1 Introduction to analog CMOS design

This chapter begins by explaining briefly why there is still a need for analog design and introduces its main tradeoffs. The need for accurate component modeling follows. Then, the essentials of p–n junctions and bipolar and field-effect transistors (FETs) for circuit design are recalled. The main differences between bipolar-transistor and FET operations are emphasized and drawbacks of some popular FET models for circuit design are commented on. Finally, two single-stage amplifiers, one in bipolar and another in MOS technology, are designed in order to make the differences between these technologies clear.

1.1. Analog design

1.1.1. The need for analog design

Analog circuits such as audio and radio amplifiers have been in use since the early days of electronics. Analog systems carry the signals in the form of physical variables such as voltages, currents, or charges, which are continuous functions of time. The manipulation of these variables must often be carried out with high accuracy.

On the other hand, in digital systems the link of the variables with the physical world is indirect, since each signal is represented by a sequence of numbers. Clearly, the types of electrical performance that must be achieved by analog and digital electronic circuits are quite different, and for this reason they are generally studied in separate university courses.

Nowadays, analog circuits continue to be used for direct signal processing in some very-high-frequency or specialized applications, but their main use is in interfacing computers to the analog world. The development of the very-large-scale-integration (VLSI) technology has led to computers being pervasive in telecommunications, consumer electronics, biomedicine, robotics, the automotive industry, etc. As a consequence, the analog circuits needed around them are also pervasive.

Interfacing computers or digital signal processors to the analog world requires various analog functions, among them amplification, filtering, sampling, (de)multiplexing, and analog-to-digital (A/D) and digital-to-analog (D/A) conversions. Since analog circuits are needed together with digital ones in almost any complex chip and the technology for VLSI is the complementary metal–oxide–semiconductor (CMOS), most of the current analog circuits are CMOS circuits. For this reason, this book is focused on CMOS analog circuits and the treatment of bipolar transistors is essentially limited to this introductory

chapter and to a few sections that cover the use of bipolar transistors compatible with CMOS technologies at appropriate points in the book.

The spread of analog circuits continues to increase with the evolution of technology. The current extensive research efforts directed toward sensors and actuators [1] for applications in numerous industrial products will lead to a demand for analog circuits in all subsequent products.

1.1.2. Tradeoffs in analog design

Analog circuits present a large variety of circuit functions, performance objectives, and specification methodologies [2]. The basic elements of modern analog circuits are MOS transistors, which are highly non-linear devices.

Even a simple cell as an operational amplifier (op amp) has many different but interrelated specifications including noise, distortion, power consumption, gain, phase margin, common-mode range, offset, temperature stability, supply sensitivity, and decoupling from other circuitry.

Numerous circuit topologies are often considered in the design phase. The broad choice of geometries and operating conditions for each device increases enormously the parametric complexity of this apparently simple design. The final designed circuit must satisfy the specifications considering the possible variations in the fabrication process, operating temperature, and power-supply voltage. As a consequence of the variety of performance objectives and of the design complexity, there are numerous op-amp designs available [3], each of them suitable for a specific situation.

The high degree of accuracy required and the difficulties involved in precisely modeling the device characteristics and interferences make the analog design problem even more complicated. Each design involves many complex, multi-variable interactions, and no widely applicable systematic design procedure is available [4].

Following the design, the actual performance of an analog circuit is dependent on the detailed electrical parameters of each of the many devices constituting the integrated circuit.

From the considerations summarized in this section, it is clear that analog-circuit-design expertise is not easy to achieve. To cope with the complexity of the analog design area, this book focuses on CMOS analog circuits and emphasizes the basic analog building blocks and their application to operational amplifier design.

1.1.3. The importance of component modeling

Additional difficulties in designing analog circuits come from inadequate device modeling [5] and poor knowledge of the technological parameters relevant to analog design, such as those related to excess noise and to device mismatch. Many of the device models used for hand design and for circuit simulation are inadequate for analog design and, sometimes, can cause a design to fail. This is particularly true for certain models of the MOSFET.

Analog circuits rely on details of the device characteristics to a much greater extent than digital circuits. In digital circuits operating with a relatively large supply voltage

(say above 3 V), transistors operate essentially as switches, and an approximate model is sufficient for the transition between the on and off states. For low-voltage digital circuits (say below 1 V), transistors must be considered as dimmers rather than switches; thus, an accurate model of the transistor for all the operating regions becomes necessary. Design of analog circuits requires, in general, very careful device modeling in all phases of the design procedure. An initial design with simple models is the first step in the design procedure. In this initial design, all transistor currents and sizes must be determined in order to satisfy the specifications. Transistor sizing and current levels can be easily derived from simple expressions, as we will show at the end of this chapter in Section 1.3.2. A systematic presentation of the design models is the subject of Chapter 2.

At the end of the design process, complicated and accurate MOSFET models are employed for design verification using circuit simulators such as SPICE or ELDO. MOSFET models for circuit simulation and the basic parameter extraction for design are the subjects of Chapter 11.

1.2. Bipolar and metal–oxide–semiconductor field-effect transistors

Transistors are semiconductor devices that constitute the basic building blocks of modern electronic circuits and systems. Transistors are essentially of two types: bipolar junction transistors (BJTs) and field-effect transistors (FETs). Analog bipolar circuits have been under development since the invention of the transistor. The first electronic products in bipolar technology, hearing aids and AM radios, appeared in the mid 1950s. The modern metal–oxide–semiconductor field-effect transistor (MOSFET) appeared in 1960 and MOS digital integrated circuits (ICs), memories, and microprocessors became available in the early 1970s. The first analog MOS circuits, converters, and switched capacitor filters were launched in the mid and late 1970s. Since MOS analog circuits began to be developed when analog bipolar technology was already mature, many concepts and techniques from analog bipolar technology were transferred to analog MOS circuits. Bipolar and MOS electronic circuits have many commonalities because transistors, whatever their type, have some basic common characteristics. One terminal, called the emitter or source, furnishes the carriers that are collected by a second terminal, called the collector or drain. The amount of carriers able to cross from the first to the second terminal is controlled by a third electrode, called the base in BJTs and the gate in FETs. The simplest model of either a BJT or an FET represents either transistor as a controlled current source. Although the similarities between transistors of different types are useful to facilitate understanding of the basic principle of operation of many circuits, for design purposes some superficial similarities can induce mistakes. Thus, for design, we must have a clear understanding and an accurate modeling of the specific behaviors of BJTs and FETs.

1.2.1. p–n Junctions

Let us begin with the p–n junctions [6]–[8], which are essential to the operation of bipolar and field-effect transistors. Electrons and holes with opposite electric charges are free to

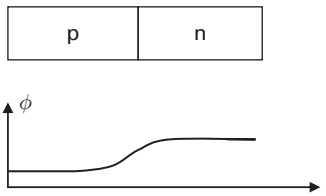

Fig. 1.1 The internal electrostatic potential of a p–n junction.

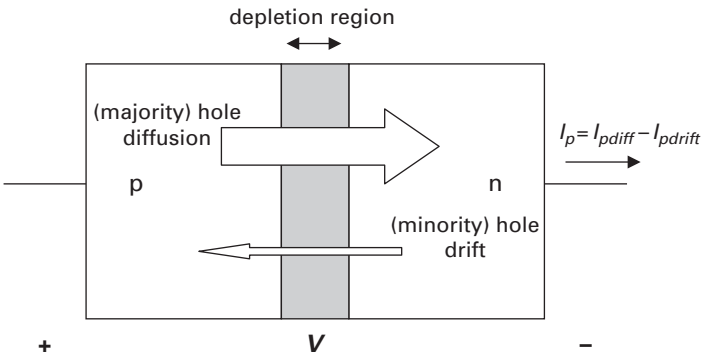

Fig. 1.2 Hole current in a forward-biased p–n junction.

move in semiconductors. Since free particles concentrate in the region where their energy is minimum, the electrostatic potential in the p-type region with plenty of holes, with positive charge, must be lower than that in the n-type region, where the electron concentration is high and holes are scarce. Similarly, electrons concentrate in the n-type region, where the potential is high and consequently their energy is low, because of their negative charge. The internal electrostatic potential in a p–n junction is shown in Figure 1.1.

In thermal equilibrium there is no net flow of carriers through the sample. Just a few holes on the p-side have enough energy to overcome the potential barrier and reach the n-side, where they recombine. Owing to these holes, a current I_{pdiff} flows but it is balanced by the current I_{pdrift} originating from the holes coming down the potential hill from the n-side. In equilibrium $I_{pdrift} = I_{pdiff} = I_{pS}$. A similar reasoning holds for the electrons. Now suppose that a positive bias V is applied to the p-region with respect to the n-region. The height of the barrier decreases by an amount V when the ohmic drops at the contacts and in the p- and n-regions are negligible. The potential-barrier region extends across the depletion region shown in Figure 1.2, where the electron (hole) carrier density is (well) below its value in the neutral n-region (p-region). The hole current injected from the p-region into the n-region is proportional to the number of holes with enough energy to overcome the potential barrier, and the reduction of the barrier by an amount V produces an exponential increase (given by the Boltzmann factor) in the number of holes with enough energy to cross the barrier. The hole current over the barrier will now be increased by the Boltzmann factor, i.e. $I_{pdiff} = I_{pS} \exp(V/\phi_t)$, where

$\phi_t \cong 26$ mV at 300 K is the thermal voltage. On the other hand, the hole current from the n-region will not have changed after the application of voltage V because it is dependent on the generation rate of hole–electron pairs which, in turn, is dependent on the local properties of the semiconductor near the junction.

As shown in Figure 1.2, the net hole current through the junction is

$$I_p = I_{pdiff} - I_{pdrift} = I_{pS}(e^{V/\phi_t} - 1). \tag{1.2.1}$$

The electron current is added to the hole current, giving the total as

$$I = I_p + I_n = I_S(e^{V/\phi_t} - 1), \tag{1.2.2}$$

where $I_S = I_{pS} + I_{nS}$ is the reverse saturation current, which is dependent both on the diode parameters and on the temperature.

The junction is a good rectifier of ac voltages higher than a few times ϕ_t. When the junction is forward-biased ($V > 0$), there is no limitation to the exponential increase in the current modeled by (1.2.2). A realistic model of the high-current region, however, must include other phenomena such as high-injection effects, series resistance, and self-heating. For a reverse-biased ($V < 0$) junction, the current is limited to I_S. A realistic model of the reverse-current region must include excess currents due both to recombination mechanisms and to high-field effects, which were not included in the simple model given above. Despite its limitations with regard to very high and very low currents, the idealized model of (1.2.2) is usually valid for a current range of up to six or seven orders of magnitude in junction diodes implemented with vertical n–p–n transistors, which is the subject of the next subsection.

1.2.2. Bipolar junction transistors

We will focus this subsection mainly on the planar n–p–n bipolar transistor [9] shown schematically in Figure 1.3, which is the most used component of the bipolar technology. The n–p–n intrinsic structure, where the main transistor action takes place, is shown by the dashed rectangle. To build the transistor dc model, we start with its operation in the so-called forward active mode, where the base–emitter junction is forward-biased and the collector–base junction is reverse-biased, as shown in Figure 1.4. Since the doping of the emitter region is much higher than that of the base region, the forward current of the base–emitter junction is constituted mostly by electrons injected from the emitter into the base. The hole current through the emitter–base junction is just a very small fraction of

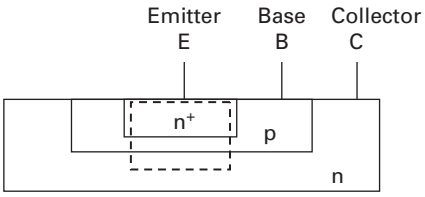

Fig. 1.3 Simplified structure of the ordinary vertical n–p–n transistor.

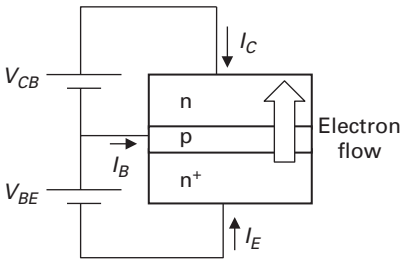

Fig. 1.4 An n–p–n bipolar junction transistor biased in the forward active region, i.e. the base–emitter junction is forward-biased and the base–collector junction is reverse-biased.

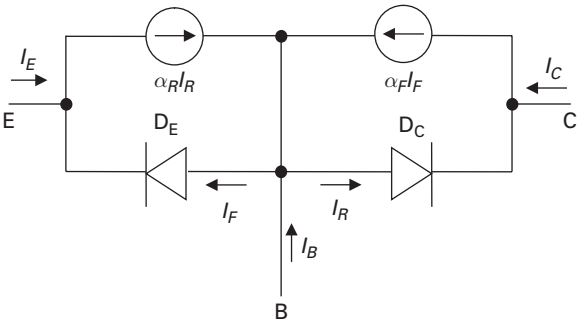

Fig. 1.5 The Ebers–Moll equivalent circuit of an n–p–n transistor.

the electron current. Most of the electrons coming from the emitter are attracted by the high potential of the collector region and diffuse through the base without recombining, since the base region is very thin. In the forward active mode, the potential barrier across the collector–base junction impedes the holes from being injected into the collector.

The collector current is a fraction of the emitter current since some of the electrons injected by the emitter into the base recombine without reaching the collector and, mainly, because a (small) part of the emitter current is composed of holes coming from the base. Thus,

$$I_C = -\alpha_F I_E, \tag{1.2.3}$$

where α_F ($\alpha_F < 1$) is the common-base current gain in the forward active mode.

The Ebers–Moll model [10] of the BJT represents the terminal currents of a bipolar transistor as a superposition of the effects of the two junctions, for all bias conditions, in terms of easily measurable transistor parameters. The equivalent circuit of the Ebers–Moll model of an n–p–n transistor is shown in Figure 1.5. I_F, the current through diode D_E, is the emitter current for $V_{BC}=0$, while I_R, the current through D_C, is the collector current for $V_{BE}=0$. The forward (I_F) and reverse (I_R) currents are described by

$$I_F = I_{ES}(e^{V_{BE}/\phi_t} - 1), \tag{1.2.4}$$

$$I_R = I_{CS}(e^{V_{BC}/\phi_t} - 1), \tag{1.2.5}$$

which are similar to (1.2.2) for the p–n junction.

The gains of the two current-controlled current sources in Figure 1.5 are α_F and α_R, the latter being the common-base current gain in the inverted mode, that is, with the collector operating as emitter and the emitter as collector. The terminal currents are

$$I_C = \alpha_F I_F - I_R, \tag{1.2.6}$$

$$I_E = \alpha_R I_R - I_F, \tag{1.2.7}$$

$$I_B = -(I_C + I_E) = (1 - \alpha_F) I_F + (1 - \alpha_R) I_R. \tag{1.2.8}$$

The Ebers–Moll model has only three independent parameters since the four parameters in the equations above are related by the reciprocity relation [9], [10]

$$\alpha_F I_{ES} = \alpha_R I_{CS} = I_S, \tag{1.2.9}$$

which holds as long as the minority-carrier density in the base is small compared with the thermal-equilibrium majority-carrier density. When the base–collector junction is reverse-biased, the combination of (1.2.4), (1.2.6), and (1.2.9) yields the familiar expression for the collector current in the forward active mode, i.e.

$$I_C \cong \alpha_F I_F = \alpha_F I_{ES} (e^{V_{BE}/\phi_t} - 1) \cong I_S e^{V_{BE}/\phi_t}. \tag{1.2.10}$$

Bipolar transistors are, in general, asymmetric devices, i.e. the collector and emitter regions are not interchangeable because they are optimized to operate in the forward active mode, typically having $\alpha_F > \alpha_R$. For the high-performance vertical BJT structure, the collector doping is much lower than the emitter doping, and typical values of the common-base current gains are $\alpha_F = 0.99$ and $\alpha_R = 0.65$. For the parasitic horizontal structure with the collector region surrounding the emitter region, common values of the common-base current gains are $\alpha_F = 0.98$ and $\alpha_R = 0.75$.

For an ideal device with common-base current gains $\alpha_F = \alpha_R = 1$, the base current is zero. Using (1.2.4) through (1.2.9) yields

$$I_C = -I_E = I_F - I_R = I_S (e^{V_{BE}/\phi_t} - e^{V_{BC}/\phi_t}). \tag{1.2.11}$$

The equation above clearly represents a (non-linear) symmetric device, i.e. one in which the output terminals are equivalent (or interchangeable). It is worth observing that the Ebers–Moll model also represents properly the not so common symmetric bipolar transistor. It is remarkable that the (more than) fifty-year-old Ebers–Moll model, although with some complements introduced by the Gummel–Poon model [11], continues to be the basic framework for BJT modeling and bipolar circuit design.

1.2.3. MOS field-effect transistors

The MOS structure, consisting of a metal–oxide (SiO_2)–semiconductor (Si) sandwich, is shown in Figure 1.6. In the common strong-inversion model of the field effect it is assumed that the MOS structure operates as a linear capacitor. For the case of the p-type substrate shown in Figure 1.6, the electron charge is assumed to be zero for gate voltages below the

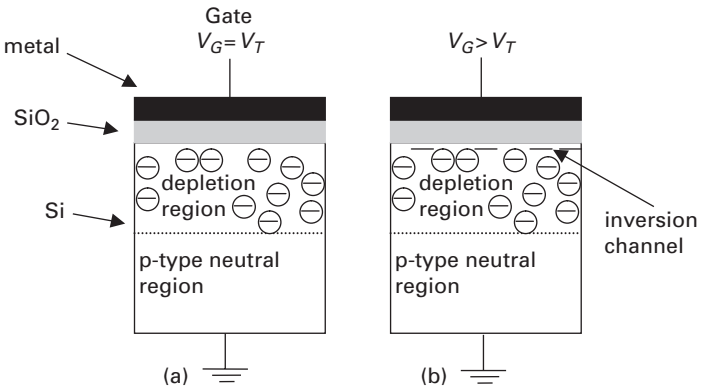

Fig. 1.6 The strong-inversion model of the field effect. (a) The MOS structure at threshold. The charge of the minority carriers (electrons) is negligible. The stored charge in the semiconductor is assumed to be due only to the depletion of majority carriers (holes). (b) The MOS structure above threshold. The depletion charge is assumed to be the same as at threshold while all the semiconductor charge variation is due to an electron (inversion) channel at the semiconductor surface.

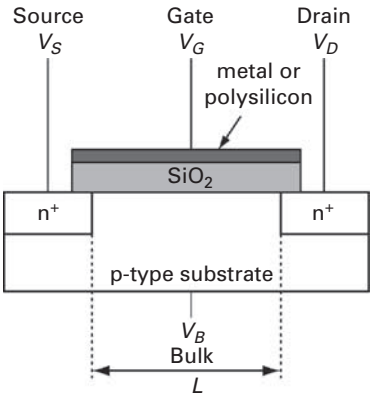

Fig. 1.7 An idealized enhancement-mode nMOS transistor. The width W is in the direction perpendicular to the page.

so-called threshold voltage and to increase linearly with the gate voltage for above-threshold operation. The basic flaw in the strong-inversion model is that the electron density follows Boltzmann's (exponential) law and, consequently, the electron charge cannot be zero for any finite applied voltage. Thus, the electron charge cannot vanish for subthreshold operation; consequently, the conventional definition of threshold cannot be applied in practice. Most of the numerous methods [12] developed to determine the threshold voltage are certainly due to the lack of a physically correct definition of threshold. Even so, since the strong-inversion model of the MOS transistor was for some time very popular, and still is, to a lesser extent, it should be summarized here and its shortcomings commented on.

A typical planar nMOS transistor structure is shown in Figure 1.7. The (parasitic) n–p–n bipolar transistor associated with this structure is inactive during the normal operation of

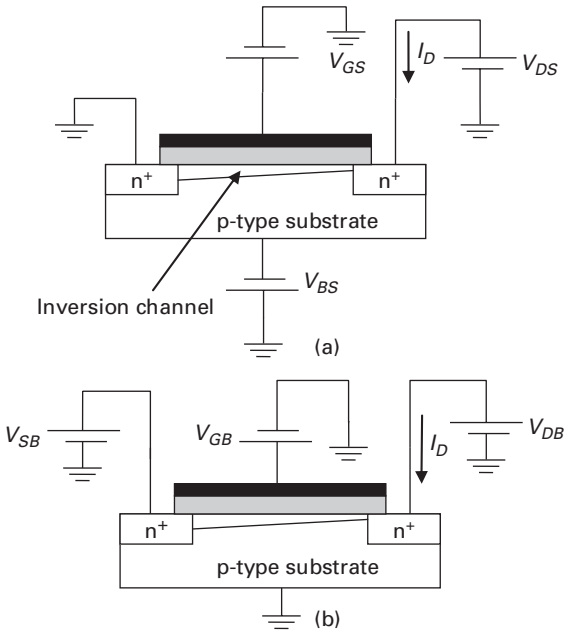

Fig. 1.8 Bias voltages applied to an nMOS transistor. (a) The common choice of the source as the reference terminal and (b) the choice of the substrate as the reference terminal. Since source and drain terminals are equivalent in the conventional MOS transistor, the choice of the substrate (or the gate) as the reference terminal preserves the symmetry between source and drain.

the MOSFET because the source–substrate and drain–substrate junctions are reverse-biased during the normal operation of the MOSFET. Conduction between source and drain occurs if an inversion (electron) channel connects them.

In the commonly used strong-inversion model [8], the applied voltages are referred to the source node, as Figure 1.8(a) shows. The inversion-channel charge density Q'_I above threshold is approximated as a linear function of the local potential to give

$$Q'_I = -C'_{ox}(V_{GS} - V - V_T), \tag{1.2.12}$$

where C'_{ox} is the oxide capacitance per unit area and V is the channel-to-source voltage, which varies from zero at the source to V_{DS} at the drain. To derive a first-approximation three-terminal model of the MOSFET, the threshold voltage V_T is assumed to be constant along the channel. As will be seen later on in this subsection, the four-terminal model is derived by including the effect of the fourth terminal (bulk) on the threshold voltage.

From expression (1.2.12) for the inversion charge density and through the integration of the equation for the electron current density along the channel assuming that diffusion is negligible, the drain current of the three-terminal device given by (1.2.14) is obtained.

For gate voltages below V_T the drain current is assumed to be zero, i.e.

$$I_D = 0 \quad \text{for } V_{GS} < V_T, \tag{1.2.13}$$

whereas

$$I_D = \beta\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}$$
$$\text{for } V_{GS} > V_T,\ V_{DS} < V_{DSsat} = V_{GS} - V_T. \tag{1.2.14}$$

It is important to note that the strong-inversion expression (1.2.14) holds only for the inequalities indicated. Note that, in this simplified model of the drain current in the three-terminal device, the transistor is represented by two parameters, namely the threshold voltage V_T and the transconductance parameter β given by

$$\beta = \mu C'_{ox}\frac{W}{L}, \tag{1.2.15}$$

where μ is the carrier mobility, W the channel width, and L the channel length.

For drain-to-source voltages above $V_{DSsat} = V_{GS} - V_T$, the drain current is assumed to saturate, i.e.

$$I_D = I_{Dsat} = \frac{\beta}{2}(V_{GS} - V_T)^2 \quad \text{for } V_{GS} > V_T,\ V_{DS} > V_{GS} - V_T. \tag{1.2.16}$$

In effect, the plot of (1.2.14) in the I_D–V_{DS} plane is a parabola with a maximum at $V_{DS} = V_{DSsat}$. For $V_{DS} > V_{DSsat}$ (1.2.14) has no physical meaning and, in a first-order approximation, the transistor is assumed to operate as an ideal current source, the value of the current being given by (1.2.16).

In the commonly used four-terminal model of the MOSFET, the threshold voltage is a function of the source-to-substrate bias V_{SB} given [8] by

$$V_T = V_{T0} + \gamma\left(\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}\right), \tag{1.2.17}$$

where ϕ_F is the Fermi potential, γ is the body-effect factor given by

$$\gamma = \frac{\sqrt{2q\varepsilon_s N_A}}{C'_{ox}}, \tag{1.2.18}$$

where $q = 1.6 \times 10^{-19}$ C is the electronic charge, ε_s is the permittivity of silicon, and N_A is the substrate doping, which is assumed to be uniform. The effect of the source–substrate bias is usually referred to as the body effect.

Expression (1.2.17) represents the effect of the depletion charge density at the source on the threshold voltage. In fact, for $V_{DS} \neq 0$, the depletion charge is non-uniform along the channel and, thus, the threshold voltage varies along the channel. Therefore, there is some inconsistency in using (1.2.17) in association with Equation (1.2.14) for the drain current since the latter has been deduced using the hypothesis that the depletion charge, and consequently V_T, is independent of the channel-to-source voltage V along the channel. As a consequence, the commonly employed model of the MOSFET which applies the body effect at the source to the whole channel does not preserve the physical interchangeability of source and drain.

To obtain a symmetric MOSFET model, the variation of the threshold voltage must be considered along the whole channel. To emphasize the symmetry between source and