1 Components of analog CMOS ICs

1.1 MOS transistors

The basic structure of an n-channel metal–oxide–semiconductor (NMOS) transistor built on a p-type substrate is shown in Fig. 1.1. The MOS transistor consists of two disjoint p-n junctions (source and drain), bridged by a MOS capacitor composed of the thin gate oxide and the poly-silicon gate electrode. If a positive voltage with sufficiently large magnitude is applied to the gate electrode, the resulting vertical electric field between the substrate and the gate attracts the negatively charged electrons to the surface. Once the electron concentration on the surface exceeds the majority hole concentration of the p-type substrate, the surface (the channel) is said to be inverted, i.e., a conducting channel is formed between the source and the drain. The carriers, i.e., the electrons in an NMOS transistor, enter the channel region underneath the gate through the source contact, leave the channel region through the drain contact, and their movement in the channel region is subject to the control of the gate voltage.

To ensure that both p-n junctions are continuously reverse biased, the substrate potential is kept lower than the source and drain terminal potentials. Note that the device structure is symmetrical with respect to the drain and source regions; the different roles of these two regions are defined only in conjunction with the applied terminal voltages and the direction of the drain current. In an n-channel MOS (NMOS) transistor, the source is defined as the n+ region which has a lower potential than the other n+ region, the drain. This means that the current flow direction is from the drain to the source. By convention, all terminal voltages of the device are defined with respect to the source potential.

The value of the gate-to-source voltage (V_{GS}) necessary to cause surface inversion (to create the conducting channel) is called the threshold voltage V_T. This quantity depends on various device and process parameters such as the work function difference between the gate and the substrate, the substrate (surface) Fermi potential, the depletion region charge concentration, the interface charge concentration, the gate oxide thickness and oxide (dielectric) permittivity, as well as the concentration of the channel implantation that is used to adjust the threshold voltage level.

If the applied gate-to-source voltage exceeds the threshold voltage of the MOS transistor, a sufficiently high concentration of electrons is achieved in the channel region, leading to surface inversion. Thus, an n-type conducting channel is formed between the source and the drain, which is capable of carrying the drain (channel) current. If a small positive voltage is applied to the drain, a current proportional to this voltage will start to flow from the drain to the source through the conducting channel. The effective

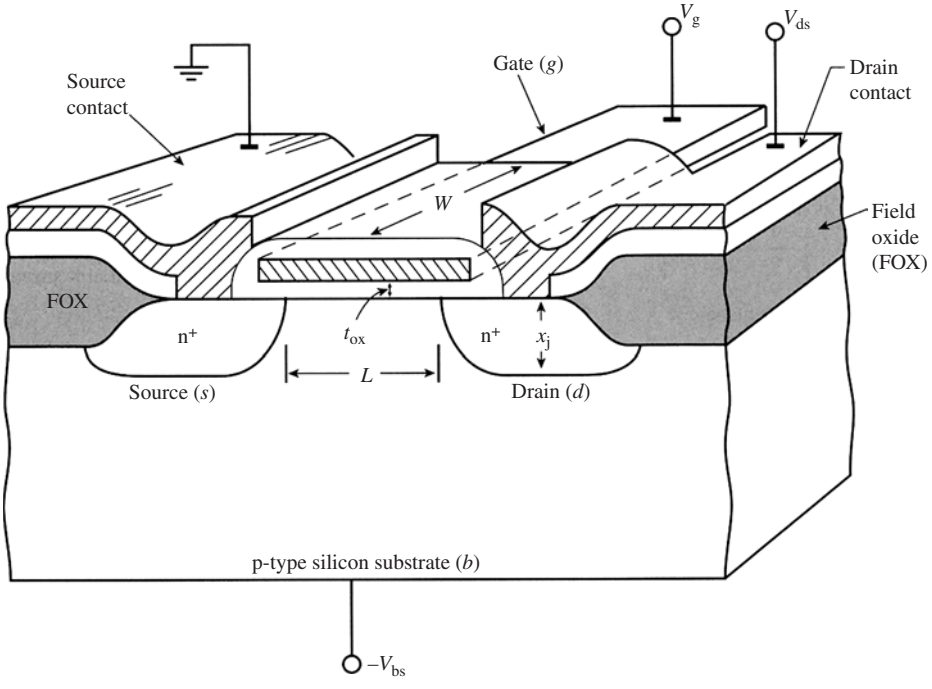

Figure 1.1 Simplified cross-section view of an n-channel MOS (NMOS) transistor (after Taur and Ning [13]).

resistivity of the continuous inversion layer between the source and the drain depends on the gate voltage. This operating mode is called the linear (or triode) mode, where the channel region acts as a voltage-controlled resistor. During this operating mode, the electron velocity in the channel is usually much lower than the drift velocity limit.

As the applied drain voltage is increased, the inversion layer charge and the channel depth at the drain end start to decrease. Eventually, when the drain voltage reaches a limit value called the saturation voltage ($V_{D(sat)}$), the inversion charge at the drain is reduced – theoretically – to zero, and the velocity of electrons – theoretically – reaches very high values, as discussed in the following sections. This event is named as the "pinch-off" of the channel. Beyond the pinch-off point, i.e., for drain voltage values larger than the saturation voltage, electrons travel in a very shallow pinched-off channel with a very high velocity, which is called the "saturation velocity". This operating regime is known as the saturation mode.

If the transistor is formed on an n-type substrate, using two p+ regions as source and drain, this structure is called a p-channel MOS (PMOS) transistor. In a PMOS transistor, the fundamental mechanisms of surface inversion and channel conduction are exactly the same as in NMOS transistors, although the majority of carriers consist of holes, not electrons. Thus, the gate-to-source voltage applied to the gate electrode to achieve surface inversion must be negative. Also, it should be taken into account that the hole mobility is considerably smaller than the electron mobility at room

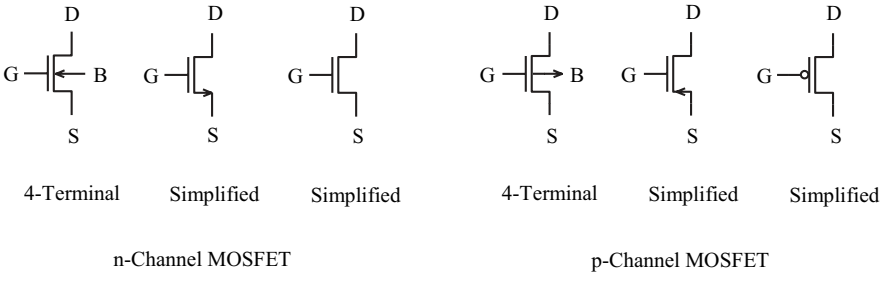

Figure 1.2 Commonly used circuit symbols for NMOS and PMOS transistors.

temperature, which leads to a smaller effective channel conductance for the PMOS transistor with the same channel dimensions. Nevertheless, the complementary nature of NMOS/PMOS biasing and operating conditions offers very useful circuit implementation possibilities, which underlines the importance and the wide-spread use of complementary MOS (CMOS) circuits in a very large range of applications.

Commonly used circuit symbols for n-channel and p-channel MOS transistors are shown in Fig. 1.2. While the four-terminal representation shows all external terminals of the device, the three-terminal symbol is usually preferred for simplicity. Unless noted otherwise, the substrate terminals are always assumed to be connected to the lowest potential for NMOS devices, and to the highest potential for PMOS devices.

1.1.1 Current–voltage relations of MOS transistors

The basic (so-called Level-1) current–voltage relations of a MOS transistor are given in most basic electronics textbooks. Since these relations contain a small number of parameters, they are convenient for hand calculations. The parameters of these expressions are:

- the mobility of electrons (or holes), μ;
- the gate capacitance per unit area, C_{ox};
- the threshold voltage of the transistor, V_T;
- the gate-length modulation coefficient, λ;
- and the aspect ratio of the transistor, (W/L).

In the following, these relations are derived with a different approach, to remind the reader of the fundamentals, and also to clarify the understanding of device behavior. In addition, the derivation presented here is based on a realistic profile of the channel-region inversion charge (as calculated from the fundamental electric field expressions), as opposed to the classical gradual channel approach which assumes linear charge profiles in the channel. This model represents the transistor under moderate to strong inversion conditions with reasonable accuracy for hand calculations, provided that the channel length is not too short and the transistor is not in the velocity saturation region.

For short-channel MOS transistors in which the carrier velocities reach saturation, i.e. approach a limit velocity, this model is no longer valid. Since a great majority of transistors realized in analog MOS integrated circuits today have channel lengths in

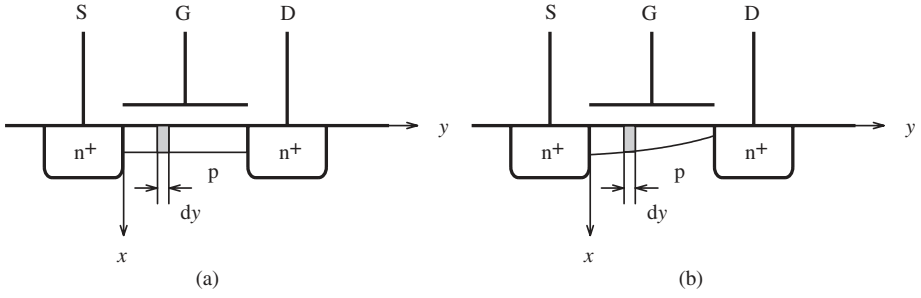

Figure 1.3 Inversion channel profiles of an NMOS transistor for (a) $V_{GS} > V_T$, $V_{DS} = 0$, (b) $V_{GS} > V_T$, $V_{DS} > 0$.

the sub-half-micron range, they may easily enter the velocity saturation region. Therefore it is necessary to derive rules to check whether a transistor is operating in the velocity saturation region or not, and to obtain expressions that are valid for velocity saturated transistors.

1.1.1.1 The basic current–voltage relations without velocity saturation

The cross-section of an NMOS transistor having an inversion layer along the channel owing to a gate–source voltage greater than the threshold voltage and zero source–drain voltage is shown in Fig. 1.3(a). Since there is no surface inversion for gate voltages smaller than the threshold voltage V_T, the value of the inversion charge density is

$$Q_i = -C_{ox}(V_{GS} - V_T)\ [\text{coulomb/cm}^2]$$

where $(V_{GS} - V_T)$ is the "effective gate voltage" for this case. The amount of the inversion charge of a transistor having a channel length L and a channel width W is

$$\bar{Q}_i = -C_{ox}WL(V_{GS} - V_T)$$

The minus signs in front of these expressions denote that this is a negative charge, since the carriers in the inversion layer of an NMOS transistor are electrons.

When we apply a positive drain voltage with respect to the source, a drain current (I_D) flows in the $-y$ direction and is constant along the channel (Fig. 1.3(b)). However, owing to the voltage drop on the channel resistance, the voltage along the channel is not constant. Although the effective gate voltage – inducing the inversion charge – at the source end of the channel is $(V_{GS} - V_T)$, it decreases along the channel and becomes equal to $(V_{GD} - V_T) = (V_{GS} - V_{DS} - V_T)$ at the drain end. If the channel voltage with respect to the source is denoted by $V_c(y)$, the effective gate voltage as a function of y can be written as

$$V_{eff}(y) = (V_{GS} - V_T) - V_c(y) \tag{1.1}$$

and the amount of the inversion charge in an infinitesimal channel segment dy is

$$d\bar{Q}_i(y) = -C_{ox}W[(V_{GS} - V_T) - V_c(y)]dy \tag{1.2}$$

Since the drain current is constant along the channel, for any y position the current can be expressed as

$$I_D = \frac{d\bar{Q}_i(y)}{dt} = \frac{d\bar{Q}_i(y)}{dy/v(y)} \tag{1.3}$$

where $v(y)$ is the velocity of electrons at position y, and can be expressed in terms of the electron mobility and the electric field strength at y:

$$v(y) = \mu_n E(y) = -\mu_n \frac{dV_c(y)}{dy} \tag{1.4a}$$

Using (1.2), (1.3) and (1.4a),

$$I_D = \mu_n C_{ox} W[(V_{GS} - V_T) - V_c(y)] \frac{dV_c(y)}{dy}$$

which gives the electric field strength as

$$E(y) = \frac{dV_c(y)}{dy} = \frac{I_D}{\mu_n C_{ox} W[(V_{GS} - V_T) - V_c(y)]} \tag{1.4b}$$

and

$$\frac{I_D}{\mu_n C_{ox} W} dy = [(V_{GS} - V_T) - V_c(y)] dV_c(y)$$

After integration from the source end ($y = 0$) to y we obtain

$$\frac{I_D}{\mu_n C_{ox} W} y = (V_{GS} - V_T) V_c(y) - \frac{1}{2} V_c^2(y) \tag{1.5}$$

From (1.5), the channel voltage $V_c(y)$ corresponding to a certain gate voltage and drain current can be deduced as

$$V_c(y) = (V_{GS} - V_T) \mp \sqrt{(V_{GS} - V_T)^2 - \frac{2I_D}{\mu_n C_{ox} W} y} \tag{1.6}$$

To fulfill the obvious physical condition $V_c(0) = 0$, the sign before the square-root term has to be taken as minus. The effective gate voltage is found from (1.1) and (1.6):

$$\begin{aligned} V_{eff}(y) &= (V_{GS} - V_T) - V_c(y) \\ &= (V_{GS} - V_T) \sqrt{1 - \frac{2I_D}{\mu_n C_{ox} W (V_{GS} - V_T)^2} y} \end{aligned} \tag{1.7}$$

It is useful to interpret (1.7) for certain cases.

(a) For $y = 0$ (at the source end of the channel) the effective channel voltage is $V_{eff}(0) = (V_{GS} - V_T)$, as expected.
(b) If $V_c(L) = V_{DS} = (V_{GS} - V_T)$, the effective channel voltage at $y = L$ is equal to zero and the channel is pinched-off at the drain end of the channel. For this case

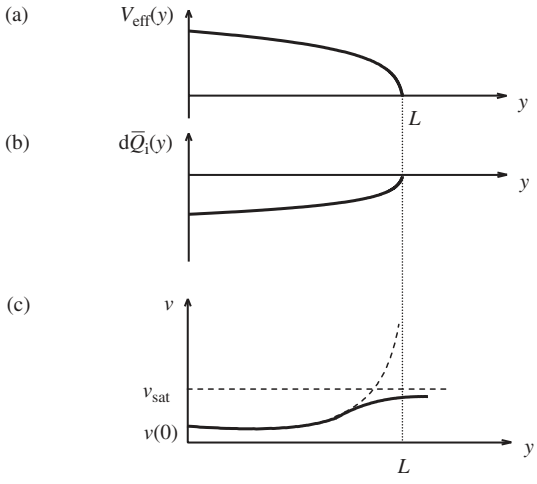

Figure 1.4 The variation of (a) the effective channel voltage, (b) the corresponding inversion charge and (c) the velocity of electrons along the channel for a transistor pinched-off at the drain end of the channel.

the value of the drain current can be found as

$$I_D = I_{D\,(sat)} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)^2 \tag{1.8a}$$

In this expression $\mu_n C_{ox}$ is a technology-dependent parameter and has the same value for all NMOS transistors on a chip.[1] For a given technology (1.8a) can be written as

$$I_D = I_{D\,(sat)} = \frac{1}{2}KP_n\frac{W}{L}(V_{GS} - V_T)^2 \tag{1.8b}$$

$I_{D(sat)}$ is called the "saturation current"[2] corresponding to a given gate voltage. The variations of the effective channel voltage along the channel and the corresponding inversion charge for a "saturated" transistor are plotted in Fig. 1.4(a) and (b), respectively, based on (1.2) and (1.7). Note that the inversion charge profile in Fig. 1.4(b) is found to be a second-order function of the distance y, and not a linear profile as usually assumed in the conventional gradual channel approach. While this does not influence the current–voltage relationship, the realistic charge profile will later be used for a more straightforward calculation of the channel capacitance.

[1] Similarly, $KP_p = \mu_p C_{ox}$ is a parameter common to all PMOS transistors on a chip.

[2] The phrase "saturation current" is also used to express the value of the drain current per 1 micron channel width for a certain technology, when the gate and the drain are both connected to the maximum permissible voltage for this technology. This "saturation current" takes into account all secondary effects discussed in the following sections, affecting the drain current. For example, for AMS 0.35 micron, 3.3 V technology the typical value of the saturation current is given as 540 μA/micron for NMOS transistors and 240 μA/micron for PMOS transistors, respectively.

It is useful to note and to interpret an important fact: the current remains constant along the channel but the electron density is decreasing. To maintain the current constant along the channel, i.e. to carry the same amount of charge in a certain time interval along the channel, the velocities of electrons have to increase from the source end to the drain end of the channel. Even more dramatically, the electron velocity theoretically has to reach infinity in the case of pinch-off of the channel, since the inversion charge decreases to zero. But it is known that the velocities of the electrons (and holes) cannot exceed a certain limit value and approach asymptotically this "saturation velocity", v_{sat}. In Fig. 1.4(c) the velocity of electrons along the channel is plotted. The dashed curve corresponds to the theoretical behavior, without any velocity limitation. The solid curve takes into account the velocity limitation. It is obvious that, owing to this limitation, the electron density does not decrease to zero but has a finite value to maintain the drain current with the limit velocity.

(c) If $V_c(L) = V_{DS} < (V_{GS} - V_T)$, the effective channel voltage is always positive along the channel. In other words, the channel does not pinch-off. From another point of view this can be interpreted such that the distance of the pinch-off point (L') is longer than the channel length. The drain current in this case can be solved from (1.6), for $y = L$ and $V_c(y) = V_{DS}$ as

$$I_D = \mu_n C_{ox} \frac{W}{L} \left[(V_{GS} - V_T).V_{DS} - \frac{1}{2} V_{DS}^2 \right] \tag{1.9a}$$

which reduces to (1.8a) for $V_{DS} = (V_{GS} - V_T)$, as expected. The variations of the effective channel voltage, the inversion (electron) charge density and the velocities of electrons along the channel are plotted, qualitatively, in Fig. 1.5.

For this "pre-saturation" region the variation of the drain current for small V_{DS} values can be written as

$$I_D \cong \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T) V_{DS} \quad \text{for } V_{DS} \ll (V_{GS} - V_T) \tag{1.9b}$$

This means that the drain current is proportional to the drain–source voltage. In other words the transistor acts as a resistor in this region. That is why this region is also called as the "resistive region".

(d) For $V_c(L) = V_{DS} > (V_{GS} - V_T)$, the transistor is in saturation. Assume that the transistor is pinched-off at the drain end of the channel and then the drain-source voltage increases by ΔV_{DS}. The effective channel voltage becomes equal to zero (the channel voltage becomes equal to $(V_{GS} - V_T)$) at a distance L' smaller than L. Since the current (I_D) and the electron velocity (v_{sat}) are constant in the interval $L' - L$, the field strength is approximately equal to the critical field strength, E_{cr}, which implies a linear variation of the potential along the pinched-off portion of the channel. Corresponding variations of the effective channel voltage and the charge density along the channel are shown in Fig. 1.6(a) and (b). Note that the critical field strength must have a value corresponding to the saturation

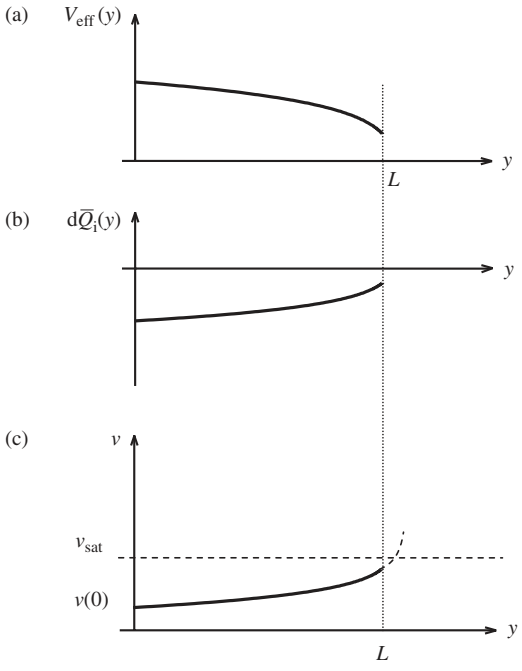

Figure 1.5 The variation of (a) the effective channel voltage, (b) the corresponding inversion charge and (c) the velocity of electrons along the channel for a transistor operating in the resistive (no pinch-off) region.

velocity, that is around $E_{cr} = 10^5$ V/cm for silicon.[3] Correspondingly, the velocity of electrons is equal to the saturation velocity along the pinched-off region of the channel, as shown in Fig. 1.6(c).

From these considerations we can conclude that

$$\Delta L = (L - L') = \frac{V_{DS} - (V_{GS} - V_T)}{E_{sat}} \tag{1.10}$$

The influence of the drain–source voltage on the drain current at a certain channel length can now be calculated:

$$I_{D\,(sat)} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)^2$$

$$\frac{dI_D}{dV_{DS}} = \frac{dI_D}{dL}\times\frac{dL}{dV_{DS}} = \left(-I_D\frac{1}{L}\right)\times\left(-\frac{1}{E_{sat}}\right) = \Lambda \,.\, I_D \tag{1.11}$$

[3] This field strength must not exceed the "breakdown field strength", which is approximately 3×10^5 V/cm for silicon. When the field strength approaches to this value, the silicon crystal structure tends to break down and the drain current increases further.

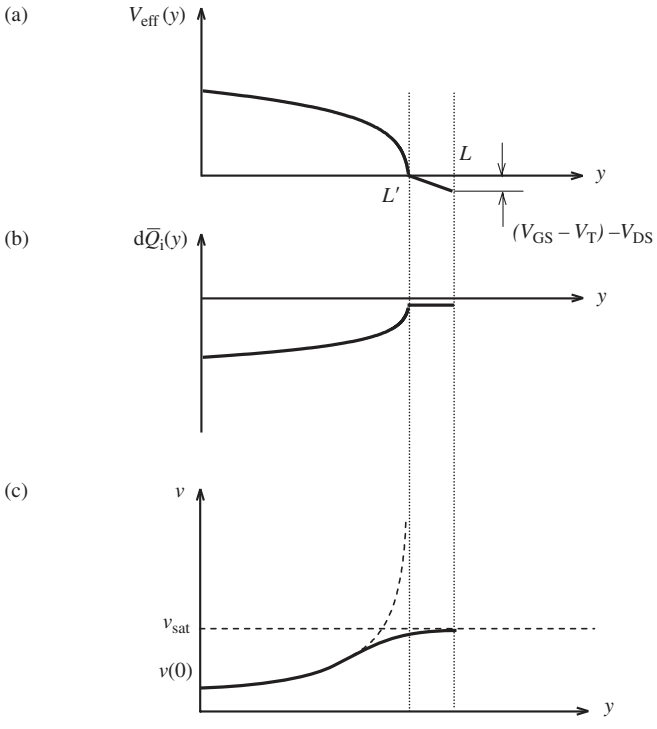

Figure 1.6 The variation of (a) the effective channel voltage, (b) the corresponding inversion charge and (c) the velocity of electrons along the channel for a transistor pinched-off before the drain end of the channel.

where

$$\Lambda = \frac{1}{L.E_{sat}} \tag{1.12}$$

Equation (1.11) gives the slope of the output characteristic curve corresponding to a certain V_{GS}, at the beginning of the saturation region, and is equal to the output conductance of the transistor (g_{ds}) for this point. Thus, the drain current corresponding to any drain–source voltage for the same gate–source voltage can be calculated as

$$I_D = I_{D\,(sat)} + g_{ds}(V_{DS} - V_{DS\,(sat)}) = I_{D\,(sat)} + g_{ds}[V_{DS} - (V_{GS} - V_T)] \tag{1.13}$$

From (1.12) and (1.13) the drain current can be written as

$$I_D = I_{D\,(sat)} \frac{1}{1 - \Lambda[V_{DS} - (V_{GS} - V_T)]} \cong I_{D\,(sat)}\{1 + \Lambda[V_{DS} - (V_{GS} - V_T)]\} \tag{1.14a}$$

Using simple linear relations as shown in Fig. 1.7, it is possible to express the drain current in terms of the conventional "lambda parameter" as

$$I_D = I_{D\,(sat)} \frac{1 + \lambda V_{DS}}{1 + \lambda(V_{GS} - V_T)} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T) \frac{1 + \lambda V_{DS}}{1 + \lambda(V_{GS} - V_T)} \tag{1.14b}$$

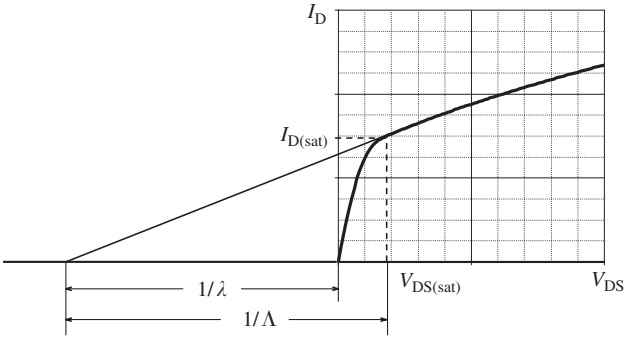

Figure 1.7 The output characteristic curve for a certain V_{GS} value and definitions of the Λ and λ parameters.

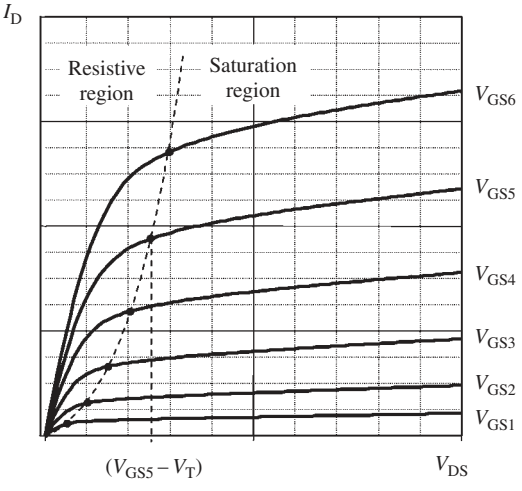

Figure 1.8 The output characteristic curves of a typical NMOS transistor. The border between the resistive region and the saturation region corresponding to $V_{DS} = (V_{GS} - V_T)$ is plotted as a dashed line.

For $V_{DS} \geq (V_{GS} - V_T)$, (1.14a) can be simplified as

$$I_D = I_{D\,(sat)}(1 + \lambda.V_{DS}) = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)(1 + \lambda.V_{DS}) \tag{1.15}$$

This expression is commonly used to model the "channel length modulation effect". However, it should be kept in mind that this approximation is not very accurate, especially for short-channel devices.

In Fig. 1.8 the output characteristic curves of an NMOS transistor covering these three characteristic features, namely the pre-saturation (or resistive) region, the onset of the pinch-off and the saturation region, are given. For PMOS transistors the