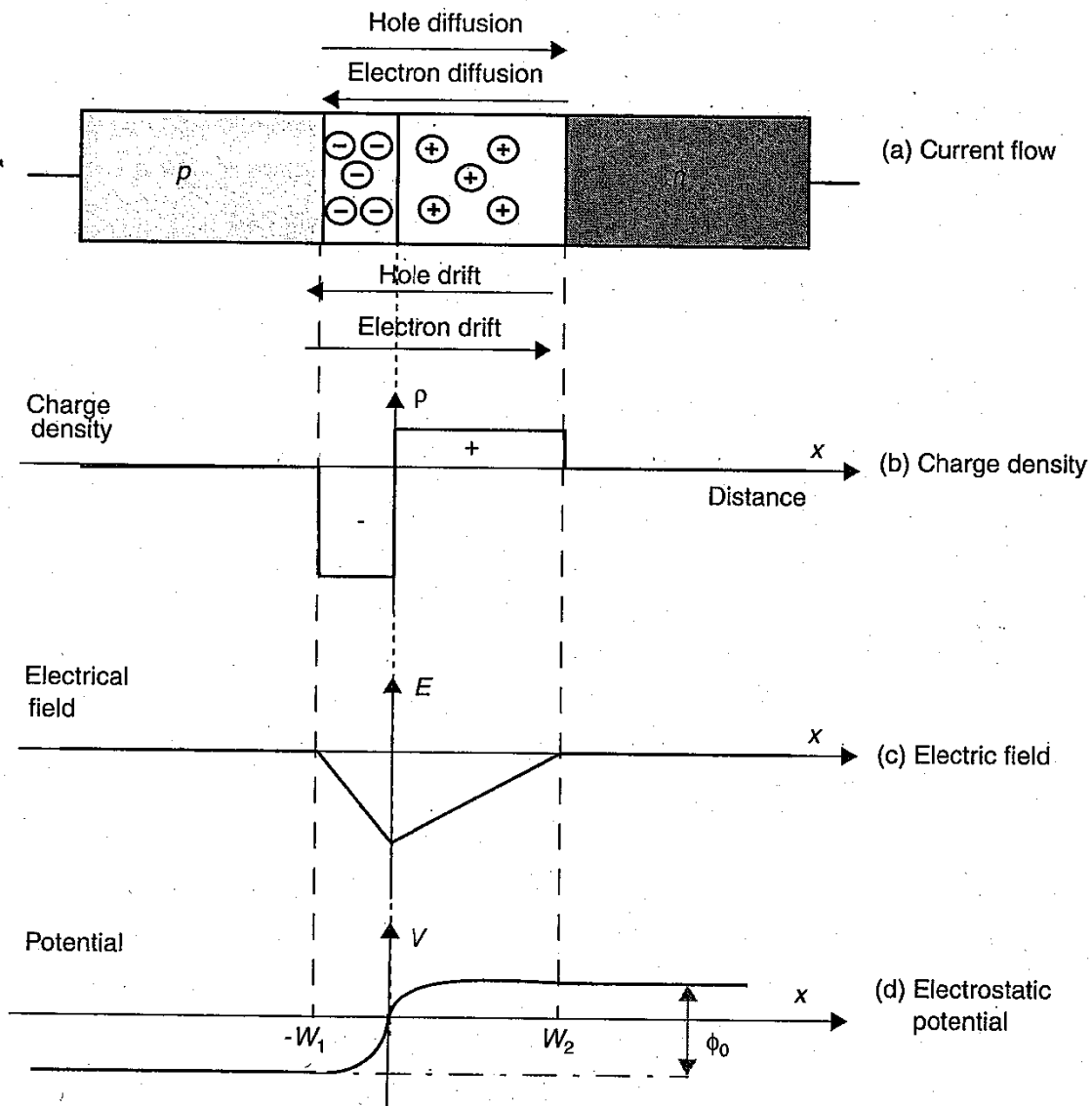
From Eq. (2.1),

$$\phi_0 = 26 \ln \left[ \frac{10^{15} \times 10^{16}}{2.25 \times 10^{20}} \right] \text{ mV} = 638 \text{ mV}$$


---

Assume now that a forward voltage  $V_D$  is applied to the junction or, in other words, that the potential of the  $p$ -region is raised with respect to the  $n$ -zone. The applied potential lowers the potential barrier. Consequently, the flow of mobile carriers across the junction increases as the diffusion current dominates the drift component. These carriers traverse the depletion region and are injected into the neutral  $n$ - and  $p$ -regions, where they become minority carriers. Under the assumption that no voltage gradient exists over the neutral regions, which is approximately the case for most modern devices, these minority carriers will diffuse through the region as a result of the concentration gradient until they get





**Figure 2.2** The abrupt  $pn$ -junction under equilibrium bias.

recombined with a majority carrier. The net result is a current flowing through the diode from the  $p$ -region to the  $n$ -region. The most important property of this current is its *exponential dependence* upon the applied bias voltage.

On the other hand, when a reverse voltage  $V_D$  is applied to the junction or when the potential of the  $p$ -region is lowered with respect to the  $n$ -region, the potential barrier is raised. This results in a reduction in the diffusion current, and the drift current becomes dominant. A current flows from the  $n$ -region to the  $p$ -region. Since the number of minority carriers in the neutral regions (electrons in the  $p$ -zone, holes in the  $n$ -region) is very small, this drift current component is virtually ignorable. It is fair to state that in the reverse-bias mode the diode operates as a nonconducting, or blocking, device. The diode thus acts as a one-way conductor. This is illustrated in Figure 2.3, which plots the diode current  $I_D$  as a function of the bias voltage  $V_D$ . The exponential behavior for positive-bias voltages is shown in Figure 2.3b, where the current is plotted on a logarithmic scale. The current increases by a factor of 10 for every extra 60 mV ( $\approx 2.3 \phi_T$ ) of forward bias. At small voltage levels ( $V_D < 0.15$  V), a deviation from the exponential dependence can be observed,

### 2.2.5 The SPICE Diode Model

In the preceding sections, we have presented a model for manual analysis of a diode circuit. For more complex circuits, or when a more accurate modeling of the diode that takes into account second-order effects is required, manual circuit evaluation becomes intractable, and computer-aided simulation is necessary. While different circuit simulators have been developed over the last decades, the SPICE program, developed at the University of California at Berkeley, is definitely the most successful [Nagel75]. Simulating an integrated circuit containing active devices requires a mathematical model for those devices (which is called the *SPICE model* in the rest of the text). The accuracy of the simulation depends directly upon the quality of this model. For instance, one cannot expect to see the result of a second-order effect in the simulation if this effect is not present in the device model. Creating accurate and computation-efficient SPICE models has been a long process and is by no means finished. Every major semiconductor company has developed their own proprietary models, which it claims have either better accuracy or computational efficiency and robustness.

The standard SPICE model for a diode is simple, as shown in Figure 2.14. The steady-state characteristic of the diode is modeled by the nonlinear current source  $I_D$ , which is a modified version of the ideal diode equation

$$I_D = I_S(e^{V_D/n\phi_T} - 1) \quad (2.34)$$

The extra parameter  $n$  is called the *emission coefficient*. It equals 1 for most common diodes but can be somewhat higher than 1 for others. The resistor  $R_s$  models the series resistance contributed by the neutral regions on both sides of the junction. For higher current levels, this resistance causes the internal diode  $V_D$  to differ from the externally applied voltage, hence causing the current to be lower than what would be expected from the ideal diode equation.

The dynamic behavior of the diode is modeled by the nonlinear capacitance  $C_D$ , which combines the two different charge-storage effects in the diode: the excess minority carrier charge and the space charge.

$$C_D = \frac{\tau_T I_S}{\phi_T} e^{V_D/n\phi_T} + \frac{C_{j0}}{(1 - V_D/\phi_0)^m} \quad (2.35)$$

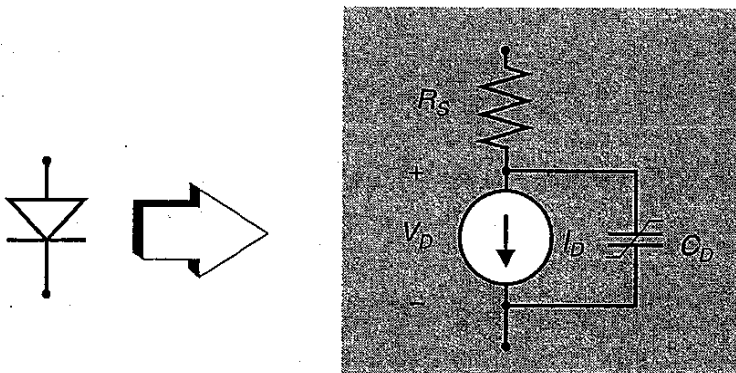


Figure 2.14 SPICE diode model.

We can verify that this equation is, aside from the introduction of the emission coefficient, nothing else than a combination of Eq. (2.23) and Eq. (2.16). The parameter  $\tau_T$  is called the *transit time* and represents, depending upon the diode type, the excess minority carrier lifetime ( $\tau_n, \tau_p$ ) for long-base diodes, or the mean transit time  $\tau_T$  for short-base diodes.

A listing of the parameters used in the diode model is given in Table 2.1. Besides the parameter name, symbol, and SPICE name, the table contains also the default value used by SPICE in case the parameter is left undefined. Observe that this table is by no means complete. Other parameters are available to govern second-order effects such as breakdown, high-level injection, and noise. To be concise, we chose to limit the listing to the parameters of direct interest to this text. For a complete description of the device models (as well as the usage of SPICE), we refer to the numerous textbooks devoted to SPICE (e.g., [Banhzaf92], [Thorpe92]).

Table 2.1 First-order SPICE diode model parameters.

Parameter Name	Symbol	SPICE Name	Units	Default Value
Saturation current	$I_S$	IS	A	1.0 E-14
Emission coefficient	$n$	N	—	1
Series resistance	$R_S$	RS	$\Omega$	0
Transit time	$\tau_T$	TT	sec	0
Zero-bias junction capacitance	$C_{j0}$	CJ0	F	0
Grading coefficient	$m$	M	—	0.5
Junction potential	$\phi_0$	VJ	V	1

## 2.3 The MOS(FET) Transistor

The metal-oxide-semiconductor field-effect transistor (MOSFET or MOS, for short) is certainly the workhorse of contemporary digital design. Its major assets are its integration density and a relatively simple manufacturing process, which make it possible to produce large and complex circuits in an economical way.

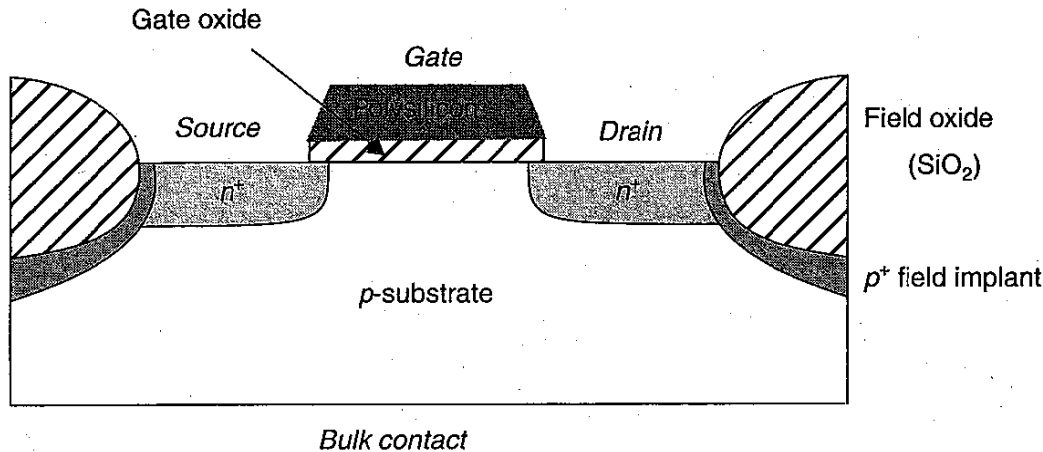
We restrict ourselves in this section to a general overview of the device and its parameters, as we did for the diode—after a generic overview of the device, we present an analytical description of the transistor from a static (steady-state) and dynamic (transient) viewpoint. The discussion concludes with an enumeration of some second-order effects and the introduction of the SPICE MOS transistor models.

### 2.3.1 A First Glance at the Device

A cross section of a typical  $n$ -channel MOS transistor (NMOS) is shown in Figure 2.15. Heavily doped  $n$ -type *source* and *drain* regions are implanted (or diffused) into a lightly doped  $p$ -type substrate (often called the *body*). A thin layer of silicon dioxide ( $\text{SiO}_2$ ) is grown over the region between the source and drain and is covered by a conductive material, most often polycrystalline silicon (or polysilicon, for short). The conductive material



forms the *gate* of the transistor. Neighboring devices are insulated from each other with the aid of a thick layer of  $\text{SiO}_2$  (called the *field oxide*) and a reverse-biased *np*-diode, formed by adding an extra  $p^+$  region, called the *channel-stop implant* (or *field implant*).



**Figure 2.15** Cross section of NMOS transistor.

At the most superficial level, the NMOS transistor can be considered to act as a switch. When a voltage is applied to the gate that is larger than a given value called the *threshold voltage*  $V_T$ , a conducting channel is formed between drain and source. In the presence of a voltage difference between drain and source, current flows between the two. The conductivity of the channel is modulated by the gate voltage—the larger the voltage difference between gate and source, the smaller the channel resistance and the larger the current. When the gate voltage is lower than the threshold, no such channel exists, and the switch is considered open.

In an NMOS transistor, current is carried by electrons moving through an *n*-type channel between source and drain. This is in contrast with the *pn*-junction diode, where current is carried by both holes and electrons. MOS devices can also be made by using an *n*-type substrate and  $p^+$  drain and source regions. In such a transistor, current is carried by holes moving through a *p*-type channel. Such a device is called a *p*-channel MOS, or PMOS transistor. In a complementary MOS technology (CMOS), both devices are present. In a pure NMOS or PMOS technology, the substrate is common to all devices and invariably connected to dc power supply voltage. In CMOS technology, PMOS and NMOS devices are fabricated in separate isolated regions called *wells* that are connected to different power supplies. Figure 2.16 shows a cross-section of a CMOS device, where PMOS transistors are implemented in a *n*-type area embedded in a *p*-type substrate. For obvious reasons, such a fabrication approach is called an *n-well* technology.

Circuit symbols for the various MOS transistors are shown in Figure 2.17. In general, the device is considered to be a three-terminal one with gate, drain, and source ports. In reality, the MOS transistor has a fourth terminal, the substrate. Since the substrate is generally connected to a dc supply that is identical for all devices of the same type (GND for NMOS,  $V_{dd}$  for PMOS), it is most often not shown on the schematics. In case a design deviates from that concept, a four-terminal symbol is also available as shown in Figure 2.17c. **If the fourth terminal is not shown, it is assumed that the substrate is connected to the appropriate supply.**

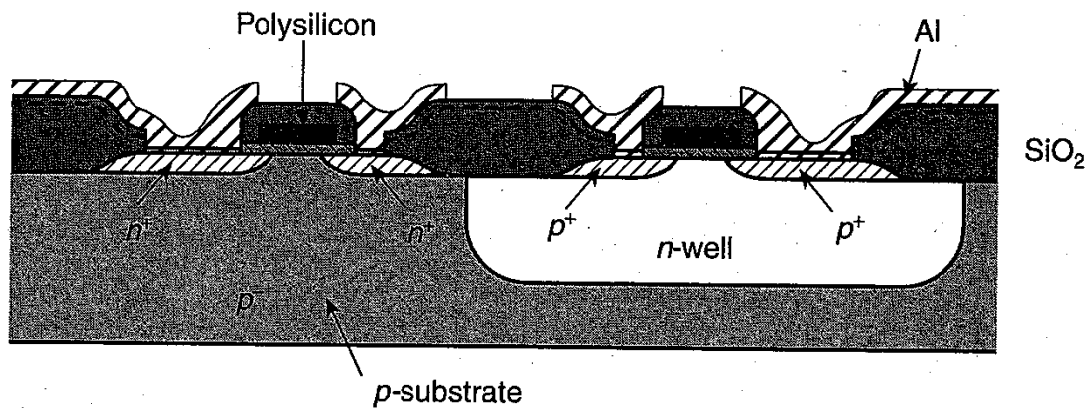


Figure 2.16 Cross section of CMOS *n*-well technology.

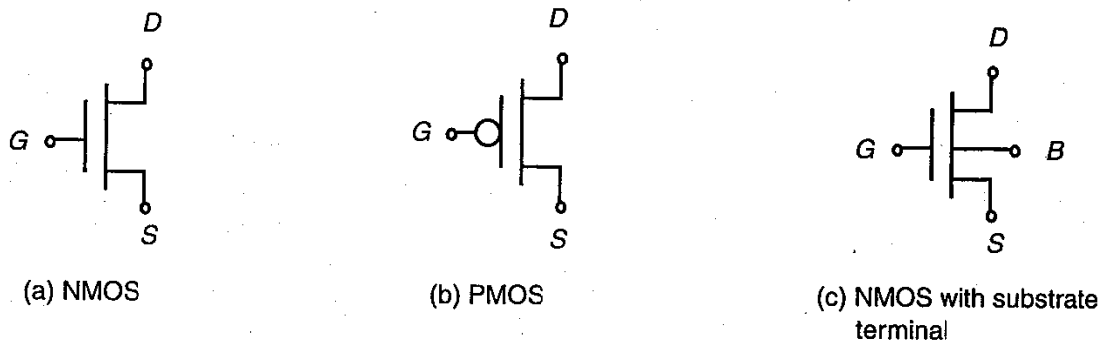


Figure 2.17 Circuit symbols for MOS transistors.

### 2.3.2 Static Behavior

In the derivation of the static model of the MOS transistor, we concentrate on the NMOS device. All the arguments made are valid for PMOS devices as well as will be discussed at the end of the section.

#### The Threshold Voltage

Consider first the case where  $V_{GS} = 0$  and drain, source, and bulk are connected to ground. The drain and source are connected by back-to-back *pn*-junctions (substrate-source and substrate-drain). Under the mentioned conditions, both junctions have a 0 V bias and can be considered off, which results in an extremely high resistance between drain and source.

Assume now that a positive voltage is applied to the gate (with respect to the source), as shown in Figure 2.18. The gate and substrate form the plates of a capacitor with the gate oxide as the dielectric. The positive gate voltage causes positive charge to accumulate on the gate electrode and negative charge on the substrate side. The latter manifests itself initially by repelling mobile holes. Hence, a depletion region is formed below the gate. This depletion region is similar to the one occurring in a *pn*-junction diode. Consequently, similar expressions hold for the width and the space charge per unit area. Compare these expressions to Eq. (2.11) and Eq. (2.12).

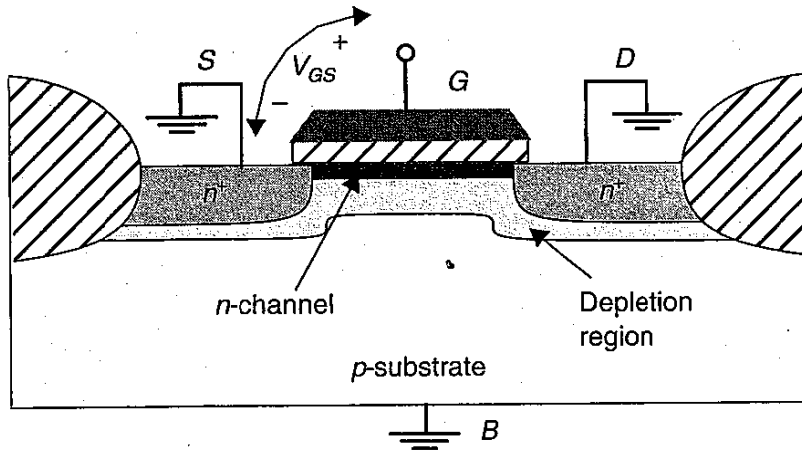


Figure 2.18 NMOS transistor for positive  $V_{GS}$ , showing depletion region and induced channel.

$$W_d = \sqrt{\frac{2\epsilon_{si}\phi}{qN_A}} \quad (2.36)$$

and

$$Q_d = \sqrt{2qN_A\epsilon_{si}\phi} \quad (2.37)$$

with  $N_A$  the substrate doping and  $\phi$  the voltage across the depletion layer (i.e., the potential at the oxide-silicon boundary).

As the gate voltage increases, the potential at the silicon surface at some point reaches a critical value, where the semiconductor surface inverts to  $n$ -type material. This point marks the onset of a phenomenon known as *strong inversion* and occurs at a voltage equal to twice the *Fermi Potential* ( $\phi_F \approx -0.3$  V for typical  $p$ -type silicon substrates).

Further increases in the gate voltage produce no further changes in the depletion-layer width, but result in additional electrons in the thin inversion layer directly under the oxide. These are drawn into the inversion layer from the heavily doped  $n+$  source region. Hence, a continuous  $n$ -type channel is formed between the source and drain regions, whose conductivity is modulated by the gate-source voltage.

In the presence of an inversion layer, the charge stored in the depletion region is fixed and equals

$$Q_{B0} = \sqrt{2qN_A\epsilon_{si}|-2\phi_F|} \quad (2.38)$$

In the presence of a substrate bias voltage  $V_{SB}$  ( $V_{SB}$  is normally positive for  $n$ -channel devices), the surface potential required for strong inversion increases and becomes  $-2\phi_F + V_{SB}$ . The charge stored in the depletion region then is expressed by Eq. (2.39)

$$Q_B = \sqrt{2qN_A\epsilon_{si}(-2\phi_F + V_{SB})} \quad (2.39)$$

The value of  $V_{GS}$  where strong inversion occurs is called the *threshold voltage*  $V_T$ . The expression of  $V_T$  consists of several components:

1. A flat-band voltage  $V_{FB}$  that represents the built-in voltage offset across the MOS structure. It consists of the work-function difference  $\phi_{ms}$ , which exists between the gate polysilicon and the silicon, and some extra components to compensate for the (undesired) fixed charge  $Q_{ox}$ , sitting at the oxide-silicon interface, and the threshold-adjusting implanted impurities  $Q_I$ .
2. A second term  $V_B$  represents the voltage drop across the depletion region at inversion and equals  $-2\phi_F$ .
3. A final component  $V_{ox}$  stands for the potential drop across the gate oxide and is equal to  $Q_B/C_{ox}$ , with  $C_{ox}$  representing the gate-oxide capacitance per unit area.

$$V_T = V_{FB} + V_B + V_{ox} = \left( \phi_{ms} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_I}{C_{ox}} \right) - 2\phi_F - \frac{Q_B}{C_{ox}} \quad (2.40)$$

While most of the terms in this expression are pure material or technology parameters,  $Q_B$  is a function of  $V_{SB}$ . It is therefore customary to reorganize the threshold equation in the following manner,

$$V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|})$$

with

$$V_{T0} = \phi_{ms} - 2\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_I}{C_{ox}} \quad (2.41)$$

and

$$\gamma = \frac{\sqrt{2q\epsilon_{si}N_A}}{C_{ox}}$$

where  $V_{T0}$  is the threshold voltage for  $V_{SB} = 0$ , and the parameter  $\gamma$  (gamma) is called the *body-effect coefficient*. The gate capacitance per unit area  $C_{ox}$  is expressed by Eq. (2.42).

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \quad (2.42)$$

with  $\epsilon_{ox} = 3.97 \times \epsilon_0 = 3.5 \times 10^{-13}$  F/cm the oxide permittivity. The expression  $t_{ox}$  stands for the oxide thickness, which is 20 nm (= 200 Å) or smaller for contemporary processes. This translates into an oxide capacitance of 1.75 fF/ $\mu\text{m}^2$ .

Observe that the threshold voltage has a **positive** value for a typical **NMOS** device, while it is **negative** for a normal **PMOS** transistor.

---

#### Example 2.8 Threshold Voltage of an NMOS Transistor

An MOS transistor has a threshold voltage of 0.75 V, while the body-effect coefficient equals 0.54. Compute the threshold voltage for  $V_{SB} = 5$  V.  $2\phi_F = -0.6$  V.

Using Eq. (2.41), we obtain  $V_T(5 \text{ V}) = 0.75 \text{ V} + 0.86 \text{ V} = 1.6 \text{ V}$ , which is more than twice the threshold under zero-bias conditions!

---

## Current-Voltage Relations

Assume now that  $V_{GS} > V_T$ . A voltage difference  $V_{DS}$  causes a current  $I_D$  to flow from drain to source (Figure 2.19). Using a simple first-order analysis, an expression of the current as a function of  $V_{GS}$  and  $V_{DS}$  can be obtained.

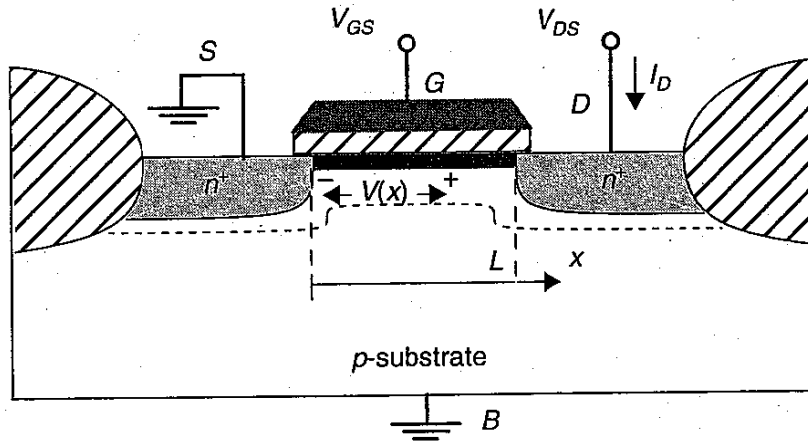


Figure 2.19 NMOS transistor with bias voltages.

At a point  $x$  along the channel, the voltage is  $V(x)$ , and the gate-to-channel voltage at that point equals  $V_{GS} - V(x)$ . Under the assumption that this voltage exceeds the threshold voltage all along the channel, the induced channel charge per unit area at point  $x$  can be computed.

$$Q_i(x) = -C_{ox}[V_{GS} - V(x) - V_T] \quad (2.43)$$

The current is given as the product of the drift velocity of the carriers  $v_n$  and the available charge. Due to charge conservation, it is a constant over the length of the channel.  $W$  is the width of the channel in a direction perpendicular to the current flow.

$$I_D = -v_n(x)Q_i(x)W \quad (2.44)$$

The electron velocity is related to the electric field through a parameter called the *mobility*  $\mu_n$  (expressed in  $\text{cm}^2/\text{V}\cdot\text{sec}$ ). The mobility is a complex function of crystal structure, local fields, and so on. In general, an empirical value is used.

$$v_n = -\mu_n E(x) = \mu_n \frac{dV}{dx} \quad (2.45)$$

Combining Eq. (2.43) – Eq. (2.45) yields

$$I_D dx = \mu_n C_{ox} W (V_{GS} - V - V_T) dV \quad (2.46)$$

Integrating the equation over the length of the channel  $L$  yields the voltage-current relation of the NMOS transistor.

$$I_D = k_n \frac{W}{L} \left[ (V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] = k_n \left[ (V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \quad (2.47)$$



$k'_n$  is called the *process transconductance parameter* and equals

$$k'_n = \mu_n C_{ox} = \frac{\mu_n \epsilon_{ox}}{t_{ox}} \quad (2.48)$$

For a typical  $n$ -channel device with  $t_{ox} = 20$  nm,  $k'_n$  equals  $80 \mu\text{A}/\text{V}^2$ . The product of the process transconductance  $k'_n$  and the  $(W/L)$  ratio of an (NMOS) transistor is called the *gain factor*  $k_n$  of the device.

As the value of the drain-source voltage is further increased, the assumption that the channel voltage is larger than the threshold all along the channel ceases to hold. This happens when  $V_{GS} - V(x) < V_T$ . At that point, the induced charge is zero, and the conducting channel disappears or is *pinched off*. This is illustrated in Figure 2.20, which shows (in an

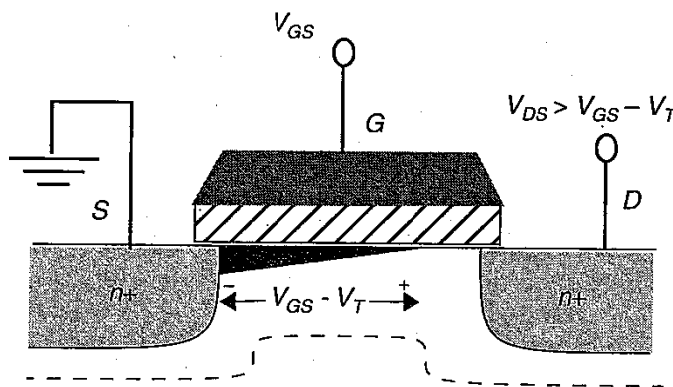


Figure 2.20 NMOS transistor under pinch-off conditions.

exaggerated fashion) how the channel thickness gradually is reduced from source to drain until pinch-off occurs. No channel exists in the vicinity of the drain region. Obviously, for this phenomenon to occur, it is essential that the pinch-off condition be met at the drain region, or

$$V_{GS} - V_{DS} \leq V_T \quad (2.49)$$

Under those circumstances, the transistor is in the *saturation* region. When a continuous channel exists between source and drain, the device operates in the *triode* (or *linear*) mode.

In the saturation region, Eq. (2.47) no longer holds. The voltage difference over the induced channel (from the pinch-off point to the source) remains fixed at  $V_{GS} - V_T$  and consequently, the current remains constant (or saturates). Replacing  $V_{DS}$  by  $V_{GS} - V_T$  in Eq. (2.47) yields the drain current for the saturation mode.

$$I_D = \frac{k'_n W}{2 L} (V_{GS} - V_T)^2 \quad (2.50)$$

This equation is not entirely correct. The position of the pinch-off point, and hence the effective length of the conductive channel, is modulated by the applied  $V_{DS}$ . As can be observed from Eq. (2.50), the current increases when the length factor  $L$  is decreased. A more accurate description of the current of the saturated MOS transistor is given in Eq. (2.51).

$$I_D = \frac{k'_n W}{2 L} (V_{GS} - V_T)^2 (1 + \lambda V_{DS}) \quad (2.51)$$

with  $\lambda$  an empirical constant parameter, called the *channel-length modulation*.<sup>3</sup>

Figure 2.21 plots  $I_D$  versus  $V_{DS}$  (with  $V_{GS}$  as a parameter) for an NMOS transistor. In the triode region, the transistor behaves like a voltage-controlled resistor, while in the saturation region, it acts as a voltage-controlled current source (when the channel-length modulation effect is ignored). Also shown is a plot of  $\sqrt{I_D}$  as a function of  $V_{GS}$  (with  $V_{DS}$  a constant). As expected a linear relationship is observed for values of  $V_{GS} \gg V_T$ . Notice also how the current does not drop abruptly to 0 at  $V_{GS} = V_T$ . At that point, the device goes into *subthreshold operation*. To turn the device completely off, the gate-source voltage has to be substantially lower than  $V_T$ . Subthreshold conduction is discussed in more detail later in the chapter, when we discuss some second-order effects in MOS transistors.

All the derived equations hold for the PMOS transistor as well. The only difference is that for PMOS devices, the polarities of all voltages and currents are reversed.

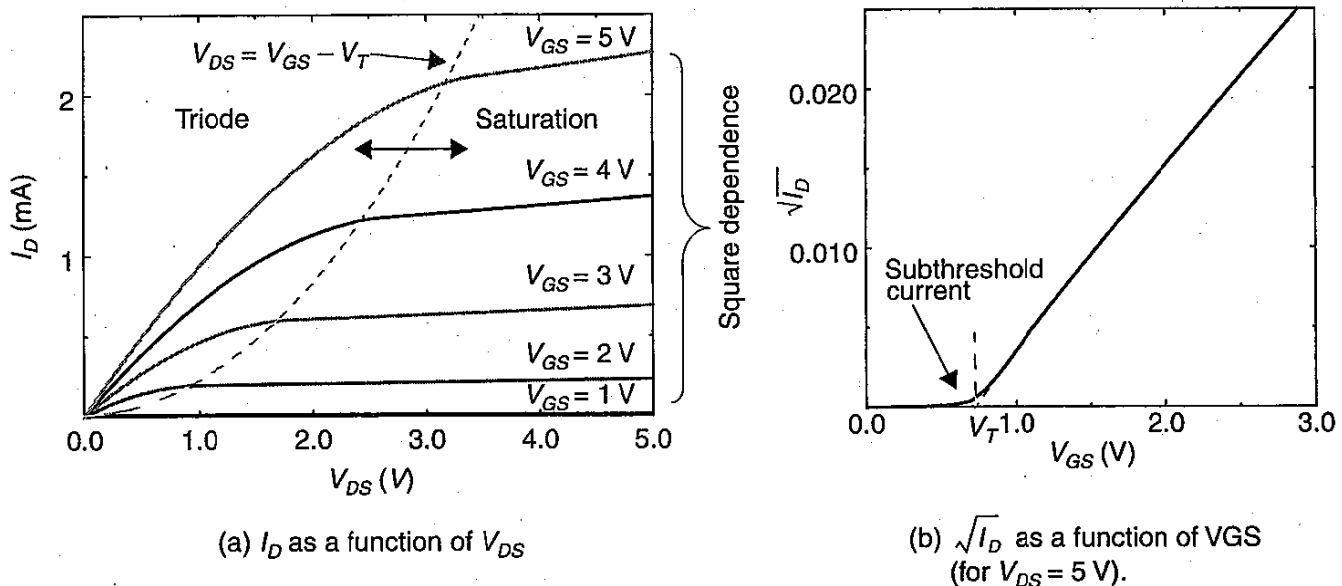


Figure 2.21  $I$ - $V$  characteristics of NMOS transistor ( $W = 100\ \mu\text{m}$ ,  $L = 20\ \mu\text{m}$  in a  $1.2\ \mu\text{m}$  CMOS technology).

### Problem 2.2 PMOS $I$ - $V$ Characteristic

Draw the  $I$ - $V$  curves of a PMOS transistor. Derive expressions for the PMOS drain current in the saturation and triode regions, which take the polarities of voltages and currents into account.

### A Model for Manual Analysis

The derived equations can be combined into a simple device model, which we will employ for the manual analysis of MOS circuits in the rest of the book. It is summarized in Figure 2.22.

<sup>3</sup> Analytical expressions for  $\lambda$  have proven to be complex and inaccurate. Device experiments and simulations indicate that  $\lambda$  varies roughly with the inverse of the channel length.

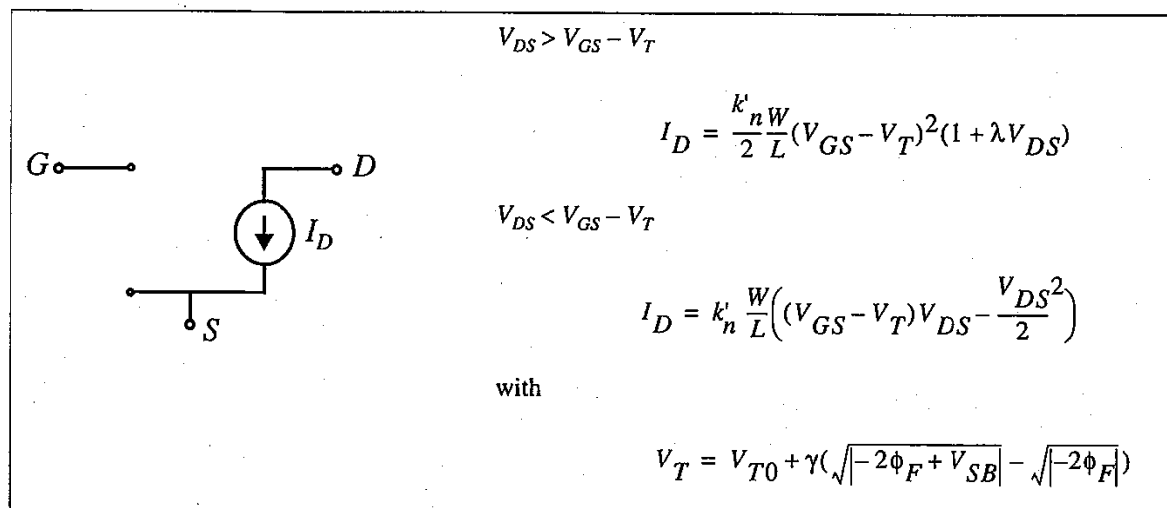


Figure 2.22 An MOS model for manual analysis.

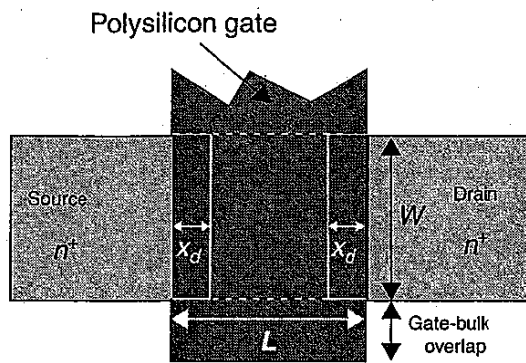
### 2.3.3 Dynamic Behavior

The transient behavior of the  $pn$ -junction diode is dominated by the moving of the excess minority carrier charge in the neutral zones and, to a second degree, by the space charge in the depletion region. Since the MOSFET is a majority carrier device, its dynamic response is solely determined by the time to (dis)charge the capacitances between the device ports and from the interconnecting lines. An accurate analysis of the nature and behavior of these capacitances is essential when designing high-performance digital circuits. They originate from three sources: the basic MOS structure, the channel charge, and the depletion regions of the reverse-biased  $pn$ -junctions of drain and source. Aside from the MOS structure capacitances, all capacitors are nonlinear and vary with the applied voltage. We discuss each of the components in turn.

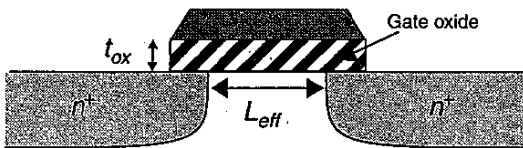
#### MOS Structure Capacitances

The gate of the MOS transistor is isolated from the conducting channel by the gate oxide that has a capacitance per unit area equal to  $C_{ox} = \epsilon_{ox} / t_{ox}$ . From the  $I$ - $V$  equations, we learned that it is useful to have  $C_{ox}$  as large as possible, or to keep the oxide thickness very thin. The total value of this capacitance is called the *gate capacitance*  $C_g$  and equals  $C_{ox}WL$ . This gate capacitance can be decomposed into a number of elements, each with a different behavior. Obviously, one part of  $C_g$  contributes to the channel charge, and is discussed in a subsequent section. Another part is solely due to the topological structure of the transistor. This component is the subject of the remainder of this section.

Consider the transistor structure of Figure 2.23. Ideally, the source and drain diffusion should end right at the edge of the gate oxide. In reality, both source and drain tend to extend somewhat below the oxide by an amount  $x_d$ , called the *lateral diffusion*. Hence, the effective channel of the transistor  $L_{eff}$  becomes shorter than the drawn length (or the length the transistor was originally designed for) by a factor of  $2x_d$ . It also gives rise to a parasitic capacitance between gate and source (drain) that is called the *overlap capacitance*. This capacitance is strictly linear and has a fixed value



(a) Top view



(b) Cross section

Figure 2.23 MOSFET overlap capacitance.

$$C_{gsO} = C_{gdO} = C_{ox}x_dW = C_OW \quad (2.52)$$

Since  $x_d$  is a technology-determined parameter, it is customary to combine it with the oxide capacitance to yield the overlap capacitance per unit transistor width  $C_O$ .

### Channel Capacitance

The gate-to-channel capacitance can be decomposed into three components:  $C_{gs}$ ,  $C_{gd}$ , and  $C_{gb}$ , being the capacitance between the gate and the source, drain, and bulk regions, respectively. All those components are nonlinear, and their value depends upon the operation region. To simplify the analysis, estimated and average values are used. For instance, in the cut-off mode, no channel exists, and the total capacitance  $C_{ox}WL_{eff}$  appears between gate and bulk. In the triode region, an inversion layer is formed, which acts as a conductor between source and drain. Consequently,  $C_{gb} = 0$  as the bulk electrode is shielded from the gate by the channel. Symmetry dictates that  $C_{gs} \approx C_{gd} \approx C_{ox}WL_{eff}/2$ . Finally, in the saturation mode, the channel is pinched off. The capacitance between gate and drain is thus approximately zero, and so is the gate-bulk capacitance. A careful analysis of the channel charge, taking into account the potential variations over the channel, indicates that  $C_{gs}$  averages  $2/3 C_{ox}WL_{eff}$ . Although these expressions are approximations, they are adequate for the initial design estimates. The derived values are summarized in Table 2.2.

Table 2.2 Average channel capacitances of MOS transistor for different operation regions.

Operation Region	$C_{gb}$	$C_{gs}$	$C_{gd}$
Cutoff	$C_{ox}WL_{eff}$	0	0
Triode	0	$C_{ox}WL_{eff}/2$	$C_{ox}WL_{eff}/2$
Saturation	0	$(2/3)C_{ox}WL_{eff}$	0

### Junction Capacitances

A final capacitive component is contributed by the reverse-biased source-bulk and drain-bulk  $pn$ -junctions. The depletion-region capacitance is nonlinear and decreases when the reverse bias is raised as discussed earlier. To understand the components of the junction capacitance (often called the *diffusion capacitance*), we must look at the source (drain) region and its surroundings. The detailed picture, shown in Figure 2.24, shows that the junction consists of two components:

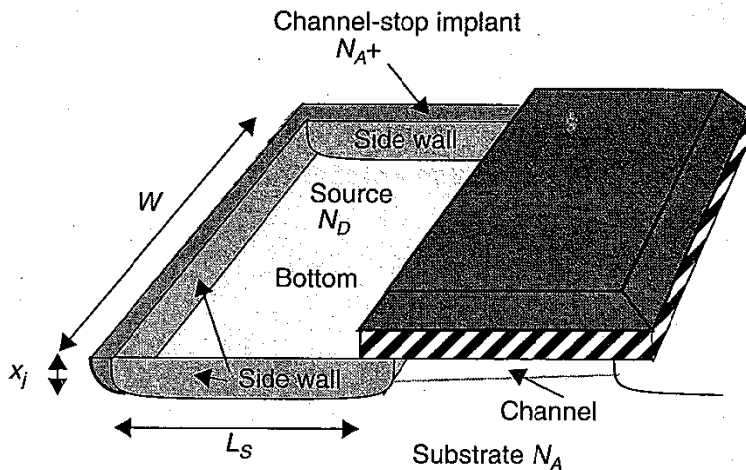


Figure 2.24 Detailed view of source junction.

- The *bottom-plate* junction, which is formed by the source region (with doping  $N_D$ ) and the substrate with doping  $N_A$ . The total depletion region capacitance for this component equals  $C_{bottom} = C_j W L_s$ , with  $C_j$  the junction capacitance per unit area as given by Eq. (2.16). As the bottom-plate junction is typically of the abrupt type, the grading coefficient  $m$  is set to 0.5.
- The *side-wall* junction, formed by the source region with doping  $N_D$  and the  $p^+$  channel-stop implant with doping level  $N_A^+$ . The doping level of the stopper is usually larger than that of the substrate, resulting in a larger capacitance per unit area. The side-wall junction resembles the graded type, which sets  $m$  to 1/3. Its total value equals  $C_{sw} = C'_{jsw} x_j (W + 2 \times L_s)$ . Notice that no side-wall capacitance is counted for the fourth side of the source region, as this represents the conductive channel. Since  $x_j$ , the junction depth, is a technology parameter, it is normally combined with  $C'_{jsw}$  into a capacitance per unit perimeter  $C_{jsw} = C'_{jsw} x_j$ .

An expression for the total junction capacitance can now be derived,

$$\begin{aligned} C_{diff} &= C_{bottom} + C_{sw} = C_j \times \text{AREA} + C_{jsw} \times \text{PERIMETER} \\ &= C_j L_s W + C_{jsw} (2L_s + W) \end{aligned} \quad (2.53)$$

Since all these capacitances are small-signal capacitances, we normally linearize them and use average capacitances along the lines of Eq. (2.17).



## Capacitive Device Model

All the above contributions can be combined in a single capacitive model for the MOS transistor, which is shown Figure 2.25. Its components are readily identified on the basis of the preceding discussions.

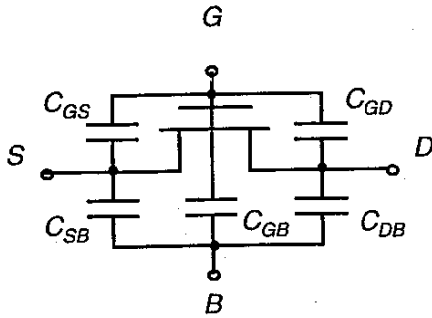


Figure 2.25 MOSFET capacitance model.

$$C_{GS} = C_{gs} + C_{gs0}; C_{GD} = C_{gd} + C_{gd0}; C_{GB} = C_{gb}$$

$$C_{SB} = C_{Sdiff}; C_{DB} = C_{Ddiff} \quad (2.54)$$

A good understanding of this model as well as of the relative values of its components is essential in the design and optimization of high performance digital circuits. The dynamic performance of these circuits is directly proportional to the capacitance.

### Example 2.9 MOS Transistor Capacitances

Consider an NMOS transistor with the following parameters:  $t_{ox} = 20$  nm,  $L = 1.2$   $\mu\text{m}$ ,  $W = 1.8$   $\mu\text{m}$ ,  $L_D = L_S = 3.6$   $\mu\text{m}$ ,  $x_d = 0.15$   $\mu\text{m}$ ,  $C_{j0} = 3 \times 10^{-4}$  F/m<sup>2</sup>,  $C_{jsw0} = 8 \times 10^{-10}$  F/m. Determine the zero-bias value of all relevant capacitances.

The gate capacitance per unit area is easily derived as  $(\epsilon_{ox}/t_{ox})$  and equals 1.75 fF/ $\mu\text{m}^2$ . The total gate capacitance  $C_g$  is equal to  $WLC_{ox} = 3.78$  fF. This is divided into the overlap capacitances ( $C_{GSO} = C_{GDO} = Wx_dC_{ox} = 0.47$  fF) and the channel capacitance, which splits between source, drain, and bulk terminals, dependent upon the operation region, and equals  $3.78 - 2 \times 0.47 = 2.84$  fF.

The diffusion capacitance consists of the bottom and the side-wall capacitances. The former is equal to  $C_{j0}L_DW = 1.95$  fF, while the side-wall capacitance under zero-bias conditions evaluates to  $C_{jsw0}(2L_D + W) = 7.2$  fF.

The diffusion capacitance seems to dominate the gate capacitance. This is a worst-case condition, however. When increasing the value of the reverse bias, the diffusion capacitance is substantially reduced (to about 50% of its value). In general, it can be stated that both contributions are virtually similar in value. Observe also the large value of the side-wall versus bottom-plate capacitance for this particular process. Advanced processes reduce the diffusion capacitances by using materials such as SiO<sub>2</sub> to isolate the devices. This approach is called *trench isolation*.

### 2.3.4 The Actual MOS Transistor—Secondary Effects

Up to this point, we have discussed the behavior of an ideal MOS device. The operation of an actual transistor can deviate substantially from this model. This is especially true when

the dimensions of the device reach the  $\mu\text{m}$ -range or below. At that point, the channel length becomes comparable to other device parameters such as the depth of drain and source junctions, and the width of their depletion regions. Such a device is called a *short-channel* transistor, in contrast to the *long-channel* devices discussed so far. The behavior of a long-channel device is adequately described by a one-dimensional model, where it is assumed that all current flows on the surface of the silicon and the electrical fields are oriented along that plane. In short-channel devices, those assumptions are no longer valid and a two-dimensional model is more appropriate. This results in important deviations from the ideal model.

The understanding of some of these second-order effects and their impact on the device behavior is essential in the design of contemporary digital circuits and therefore merits some discussion. One word of warning, though. Trying to take all those effects into account in a manual, first-order analysis results in intractable and opaque circuit models. It is therefore advisable to analyze and design MOS circuits first using the ideal model. The impact of the nonidealities can be studied in a second round using computer-aided simulation tools with more precise transistor models.

### Threshold Variations

Eq. (2.41) states that the threshold voltage is only a function of the manufacturing technology and the applied body bias  $V_{SB}$ . The threshold can therefore be considered as a constant over all NMOS (PMOS) transistors in a design. As the device dimensions are reduced, this model becomes inaccurate, since the threshold potential becomes a function of  $L$ ,  $W$ , and  $V_{DS}$ . For instance, in the derivation of  $V_{T0}$  it was assumed that all depletion charge beneath the gate originates from the MOS field effects. This ignores the depletion regions of the source and reverse-biased drain junction that become relatively more important with shrinking channel lengths. Since a part of the region below the gate is already depleted (by the source and drain fields), a smaller threshold voltage suffices to cause strong inversion. In other words,  $V_{T0}$  decreases with  $L$  for short-channel devices (Figure 2.26a). A similar

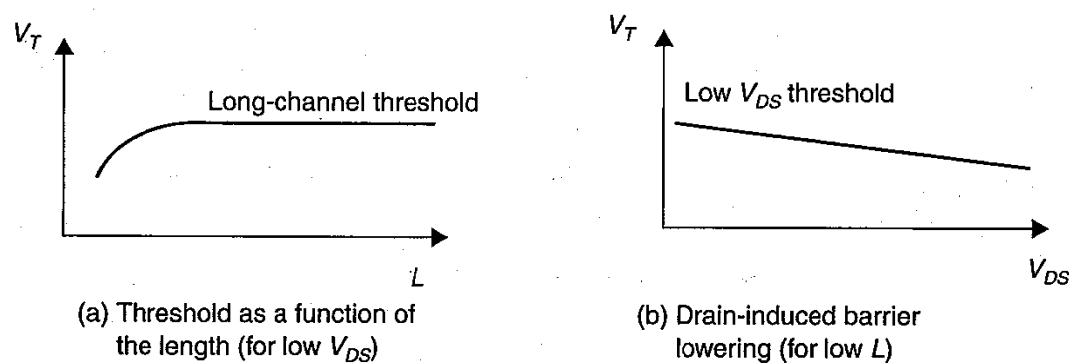


Figure 2.26 Threshold variations.

effect can be obtained by raising the drain-source (bulk) voltage, as this increases the width of the drain-junction depletion region. Consequently, the threshold decreases with increasing  $V_{DS}$ . This effect, called the *drain-induced barrier lowering*, or *DIBL*, causes the threshold potential to be a function of the operating voltages (Figure 2.26b). For high enough

values of the drain voltage, the source and drain regions can even be shorted together, and normal transistor operation ceases to exist. This effect is called *punchthrough*.

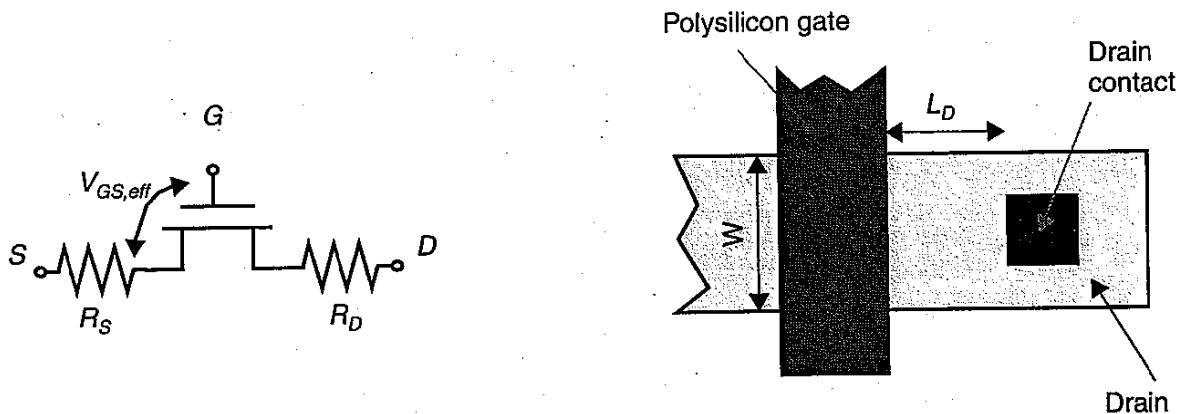
Since the majority of the transistors in a digital circuit are designed at the minimum channel length, the variation of the threshold voltage as a function of the length is almost uniform over the complete design, and is therefore a minor issue. More troublesome is the DIBL, as this effect varies with the operating voltage. This is, for instance, a problem in dynamic memories, where the leakage current of a cell (being the subthreshold current of the access transistor) becomes a function of the voltage on the data-line, which is shared with many other cells.

Besides varying over a design, threshold voltages in short-channel devices also have the tendency to *drift over time*. This is the result of the *hot-carrier* effect [Hu92]. Over the last decades, device dimensions have been scaled down continuously, while the power supply and the operating voltages were kept constant. The resulting increase in the electrical field strength causes an increasing velocity of the electrons, which can leave the silicon and tunnel into the gate oxide upon reaching a high enough energy level. Electrons trapped in the oxide change the threshold voltage, typically increasing the thresholds of NMOS devices, while decreasing the  $V_T$  of PMOS transistors. For an electron to become hot, an electrical field of at least  $10^4$  V/cm is necessary. This condition is easily met in devices with channel lengths around or below  $1\ \mu\text{m}$ . The hot-electron phenomenon can lead to a long-term reliability problem, where a circuit might degrade or fail after being in use for a while.

### Source-Drain Resistance

When transistors are scaled down, their junctions are shallower, and the contact openings become smaller. This results in an increase in the parasitic resistance in series with the drain and source regions, as shown in Figure 2.27a. The resistance of the drain (source) region can be expressed as

$$R_{S,D} = \frac{L_{S,D}}{W} R_{\square} + R_C \quad (2.55)$$



(a) Modeling the series resistance

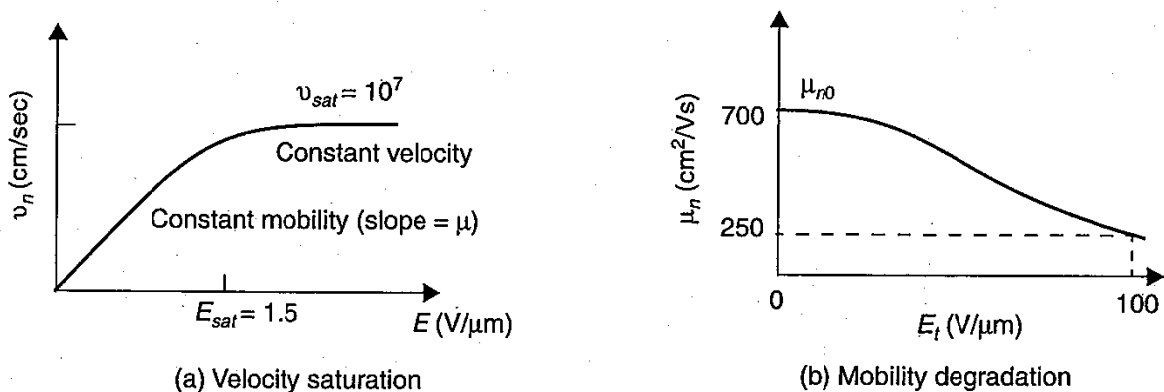
(b) Parameters of the series resistance

**Figure 2.27** Series drain and source resistance.

with  $R_C$  the contact resistance,  $W$  the width of the transistor, and  $L_{S,D}$  the length of the source or drain region (Figure 2.27b).  $R_{\square}$  is the sheet resistance per square of the drain-source diffusion, which ranges from  $50 \Omega/\square$  to  $1 \text{ k}\Omega/\square$ . Observe that the resistance of a square of material is constant, independent of its size (see Chapter 8). The series resistance causes a deterioration in the device performance, as it reduces the drain current for a given control voltage. Keeping its value as small as possible is thus an important design goal (for both the device and the circuit engineer). One option is to cover the drain and source regions with a low-resistivity material, such as titanium or tungsten. This process is called *silicidation* and effectively reduces the parasitic resistance. Silicidation is furthermore used to reduce the resistance of the polysilicon gate. Making the transistor wider than needed is another possibility as should be obvious from Eq. (2.55).

### Variations in $I$ - $V$ Characteristics

The voltage-current relations of a short-channel device deviate considerably from the ideal expressions of Eq. (2.47) and Eq. (2.51). The most important reasons for this difference are the *velocity saturation* and the *mobility degradation* effects. In Eq. (2.45), it was stated that the velocity of the carriers is proportional to the electrical field, independent of the value of that field. In other words, the carrier mobility is a constant. However, when the electrical field along the channel reaches a critical value  $E_{sat}$ , the velocity of the carriers tends to saturate as illustrated in Figure 2.28a.



**Figure 2.28** Effect of electrical field on electron velocity and mobility.

For  $p$ -type silicon, the critical field at which electron saturation occurs is  $1.5 \times 10^4$  V/cm (or  $1.5 \text{ V}/\mu\text{m}$ ), and the saturation velocity  $v_{sat}$  equals  $10^7$  cm/sec. This means that in an NMOS device with a channel length of  $1 \mu\text{m}$ , only a couple of volts between drain and source are needed to reach the saturation point. This condition is easily met in current short-channel devices. Holes in a  $n$ -type silicon saturate at the same velocity, although a higher electrical field is needed to achieve saturation ( $\geq 10^5$  V/cm).

This effect has a profound impact on the operation of the transistor. This is easily realized when computing the transistor currents under the velocity-saturated condition. Combining Eq. (2.43) and Eq. (2.44) and setting  $v_n$  to  $v_{sat}$  yields a revised current expression,

$$I_{DSAT} = v_{sat} C_{ox} W (V_{GS} - V_{DSAT} - V_T) \quad (2.56)$$

with  $V_{DSAT}$  the drain-source voltage at which velocity saturation comes into play. Observe the *linear dependence* of the saturation current with respect to the gate-source voltage  $V_{GS}$ , which is in contrast with the squared dependence in the long-channel device. Consequently, reducing the operating voltage does not have such a significant effect in submicron devices as it would have in a long-channel transistor. Furthermore,  $I_D$  is independent of  $L$  in velocity-saturated devices, suggesting that current drive cannot be further improved by decreasing the channel length as was the case in long-channel transistors. Observe that this is only true to a first degree, as it ignores the influence of  $L$  on  $V_{DSAT}$ .

Reducing the channel length has another important impact on the transistor current: even at normal electric field levels, a reduction in the electron mobility can be observed. This effect, called *mobility degradation*, can be attributed to the vertical component of the electrical field, which is no longer ignorable in these small devices. This is illustrated in Figure 2.28b, where the electron mobility is plotted as a function of the transversal electrical field.

Both effects can be combined into an approximate but manageable model for the short-channel MOSFET transistor (proposed in [Toh88])

$$\begin{aligned} I_D &= \kappa v_{sat} C_{ox} W (V_{GS} - V_T) \text{ for } V_{DS} \geq V_{DSAT} \text{ (saturated region)} \\ &= \mu_n C_{ox} \frac{W}{L} \left( (V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right) \text{ for } V_{DS} \leq V_{DSAT} \text{ (triode region)} \end{aligned} \quad (2.57)$$

with  $V_{DSAT}$  the saturation voltage, given by

$$V_{DSAT} = (1 - \kappa)(V_{GS} - V_T) \quad (2.58)$$

where  $\kappa$  is a measure of the velocity-saturation degree (with  $E$  the longitudinal electrical field):

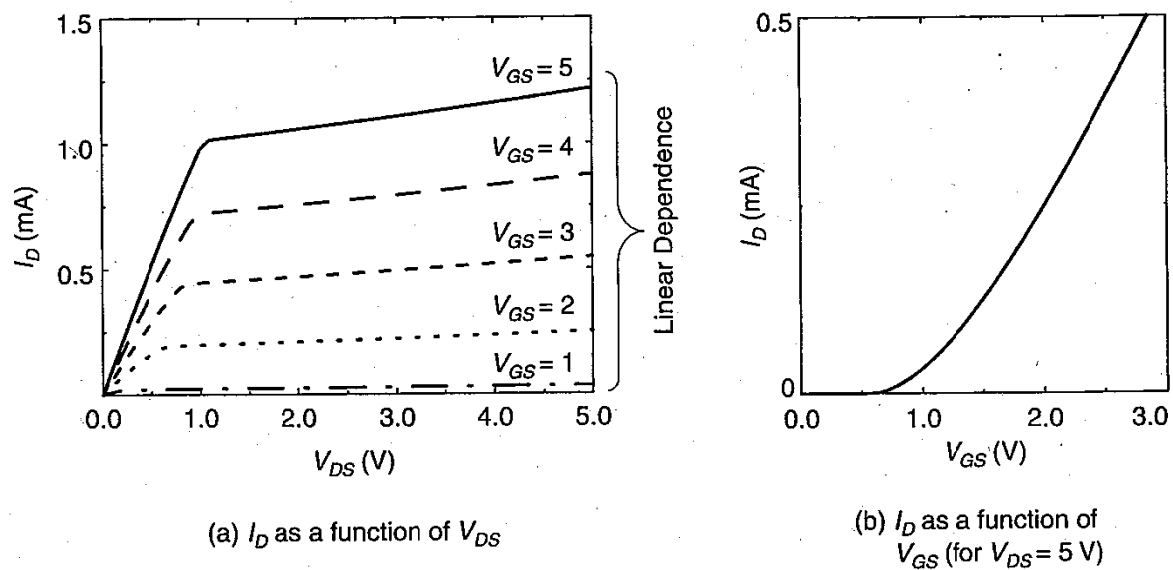
$$\kappa = \frac{1}{1 + E_{sat}/E} = \frac{1}{1 + E_{sat}L/(V_{GS} - V_T)} \quad (2.59)$$

Be aware of the fact that the mobility  $\mu_n$  in Eq. (2.57) is not a constant, but a function of the applied electrical field as well due to the mobility degradation. Also, it is easily shown that for  $E \ll E_{sat}$ , Eq. (2.57) reverts to the long-channel model, taking into account that the electric field at saturation is related to the maximum velocity by the following expression:  $E_{sat} = 2 v_{sat}/\mu_n$  [Toh88].

From Eq. (2.58), it can be observed that a short-channel MOS has an extended saturation region as compared to a long-channel transistor (as  $0 < \kappa < 1$ ). Both the extended saturation region and the linear dependence upon  $V_{GS}$  are apparent in the simulated  $I$ - $V$  response of a short-channel device, shown in Figure 2.29a and b. The simulated device uses the same technology as the example of Figure 2.21, but the channel length is set to the minimum allowed value of 1.2  $\mu\text{m}$ .

For detailed, excellent descriptions of short-channel effects in MOS transistors, please refer to [Muller86, Chen90, Sze81, Ko89].





**Figure 2.29** Short-channel NMOS transistor  $I$ - $V$  characteristic ( $W = 4.6 \mu\text{m}$ ,  $L = 1.2 \mu\text{m}$  in a  $1.2 \mu\text{m}$  CMOS technology).

### Example 2.10 Impact of Velocity Saturation

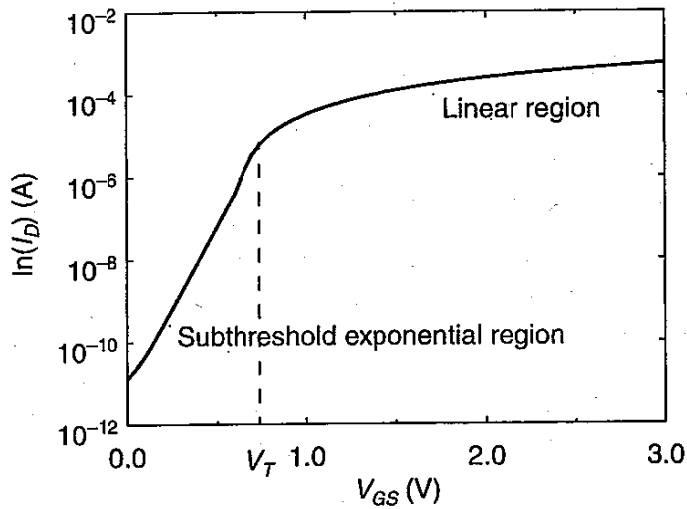
The devices of Figure 2.21 and Figure 2.29 have identical effective ( $W/L$ ) ratios (when taking into account the lateral diffusion). If the long-channel model were valid, both devices would produce identical  $I$ - $V$  characteristics. The latter transistor ( $W = 4.6 \mu\text{m}$ ,  $L = 1.2 \mu\text{m}$ ), however, suffers from velocity saturation, while this is not the case for the former device with its very long channel ( $W = 100 \mu\text{m}$ ,  $L = 20 \mu\text{m}$ ).

This results in a substantial drop in current drive for high voltage levels. For instance, at ( $V_{GS} = 5$  V,  $V_{DS} = 5$  V), the drain current of the small transistor is only 53% of the corresponding value of the longer transistor (1.2 mA versus 2.3 mA). At lower values of the drain-source voltage, the effect of velocity saturation is less significant. For instance, at ( $V_{GS} = 5$  V,  $V_{DS} = 1$  V) both devices yield approximately the same current of 0.95 mA.

### Subthreshold Conduction

From inspection of Figure 2.21b and Figure 2.29b, the reader is probably aware that the MOS transistor is already partially conducting for voltages below the threshold voltage. This effect is called *subthreshold* or *weak-inversion* conduction. The onset of strong inversion means that ample carriers are available for conduction, but by no means implies that no current at all can flow for gate-source voltages below  $V_T$ . However, the current levels are small under those conditions. The transition from the on- to the off-condition is not abrupt, but gradual.

To study this effect in somewhat more detail, we redraw the  $I_D$  versus  $V_{GS}$  curve of Figure 2.29b on a logarithmic scale as shown in Figure 2.30. This clearly demonstrates that the current does not drop to zero immediately for  $V_{GS} < V_T$ , but actually decays in an exponential fashion, similar to the operation of a bipolar transistor (which will be discussed in Section 2.4). In the absence of a conducting channel, the  $n^+$  (source) -  $p$  (bulk) -



**Figure 2.30**  $I_D$  current versus  $V_{GS}$  (on logarithmic scale), showing the exponential characteristic of the subthreshold region.

$n^+$  (drain) terminals actually form a parasitic bipolar transistor. In this operation region, the (inverse) rate of decrease of the current with respect to  $V_{GS}$  is approximated as stated in [Sze81]:

$$\left( \frac{d}{dV_{GS}} \ln(I_D) \right)^{-1} = \frac{kT}{q} \ln 10 (1 + \alpha) \quad (2.60)$$

The expression  $(kT/q) \ln(10)$  evaluates to 60 mV/decade at room temperature.  $\alpha$  equals 0 for an ideal device which means that at room temperature the subthreshold current drops by a factor of 10 for a reduction in  $V_{GS}$  of 60 mV. Unfortunately,  $\alpha$  is larger than 1 for actual devices and the current drops at a reduced rate. The current drop is further affected in a negative sense by an increase in the operating temperature (most integrated circuits operate at temperatures considerably beyond room temperature). Note that  $\alpha$  is a function of the transistor capacitances. Reducing it requires advanced and expensive technologies, such as silicon-on-insulator.

The presence of the subthreshold current detracts from the ideal switch model that we like to assume for the MOS transistor. In general, we want the current to be as close as possible to zero at  $V_{GS} = 0$ . This is especially important in the so-called *dynamic circuits*, which rely on the storage of charge on a capacitor and whose operation can be severely degraded by subthreshold leakage. This observation puts a firm lower bound on the value of the threshold voltage of the devices.

---

#### Example 2.11 Subthreshold Slope

For the example of Figure 2.30, a slope of 121 mV/decade is observed. This is equivalent to an  $\alpha$ -factor of 1.

---

### CMOS Latchup

The MOS technology contains a number of intrinsic bipolar transistors. These are especially troublesome in CMOS processes, where the combination of wells and substrates results in the formation of parasitic  $n$ - $p$ - $n$ - $p$  structures. Triggering these thyristor-like

devices leads to a shorting of the  $V_{DD}$  and  $V_{SS}$  lines, usually resulting in a destruction of the chip, or at best a system failure that can only be resolved by power-down.

Consider the  $n$ -well structure of Figure 2.31a. The  $n$ - $p$ - $n$ - $p$  structure is formed by the source of the NMOS, the  $p$ -substrate, the  $n$ -well and the source of the PMOS. A circuit equivalent is shown in Figure 2.31b. When one of the two bipolar transistors gets forward biased (e.g., due to current flowing through the well, or substrate), it feeds the base of the other transistor. This positive feedback increases the current until the circuit fails or burns out.

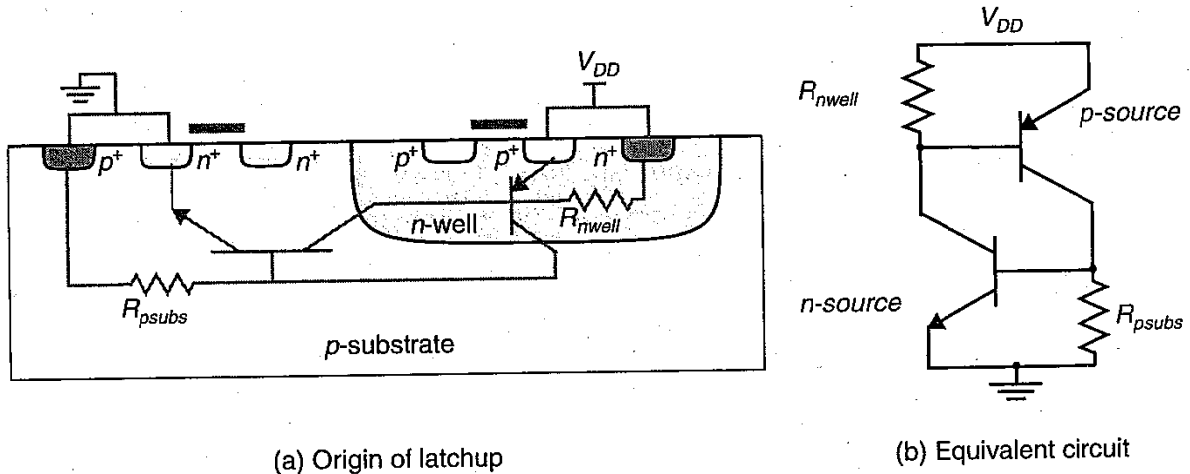


Figure 2.31 CMOS latchup.

From the above analysis the message to the designer is clear—to avoid latchup, the resistances  $R_{nwell}$  and  $R_{psubs}$  should be minimized. This can be achieved by providing numerous well and substrate contacts, placed close to the source connections of the NMOS/PMOS devices. Devices carrying a lot of current (such as transistors in the I/O drivers) should be surrounded by *guard rings*. These circular well/substrate contacts, positioned around the transistor, reduce the resistance even further and reduce the gain of the parasitic bipolars. For an extensive discussion on how to avoid latchup, please refer to [Weste93]. The latchup effect was especially critical in early CMOS processes. In recent years, process innovations and improved design techniques have all but eliminated the risks for latchup.

### 2.3.5 SPICE Models for the MOS Transistor

The complexity of the behavior of the short-channel MOS transistor and its many parasitic effects has led to the development of a wealth of models for varying degrees of accuracy and computing efficiency. In general, more accuracy also means more complexity and, hence, an increased run time. In SPICE, the choice of the device model is set by the LEVEL parameter. In this section, we introduce some of the most commonly used SPICE MOSFET models as well as their parameters. We also revisit briefly our simple model for manual analysis to accommodate some of the second-order effects.

## SPICE Models

- The LEVEL 1 SPICE model implements the *Shichman-Hodges model*, which is based on the square law long-channel expressions, derived earlier in this chapter. As it does not handle short-channel effects, it is not very appropriate for contemporary devices. Its main purpose is to verify a manual analysis.
- The LEVEL 2 model is a geometry-based model, which uses detailed device physics to define its equations. It handles effects such as velocity saturation, mobility degradation, and drain-induced barrier lowering. Unfortunately, including all 3D-effects of an advanced submicron process in a pure physics-based model becomes complex and inaccurate. The LEVEL 2 model is, therefore, virtually obsolete.
- LEVEL 3 is a semi-empirical model. It relies on measured device data to determine its parameters. However, depending upon on the specified parameters, the designer can use mixed models to calculate the threshold voltage and the drain current.
- LEVEL 4 (the Berkeley Short-Channel IGFET Model, or BSIM) provides a model that is analytically simple and is based on a small number of parameters, which are normally extracted from experimental data. Its accuracy and efficiency make it one of the most popular SPICE MOSFET models at present.
- A large number of other models are available, both from SPICE vendors and semiconductor manufacturers. A complete description of all those models would take the remainder of this book, which is, obviously, not the goal. For a good description of SPICE MOSFET models, please refer to [Antognetti88].

Table 2.3 lists the main SPICE model parameters (as used in LEVELS 1 to 3). The parameters covering the parasitic resistive and capacitive effects have been transferred to a separate table for the sake of clarity (Table 2.4). Whenever possible, we have correlated the SPICE parameter to the symbol used in this book. Observe that some of the defined parameters are redundant. For instance, PHI ( $\phi_0$ ) can be computed from process parameters such as the substrate doping. User-defined parameters always preside over the analytical value, however. The list is by no means complete, but is sufficient to cover the requirements of this and later chapters.

To conclude this lengthy enumeration, the parameters that can be defined an individual transistor have to be mentioned (Table 2.5). Not all these parameters have to be defined for each transistor. SPICE assumes default values (which are often zero!) for the missing factors. When accuracy is an issue, it is essential to painstakingly define the value of parameters such as the drain and source area or resistance. The NRS and NRD values multiply the sheet resistance RSH specified in the transistor model for an accurate representation of the parasitic series source and drain resistance of each transistor.

---

### Example 2.12 SPICE MOSFET Model

The LEVEL 2 model for a 1.2  $\mu\text{m}$  CMOS process is included. Models are provided for both the NMOS and PMOS devices. This process will serve as the generic CMOS technology in

Table 2.3 Main SPICE MOSFET model parameters.

Parameter Name	Symbol	SPICE Name	Units	Default Value
SPICE Model Index		LEVEL	–	1
Zero-Bias Threshold Voltage	$V_{T0}$	VT0	V	0
Process Transconductance	$k'$	KP	A/V <sup>2</sup>	2.E-5
Body-Bias Parameter	$\gamma$	GAMMA	V <sup>0.5</sup>	0
Channel Modulation	$\lambda$	LAMBDA	1/V	0
Oxide Thickness	$t_{ox}$	TOX	m	1.0E-7
Lateral Diffusion	$x_d$	LD	m	0
Metallurgical Junction Depth	$x_j$	XJ	m	0
Surface Inversion Potential	$2 \phi_F $	PHI	V	0.6
Substrate Doping	$N_A, N_D$	NSUB	cm <sup>-3</sup>	0
Surface-State Density	$Q_{ss}/q$	NSS	cm <sup>-3</sup>	0
Fast Surface-State Density		NFS	cm <sup>-3</sup>	0
Total Channel Charge Coefficient		NEFF	–	1
Type of Gate Material		TPG	–	1
Surface Mobility	$\mu_0$	U0	cm <sup>2</sup> /V-sec	600
Maximum Drift Velocity	$v_{max}$	VMAX	m/s	0
Mobility Critical Field	$E_{crit}$	UCRIT	V/cm	1.0E4
Critical Field Exponent in Mobility Degradation		UEXP	–	0
Transverse Field Exponent (mobility)		UTRA	–	0

Table 2.4 SPICE Parameters for parasitics (resistances, capacitances).

Parameter Name	Symbol	SPICE Name	Units	Default Value
Source Resistance	$R_S$	RS	$\Omega$	0
Drain Resistance	$R_D$	RD	$\Omega$	0
Sheet Resistance (Source/Drain)	$R_{\square}$	RSH	$\Omega/\square$	0
Zero-Bias Bulk Junction Cap	$C_{j0}$	CJ	F/m <sup>2</sup>	0
Bulk Junction Grading Coeff.	$m$	MJ	–	0.5
Zero-Bias Side-Wall Junction Cap	$C_{jsw0}$	CJSW	F/m	0
Side-Wall Grading Coeff.	$m_{sw}$	MJSW	–	0.3
Gate-Bulk Overlap Capacitance	$C_{gb0}$	CGBO	F/m	0
Gate-Source Overlap Capacitance	$C_{gs0}$	CGSO	F/m	0
Gate-Drain Overlap Capacitance	$C_{gd0}$	CGDO	F/m	0
Bulk Junction Leakage Current	$I_S$	IS	A	0
Bulk Junction Leakage Current Density	$J_S$	JS	A/m <sup>2</sup>	1E-8
Bulk Junction Potential	$\phi_0$	PB	V	0.8



Table 2.5 SPICE transistor parameters.

Parameter Name	Symbol	SPICE Name	Units	Default Value
Drawn Length	$L$	L	m	—
Effective Width	$W$	W	m	—
Source Area	AREA	AS	m <sup>2</sup>	0
Drain Area	AREA	AD	m <sup>2</sup>	0
Source Perimeter	PERIM	PS	m	0
Drain Perimeter	PERIM	PD	m	0
Squares of Source Diffusion		NRS	—	1
Squares of Drain Diffusion		NRD	—	1

the rest of the book (in both examples, problem sets, and design problems). The presented models are assumed to be the defaults, unless otherwise specified.<sup>4</sup>

\* SPICE LEVEL 2 Model for 1.2  $\mu\text{m}$  CMOS Process

```
.MODEL NMOS NMOS LEVEL=2 LD=0.15U TOX=200.0E-10
+ NSUB=5.37E+15 VTO=0.74 KP=8.0E-05 GAMMA=0.54
+ PHI=0.6 U0=656 UEXP=0.157 UCRIT=31444
+ DELTA=2.34 VMAX=55261 XJ=0.25U LAMBDA=0.037
+ NFS=1E+12 NEFF=1.001 NSS=1E+11 TPG=1.0 RSH=70.00
+ CGDO=4.3E-10 CGSO=4.3E-10 CJ=0.0003 MJ=0.66
+ CJSW=8.0E-10 MJSW=0.24 PB=0.58
```

```
.MODEL PMOS PMOS LEVEL=2 LD=0.15U TOX=200.0E-10
+ NSUB=4.33E+15 VTO=-0.74 KP=2.70E-05 GAMMA=0.58
+ PHI=0.6 U0=262 UEXP=0.324 UCRIT=65720
+ DELTA=1.79 VMAX=25694 XJ=0.25U LAMBDA=0.061
+ NFS=1E+12 NEFF=1.001 NSS=1E+11 TPG=-1.0 RSH=121
+ CGDO=4.3E-10 CGSO=4.3E-10 CJ=0.0005 MJ=0.51
+ CJSW=1.35E-10 MJSW=0.24 PB=0.64
```

Examples of a typical NMOS and PMOS transistor definition are given below. Transistor M1 is an NMOS device with its drain, gate, source, and bulk terminals connected to nodes 2, 1, 0, and 0, respectively. Its gate length is the minimum allowed in this technology (1.2  $\mu\text{m}$ ). This leads to an effective gate length  $L_{\text{eff}}$  of 0.9  $\mu\text{m}$ , as the lateral diffusion equals 0.15  $\mu\text{m}$  (LD in the transistor model). The PMOS device, connected between nodes 2, 1, 5, and 5 (D, G, S, and B, respectively) is three times wider, which reduces the series resistance, but increases the parasitic diffusion capacitances as the area and perimeter of the drain and source regions go up.

```
M1 2 1 0 0 NMOS W=1.8U L=1.2U NRS=0.333 NRD=0.333
+ AD=6.5P PD=9.0U AS=6.5P PS=9.0U
M2 2 1 5 5 PMOS W=5.4U L=1.2U NRS=0.111 NRD=0.111
+ AD=16.2P PD=11.4U AS=16.2P PS=11.4U
```

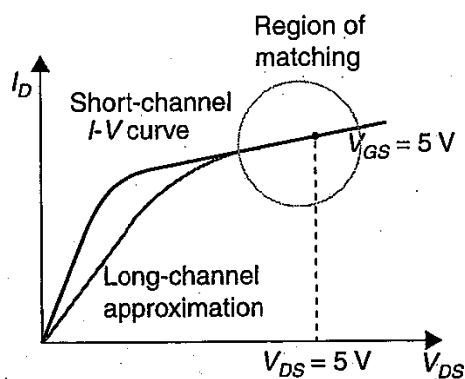
<sup>4</sup> Be aware that this represents a hypothetical model. The LEVEL 2 parameters were derived from the LEVEL 4 model of an actual process, and some deviations were introduced in the translation.

### Another Look at the Manual-Analysis Model

Even the simplified model for the velocity-saturated device (Eq. (2.57) to Eq. (2.59)) is too complex to be considered for manual analysis and would make the evaluation of any circuit with more than one transistor prohibitive. On the other hand, ignoring the short-channel effects completely leads to grossly inaccurate and overly optimistic results. Relying only on SPICE simulations leads to an *ad hoc* design strategy, where optimizations are made in a random order without a real understanding of the basic operation of the circuit and its prime parameters. As already established in the discussion on diodes, manual analysis is useful to derive a first-order solution and to build a conceptual understanding. A simple, yet reasonably accurate, model of the short-channel transistor is therefore essential.

A number of approaches to that goal are conceivable. A first one, adopted in this book, is to use the simple, long-channel equations for the transistor current (Eqs. (2.47) to (2.51)), but to adjust the dominant transistor parameters ( $k'$  and  $\lambda$ ) so that a reasonable approximation of the current is obtained in the regions that count the most. The performance of an MOS digital circuit is primarily determined by the maximum available current (i.e., the current obtained for  $V_{GS} = V_{DS} = \text{supply voltage}$ ). A good matching in this region is therefore essential.

This idea is illustrated in Figure 2.32 for a supply voltage of 5 V. Drawn is the  $I$ - $V$  curve of the short-channel device for  $V_{GS} = 5$  V. An approximative long-channel device is conceived, which yields the same current at  $V_{GS} = V_{DS} = 5$  V and whose current slope approximates the actual one in that particular region. This results in heuristic values for  $k'$  and  $\lambda$ , which on the average will yield reasonable approximations. Obviously, this model results in substantial errors when used in other regions such as for small values of  $V_{GS}$  or when the supply voltage is changed. For the latter case, a new set of heuristic parameters can be derived. Observe, also, that this model tends to yield pessimistic results, which is in general desirable.



**Figure 2.32** Retrofitting the long-channel model to short-channel devices (for a supply voltage of 5 V).

#### Example 2.13 Manual Analysis Model for 1.2 $\mu\text{m}$ CMOS Process

Using the LEVEL 2 SPICE model of Example 2.12, the drain-source current values are obtained from SPICE for both an NMOS and a PMOS device. Both devices are identical in size ( $W = 2.0 \mu\text{m}$ ,  $L = 1.2 \mu\text{m}$ ).

$$I_{DN}(V_{GS} = 5 \text{ V}, V_{DS} = 5 \text{ V}) = 0.514 \text{ mA}$$

$$I_{DN}(V_{GS} = 5 \text{ V}, V_{DS} = 4.5 \text{ V}) = 0.502 \text{ mA}$$

$$I_{DP} (V_{GS} = -5 \text{ V}, V_{DS} = -5 \text{ V}) = -0.212 \text{ mA}$$

$$I_{DP} (V_{GS} = -5 \text{ V}, V_{DS} = -4.5 \text{ V}) = -0.202 \text{ mA}$$

Plugging those numbers into the long-channel transistor model for the saturated region yields a set of simultaneous equations (for both NMOS and PMOS transistors) from which  $k'$  and  $\lambda$  can be determined.

$$I_{DN} = \frac{k'_n}{2} \left( \frac{W_n}{L_{eff}} \right) (V_{GS} - V_{T0})^2 (1 + \lambda V_{DS}) = \frac{k'_n}{2} \left( \frac{2}{0.9} \right) (5 - 0.743)^2 (1 + \lambda V_{DS})$$

and

$$I_{DP} = -\frac{k'_p}{2} \left( \frac{W_p}{L_{eff}} \right) (|V_{GS}| - |V_{T0}|)^2 (1 + \lambda |V_{DS}|) = -\frac{k'_p}{2} 2.22 (5 - 0.739)^2 (1 + \lambda |V_{DS}|)$$

Observe that  $L_{eff}$  (the effective transistor length, equal to  $L - 2x_d$ ) is used instead of the drawn length. The obtained parameters (which are equivalent to a SPICE LEVEL 1 model) are summarized in Table 2.6. Compare these values with their corresponding value in the LEVEL 2 model.

**Table 2.6** Retrofitted LEVEL 1 parameters for 1.2  $\mu\text{m}$  CMOS process.

	$V_{T0}$ (V)	$k'$ (A/V <sup>2</sup> )	$\lambda$ (V <sup>-1</sup> )
NMOS	0.743	$19.6 \times 10^{-6}$	0.06
PMOS	-0.739	$5.4 \times 10^{-6}$	0.19

Another reasonable approach is to use the model of Eqs. 2.57 to 2.59, but replace the value of  $\kappa$ , which is normally a function of  $V_{GS}$  (Eq. (2.59)) by a constant value. This is acceptable if the voltage-transition range is restricted [Toh88]. Others have modeled the effect of velocity saturation as an extra series resistance on the source side [Gray93]. However, the first approach is simple and produces acceptable results.

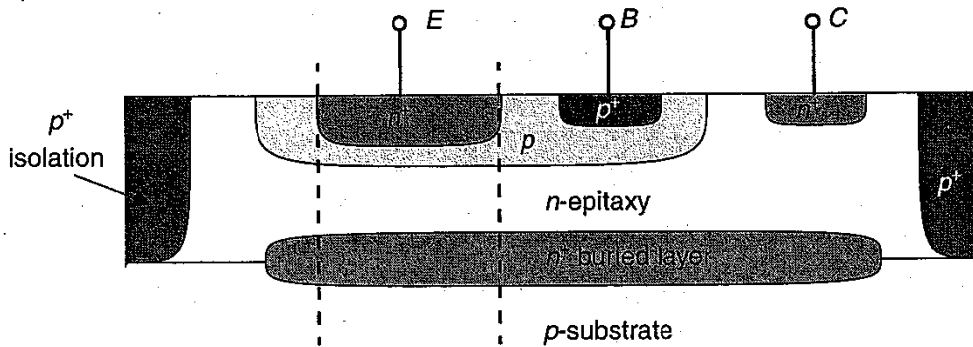
## 2.4 The Bipolar Transistor

MOS transistors took over the digital integrated circuit market in the 1970s, mainly as a consequence of their high integration density. Before that time, most digital gates were implemented in the bipolar technology. The dominance of the bipolar approach to digital design was exemplified in the wildly and widely successful TTL (Transistor-Transistor Logic) logic series, which persisted until the late 1980s. Although bipolar digital designs occupy only a small portion of the digital market at present, they still are the technology of choice when very high performance is required and are, therefore, worth studying.

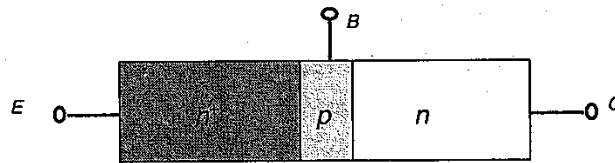
In line with the previous sections on diodes and MOSFETs, we discuss the static and the dynamic behavior of the bipolar transistor after a brief introduction to the device. We conclude with the presentation of second-order effects and simulation models.

2.4.1 A First Glance at the Device

Figure 2.33a shows a cross section of a typical *npn* bipolar (junction) transistor structure. The heart of the transistor is the region between the dashed lines and consists of two *np* junctions, connected back to back. In the following analysis, we will consider the idealized transistor structure of Figure 2.33b. The transistor is a three-terminal device, where the two *n*-regions, called the *emitter* and the *collector*, sandwich the *p*-type *base* region. In contrast to the source and drain regions of the MOSFET, the emitter and collector regions are not interchangeable, as the emitter is much more heavily doped than the collector.



(a) Cross-sectional view



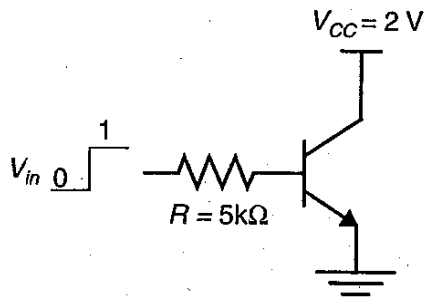
(b) Idealized transistor structure

Figure 2.33 *npn* bipolar transistor.

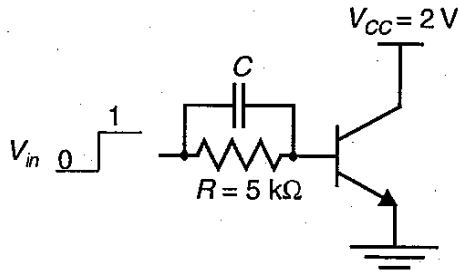
Depending upon the voltages applied over the device terminals, the emitter-base and collector-base junctions are in the *forward-* or *reverse-biased* condition. Enumeration of all possible combinations results in Table 2.7, which summarizes the operation modes of the bipolar device. In digital circuits, the transistor is operated by preference in the cut-off or forward-active mode. Operation in the saturation or reverse regions is, in general, avoided as the circuit performance in those regions tends to deteriorate.

Table 2.7 Modes of operation of the bipolar transistor.

Mode	Emitter Junction	Collector Junction
Cut-off	Reverse	Reverse
Forward-active	Forward	Reverse
Reverse-active	Reverse	Forward
Saturation	Forward	Forward



**Figure 2.61** Circuit for analyzing turn-on transient of bipolar transistor.



**Figure 2.62** Bipolar transistor circuit with speed-up capacitor.

- b. Plot  $R_B$  versus  $I_{CA}$  for the measured current values of Table 2.13. Use a log scale for  $I_{CA}$ . Assume  $I_S = 10^{-15}$  A, and  $\beta_F = 100$ .

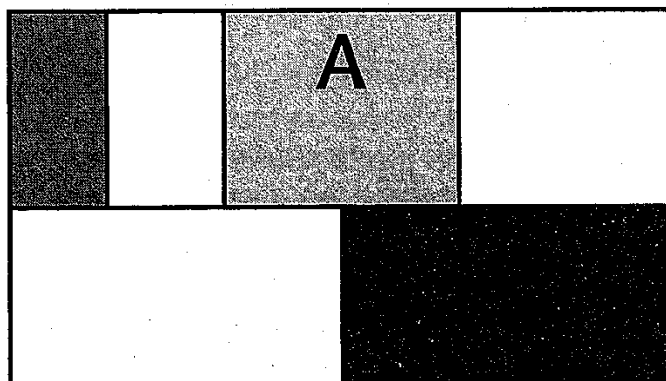
**Table 2.13** Table of measured collector currents

	$V_{BE}$ (V)	$I_{CA}$ (mA)
1	0.72	1.25
2	0.73	1.83
3	0.74	2.67
4	0.75	3.89
5	0.76	5.65
6	0.78	11.6
7	0.81	31.5
8	0.85	90.6

### DESIGN PROBLEM

Measure the  $I$ - $V$  characteristics of discrete MOS and bipolar transistors, and derive the first-order SPICE parameters. Compare the simulated characteristics with the measured ones.

## APPENDIX



# LAYOUT DESIGN RULES

*Creating a manufacturable layout*

The goal of defining a set of design rules is to allow for a ready translation of a circuit concept into an actual geometry in silicon. The design rules act as the interface between the circuit designer and the process engineer. As processes become more complex, requiring the designer to understand the intricacies of the fabrication process and interpret the relations between the different masks is a sure road to trouble.

Circuit designers in general want tighter, smaller designs, which lead to higher performance and higher circuit density. The process engineer, on the other hand, wants a reproducible and high-yield process. Design rules are, consequently, a compromise that attempts to satisfy both sides.

The design rules provide a set of guidelines for constructing the various masks needed in the patterning process. They consist of minimum-width and minimum-spacing constraints and requirements between objects on the same or on different layers.

The fundamental unity in the definition of a set of design rules is the *minimum line width*. It stands for the minimum mask dimension that can be safely transferred to the semiconductor material. In general, the minimum line width is set by the resolution of the patterning process, which is most commonly based on optical lithography. More advanced approaches use electron-beam or X-ray sources that offer a finer resolution, but are less attractive from an economical viewpoint.

Even for the same minimum dimension, design rules tend to differ from company to company, and from process to process. This makes porting an existing design between different processes a time-consuming task. One approach to address this issue is to use

advanced CAD techniques, which allow for migration between compatible processes. Another approach is to use *scalable design rules*. The latter approach, made popular by Mead and Conway [Mead80], defines all rules as a function of a single parameter, most often called  $\lambda$ . The rules are chosen so that a design is easily ported over a cross section of industrial processes. Scaling of the minimum dimension is accomplished by simply changing the value of  $\lambda$ . This results in a *linear scaling* of all dimensions. Such an approach, while attractive, suffers from some disadvantages:

1. Linear scaling is only possible over a limited range of dimensions (for instance, between 3  $\mu\text{m}$  and 1  $\mu\text{m}$ ). When scaling over larger ranges (for instance, into the submicron range), the relations between the different layers tend to vary in a nonlinear way that cannot be adequately covered by the linear scaling rules.
2. Scalable design rules are conservative. As they represent a cross section over different technologies, they have to represent the worst-case rules for the whole set. This results in overdimensioned and less-dense designs.

For these reasons, scalable design rules are normally avoided by industry. As circuit density is a prime goal in industrial designs, most semiconductor companies tend to use *micron rules*, which express the design rules in absolute dimensions and can therefore exploit the features of a given process to a maximum degree. Scaling and porting designs between technologies under these rules is more demanding and has to be performed either manually or using advanced CAD tools.

For small projects, fast prototyping, or educational use, on the other hand, the scalable design rules present a flexible and versatile design methodology. For this textbook, we have selected the MOSIS SCMOS (Scalable CMOS) scalable design rules as our preferred design medium for CMOS [Mosis]. The rest of this appendix is devoted to a short introduction and overview of the SCMOS rules (which we have slightly simplified for the sake of clarity).

As mentioned, in a  $\lambda$ -based design-rule set, all design rules are expressed as a function of  $\lambda$ . For a given process,  $\lambda$  is set to a specific absolute value, and all design dimensions are consequently translated into absolute numbers. Typically, the minimum line width of a process is set to  $2\lambda$ . For instance, for a 1.2  $\mu\text{m}$  process (i.e., a process with a minimum line width of 1.2  $\mu\text{m}$ ),  $\lambda$  equals 0.6  $\mu\text{m}$ .

A design-rule set now consists of the following entities: a set of interconnect layers, relations between objects on the same layer, and relations between objects on different layers. We discuss each of them in sequence.

## Layer Representation

The layer concept translates the intractable set of masks currently used in CMOS into a simple set of conceptual layout levels that are easier to visualize by the circuit designer. From a designer's viewpoint, all CMOS designs are based on the following entities:

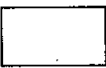
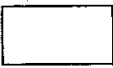
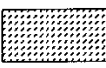
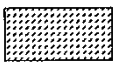
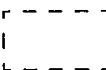
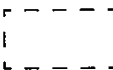
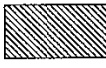
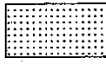




- *Substrates and/or wells*, being  $p$ -type (for NMOS devices) and  $n$ -type (for PMOS)
- *Diffusion regions* ( $n^+$  and  $p^+$ ) defining the areas where transistors can be formed. These regions are often called the *active areas*. Diffusions of an inverse type are



needed to implement contacts to the wells or to the substrate. These are called *select regions*.

- One or more *polysilicon* layers, which are used to form the gate electrodes of the transistors (but serve as interconnect layers as well).
- One or more *metal interconnect* layers (typically Al).
- *Contact* layers to provide interlayer connections.

A layout consists of a combination of polygons, each of which is attached to a certain layer. The functionality of the circuit is determined by the choice of the layers, as well as the interplay between objects on different layers. For instance, an MOS transistor is formed by the cross section of the diffusion layer and the polysilicon layer. An interconnection between two metal layers is formed by a cross section between the two metal layers and an additional contact layer. To visualize these relations, each layer is assigned a standard color (or stipple pattern for a black-and-white representation). The different layers used in the SCMOS process are represented in Figure A.1 (gray scale) or Colorplate 1 (color insert).

Layer	Representation	
well	 <i>p</i> -well	 <i>n</i> -well
active	 <i>n</i> <sup>+</sup>	 <i>p</i> <sup>+</sup>
select	 <i>p</i> <sup>+</sup>	 <i>n</i> <sup>+</sup>
polysilicon		
metal	 metal1	 metal2
contact holes	 m2p	 m2d
		 via

**Figure A.1** SCMOS layers and representations. m2p and m2d stand for metal1-to-poly and metal1-to-diffusion, respectively. See also Colorplate 1.

## Intralayer Constraints

A first set of rules defines the minimum dimensions of objects on each layer, as well as the minimum spacings between objects on the same layer. All distances are expressed in  $\lambda$ . These constraints are presented in a pictorial fashion in Figure A.2.

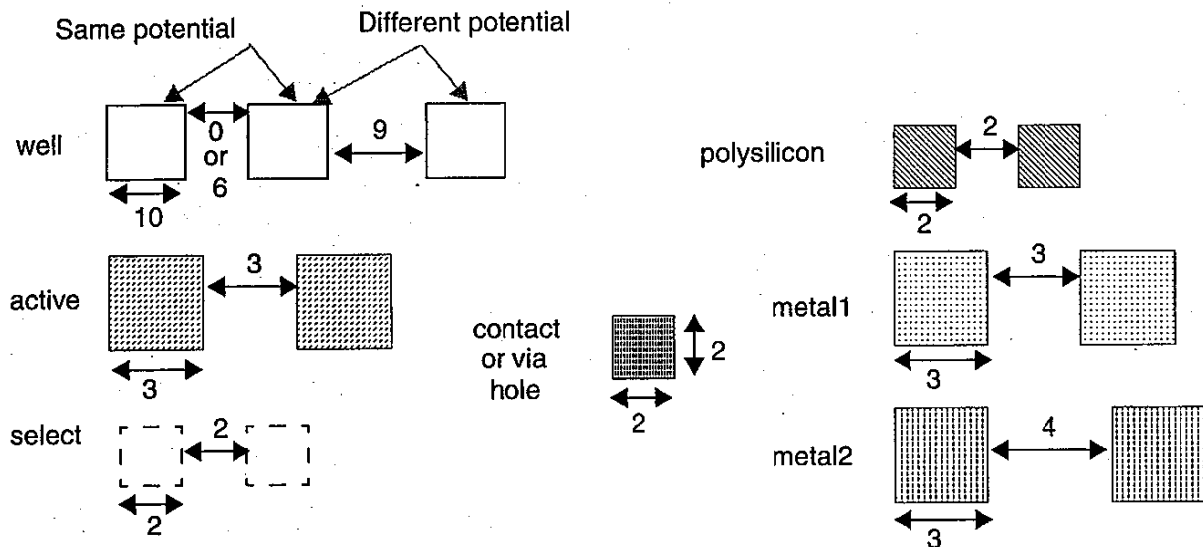


Figure A.2 Intralayer layout design rules: minimum dimensions and spacings. See also Colorplate 2.

## Interlayer Constraints

These rules tend to be more complex. The fact that multiple layers are involved makes it harder to visualize their meaning or functionality. Understanding layout requires the capability of translating the two-dimensional picture of the layout drawing into the three-dimensional reality of the actual device. This takes some practice.

We present these rules in a set of separate groupings.

1. **Transistor Rules** (Figure A.3). A transistor is formed by the overlap of the active and the polysilicon layers. From the intralayer design rules, it is already clear that the minimum length of a transistor equals  $2\lambda$  (the minimum width of polysilicon), while its minimum width is equal to  $3\lambda$  (the minimum width of diffusion). Extra rules include the spacing between the active area and the well boundary, the gate overlap of the active area, and the active overlap of the gate.
2. **Contact and Via Rules** (Figure A.4). A contact (which forms an interconnection between metal1 and active or polysilicon) or a via (which connects metal1 and metal2) is formed by overlapping the two interconnecting layers and providing a contact hole, filled with metal, between the two. In the SCMOS rules, both interconnecting layers have to extend at least one  $\lambda$  beyond the area of the contact hole, which sets the minimum size of a contact to  $4\lambda \times 4\lambda$ . This is larger than the dimensions of a minimum-size transistor! Multiple changes between interconnect layers are thus to be avoided. The figure, furthermore, points out the minimum spacings between contact and via holes, as well as their relationship with layers.

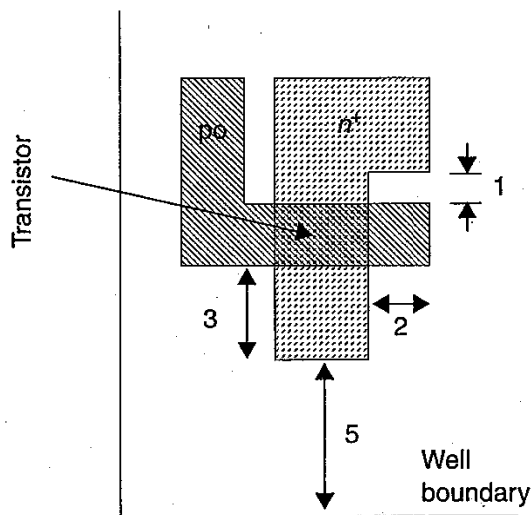


Figure A.3 Design rules concerning transistor layout. See also Colorplate 3.

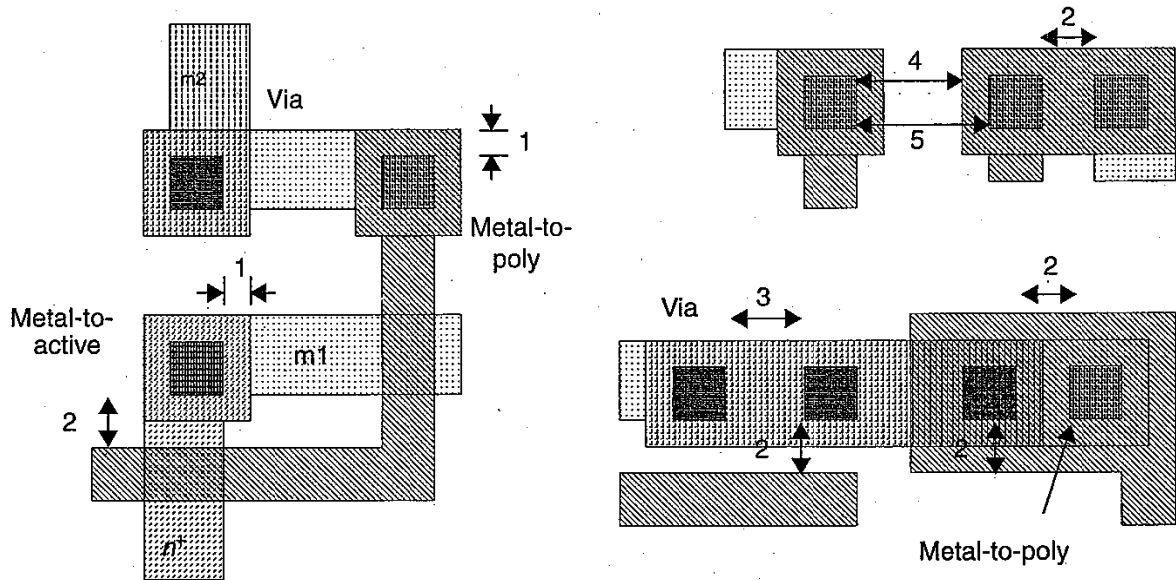
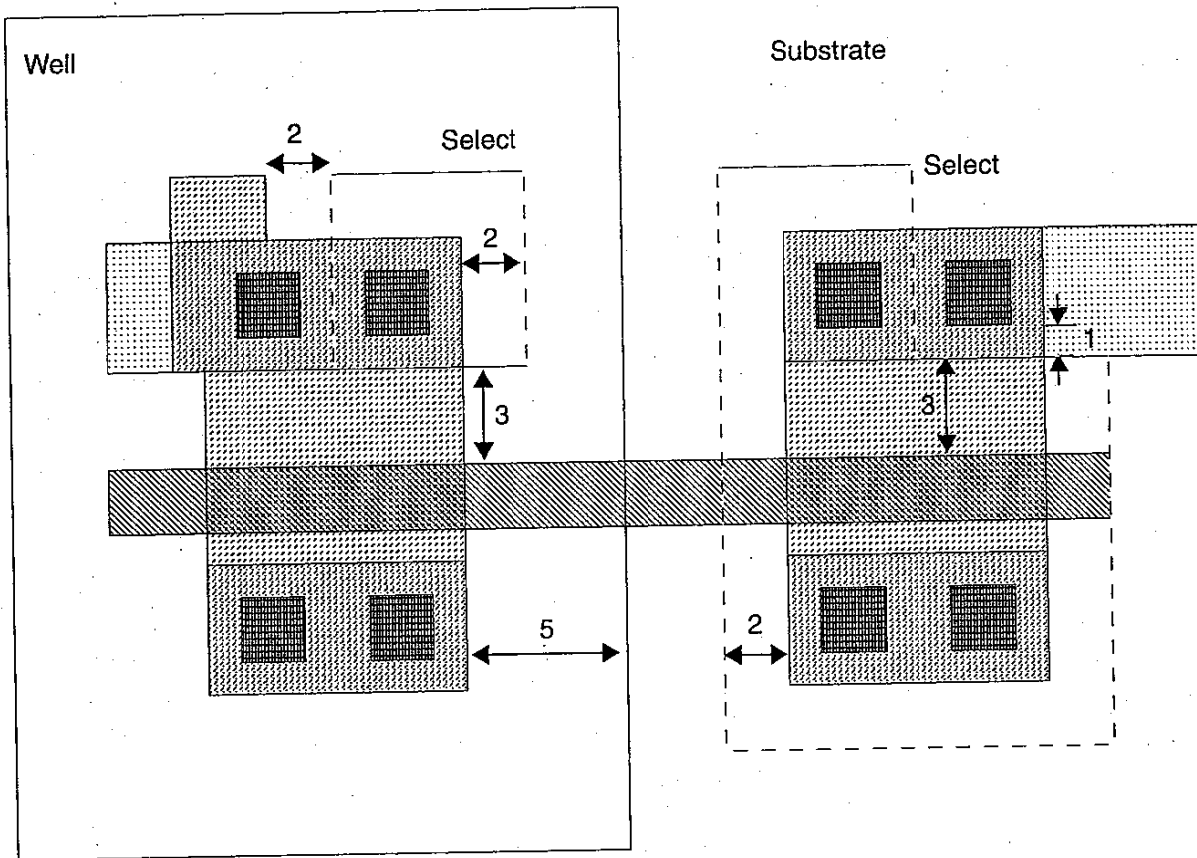


Figure A.4 Design rules regarding contacts and vias. Overlapping layers are marked by shaded regions. See also Colorplate 4.

*Well and Substrate Contacts* (Figure A.5). For robust digital circuit design, it is important for the well and substrate regions to be adequately connected to the supply voltages. Failing to do so results in a resistive path between the substrate contact of the transistors and the supply rails, and can lead to possibly devastating parasitic effects, such as latchup. It is therefore advisable to provide numerous substrate (well) contacts spread over the complete region. To establish an ohmic contact between a supply rail, implemented in metal1, and a  $p$ -type material, a  $p^+$  diffusion region must be provided. This is enabled by the *select* layer, which reverses the type of diffusion. A number of rules regarding the use of the *select* layer are illustrated in Figure A.5.

Consider an  $n$ -well process, which implements the PMOS transistors into an  $n$ -type well diffused in a  $p$ -type material. The nominal diffusion is  $p^+$ . To invert the polarity of the diffusion, an  $n$ -select layer is provided that helps to establish the  $n^+$  diffusions for the well-



**Figure A.5** Design rules regarding the select layer. See also Colorplate 5.

contacts in the  $n$ -region as well as the  $n^+$  source and drain regions for the NMOS transistors in the substrate.

Ensuring that none of the design rules is violated is a fundamental requirement of the design process. Failing to do so will almost surely lead to a nonfunctional design. Doing so for a complex design that can contain millions of transistors is no sinecure, especially when taking into account the complexity of some design-rule sets. While design teams used to spend numerous hours staring at room-size layout plots, most of this task is now done by computers. Computer-aided *Design-Rule Checking* (called *DRC*) is an integral part of the design cycle for virtually every chip produced today. A number of layout tools even perform *on-line DRC* and check the design in the background during the time of conception.

### Example A.1 Layout Example

An example of a complete layout is shown in Figure A.6. To help the visualization process, a vertical cross section of the process along the design center is included as well as a circuit schematic.

It is left as an exercise for the reader to determine the sizes of both the NMOS and the PMOS transistor, as well as the maximum current they can carry. Assume a supply voltage of 5 V. Also assume the design is implemented in the generic 1.2  $\mu\text{m}$  CMOS process described in Chapter 2.

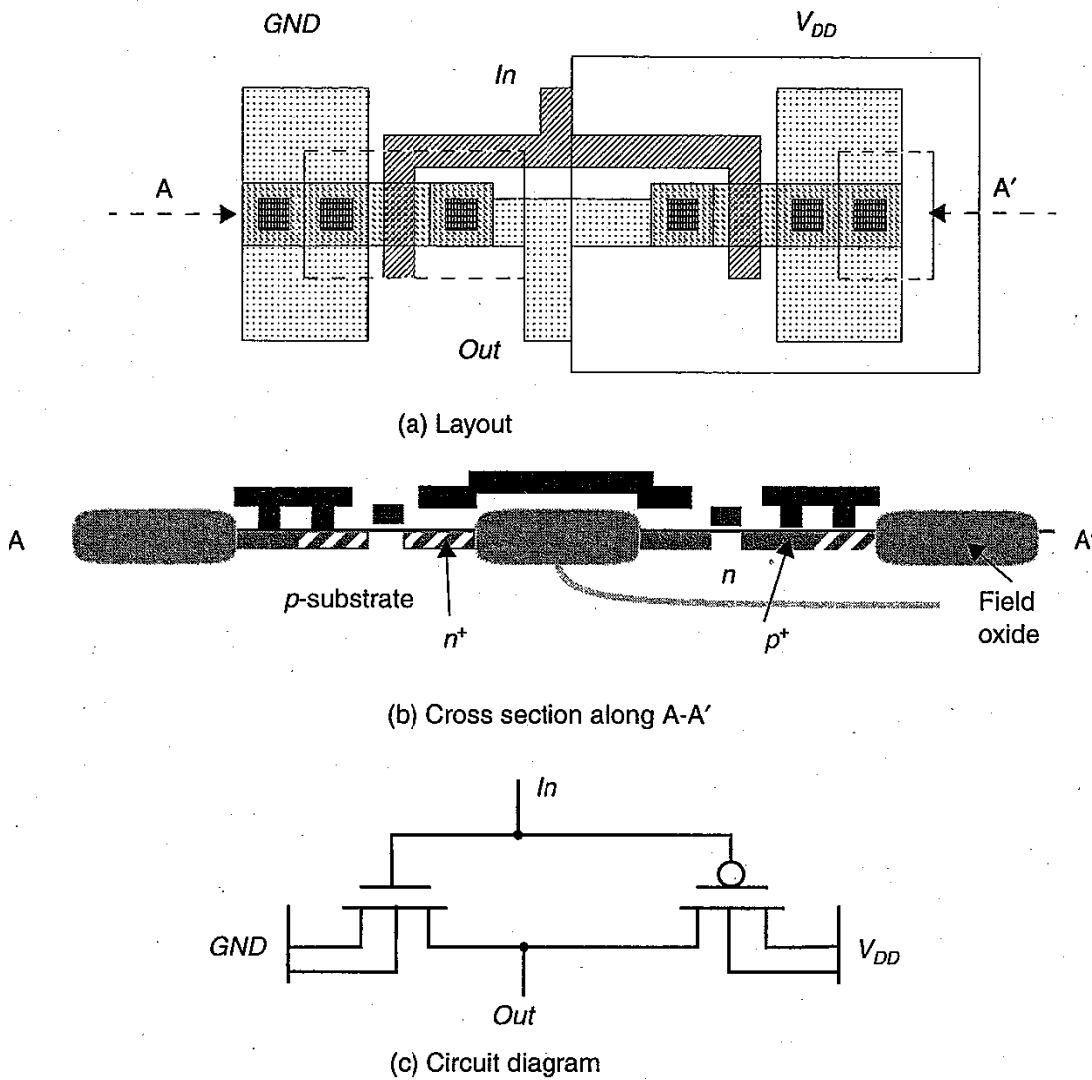


Figure A.6 A detailed layout example, including vertical process cross section and circuit diagram.

**To Probe Further**

Complete information on the Mosis SCMOS design rules is available via the Internet by sending electronic mail to [mosis@mosis.edu](mailto:mosis@mosis.edu). On the worldwide web, the information is available at the following address: <http://info.broker.isi.edu/1/mosis>. A more detailed version is also included in the instructor's manual and the www-page of the book.

For novice designers, extensive introductions on design rules can be found in the following references; [Wolf94] offers an especially comprehensive and well-illustrated discussion.

**REFERENCES**

[Mead80] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.  
 [Mosis] *MOSIS Scalable and Generic CMOS Design Rules*, Rev.6, ISI, February 1988.

[Weste85] N. Weste, and K. Eshragian, *Principles of CMOS VLSI Design: A Systems Perspective*, Addison-Wesley, 1985 (second edition 1993).

[Wolf94] W. Wolf, *Modern VLSI Design—A Systems Approach*, Prentice Hall, 1994.

### Exercises

1. [M&D, None, App. A] Figure A.7 shows an inverter layout containing several examples of either poor layout technique or design-rule violations. Find as many of these as possible and give a qualitative explanation of how each instance could be detrimental to the yield, functionality, or performance of the design. The layout is to scale and a one-lambda grid is provided in the figure.

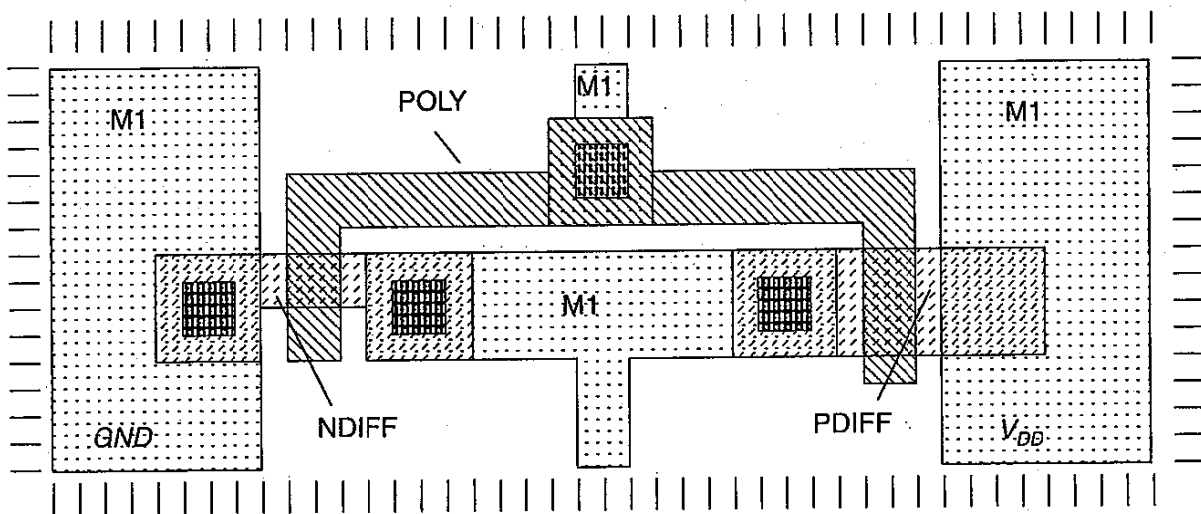
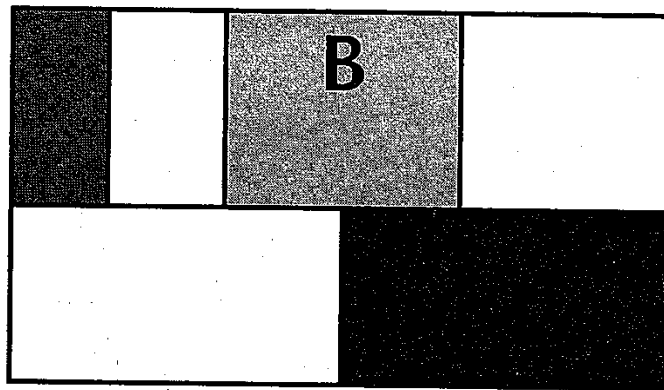


Figure A.7 Inverter layout with design-rule violations

## APPENDIX



# SMALL-SIGNAL MODELS

*Small-signal models of diodes, MOS, and bipolar transistors*

Circuits often operate with signal levels that are small compared to the bias currents and voltages. Under those conditions, a small-signal model can be employed that linearizes the device behavior and allows for the calculation of the circuit gain, terminal impedances, and frequency response without including the bias quantities.

Such models are especially useful in analog circuits, where signal excursions tend to be small compared to the supply voltage. In digital circuits, signal swings span a large fraction of the supply voltage, and the linearized, small-signal model is of a limited value. In spite of this observation, the linearized model comes in handy when analyzing some specific properties of a digital gate, for example, the calculation of the gain of a gate in the transient region.

A short overview of the small-signal models of the devices, which were introduced in Chapter 2, is therefore appropriate. The models are presented without derivation. We refer the interested reader to other textbooks such as [Sedra87] and [Gray93]. It suffices to state that the small-signal parameters represent the derivative of the device-current equations with respect to the controlling voltages ( $V_D$  for a diode,  $V_{GS}$  and  $V_{DS}$  for an MOS transistor, and  $V_{BE}$  and  $V_{CE}$  for a bipolar device). Observe also that the presented models are intended for the static analysis and do not include the capacitive effects. A dynamic model is obtained by adding the linearized small-signal capacitances that were derived for each device in Chapter 2.



### Diode

The static small-signal model of the diode is extremely simple and consists of a single resistance (Figure B.1).  $v_d$  represents the small-signal value of the diode voltage, that is, the deviation of the diode voltage from a bias voltage  $V_D$ . The value of the small-signal resistance  $r_o$  is given in Eq. (B.1).

$$r_o = \left( \frac{dI_D}{dV_D} \right)^{-1} = \frac{\phi_T}{I_D} \tag{B.1}$$

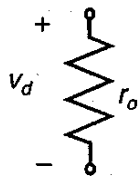


Figure B.1 Static small-signal model of diode.

### MOS Transistor

The (static) small-signal model of the MOS transistor is given in Figure B.2.  $v_{gs}$  (lowercase characters) stands for the small-signal value of the gate-source voltage, that is, the deviation of the gate-source voltage from a bias voltage  $V_{GS}$ . The values of the transconductance  $g_m$  and the output resistance  $r_o$  depend upon the operation region and the bias conditions. They are summarized in Table B.1. Observe that the presented model ignores the small-signal dependence on the substrate voltage (node  $B$ ), as this is not an issue in the context of this textbook.

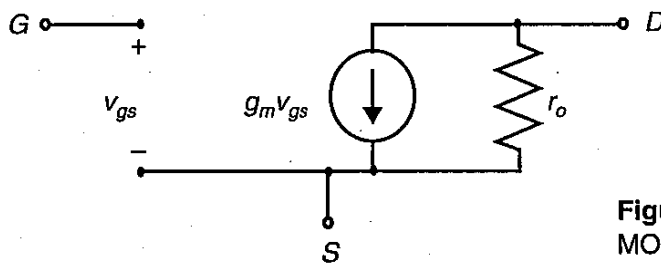


Figure B.2 Static small-signal model of MOS transistor.

Table B.1 Small-signal parameters of an MOS transistor.

	$g_m$	$r_o$
Linear	$kV_{DS}$	$[k(V_{GS} - V_T - V_{DS})]^{-1}$
Saturation	$k(V_{GS} - V_T)$	$1/\lambda_D$

### Bipolar Transistor

A simple static model, ignoring the parasitic resistances, is given in Figure B.3, where  $v_{be}$  (lowercase characters) stands for the small-signal value of the base-emitter voltage. The

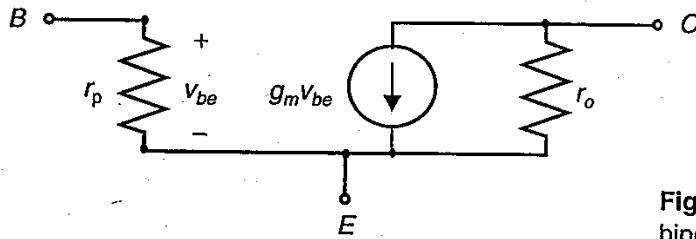


Figure B.3 Static small-signal model of bipolar transistor.

values of the transconductance  $g_m$ , the input resistance  $r_{\pi}$ , and the output resistance  $r_o$  depend upon the operation region and the bias conditions, and are summarized in Table B.2 for the forward-active region (the only region we will really use). Once again, the dynamic small-signal model can be obtained by adding the small signal-capacitances derived in Chapter 2.

Table B.2 Basic small-signal parameters of a bipolar transistor.

	$g_m$	$r_{\pi}$	$r_o$
Forward-active	$I_C/\phi_T$	$\beta_F/g_m$	$(V_A/\phi_T)/g_m$

## REFERENCES

[Gray93] P. Gray and R. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed., John Wiley and Sons, 1993.

[Sedra87] A. Sedra and K. Smith, *Microelectronic Circuits*, 2nd ed., Holt, Rinehart and Winston, 1987.

## Exercises

- [E, None, App. B] Table B.1 presents equations for the small-signal values  $g_m = \partial I_D / \partial V_{GS}$  and  $r_o = (\partial I_D / \partial V_{DS})^{-1}$  for an MOS device without derivation. Derive these expressions both for  $\lambda = 0$  and for nonzero  $\lambda$ 's.
- [E, None, App. B] Do the same as in Exercise 1 for the small-signal parameters of the bipolar device given in Table B.2.