1 Active device models and layouts

1.1 MOSFET layouts

Layout description

The layout of a planar MOSFET including drain and source contacts is shown in the bottom-left corner of Fig.1.1. The longitudinal (AA') and transverse (BB') cross sections of the device are shown on the top and bottom-left corners, respectively. These pictures are representative of both the p and n devices. For this reason, the type of doping is not specified. Clearly, doping of the drain/source regions are of opposite type with respect of the body region.

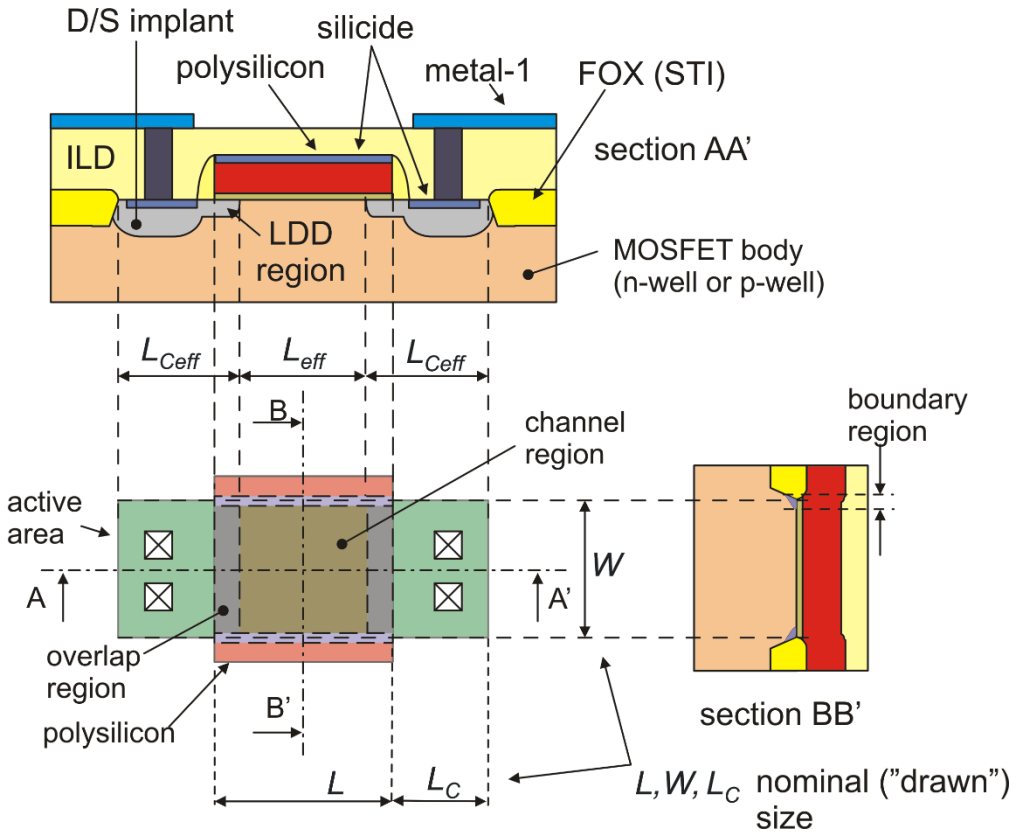

Fig. 1.1. Planar MOSFET layout with longitudinal (AA') and transverse (BB') cross-sections. Metal 1 and D/S implant layers are not shown in the layout for simplicity.

Ideally, a MOSFET is formed when an active area is crossed by a polysilicon line. The width of the polysilicon line determine the channel length (L), while the width of the active area is the channel width (W). The source and drain areas coincides with the two portions of the active area that are not covered by polysilicon and face each other across the channel. The drain/source regions are nominally rectangles

of L_C size along the longitudinal direction and W size along the transverse one. Dimension L_C can be indicated as drain/source length. Note that, generally, L_C is set to the minimum value allowed by the design rules, in order to minimize parasitic junction capacitances. The drain/source areas are filled with the maximum number of contacts allowed by the design rules in order to reduce the drain/source series resistance.

In a real device, extension of the drain-source doped regions under the gate reduces the effective channel length (L_{eff}) with respect to the nominal (drawn) value (L). For the same reason, the actual drain/source lengths (L_{Ceff}) are longer than L_C. Note that the drain/source doped areas extends under the gate mainly with their LDD (Light Doping) portions.

Similarly, the effective width of the MOSFET (W_{eff}) is slightly different from the drawn geometry (W). The reason is less intuitive than for the channel length reduction. The effect is caused by the presence of boundary regions located at the interface of the active area with the field oxide. In these points, the distance between the polysilicon gate and the MOSFET body gradually increases as the polysilicon line goes out of the active area. As a result, there are channel portions (within the drawn width W) where inversion is less effective (smaller density of mobile charge) and zones, beyond the nominal width, where the gate induces depletion charges into the substrate (fixed charge). The presence of these boundary regions generally result in reduction of the effective width ($W_{\text{eff}} < W$).

Generally, it is not necessary for the layout designer to draw all the masks required to complete the MOSFET fabrication. For example, in many process PDK, the spacer oxide and the LDD doping is automatically drawn when the polysilicon, active area and drain / source doping layers are drawn. Simplified layout for the n-MOSFET and p-MOSFET are shown in Fig. 1.2 and Fig.1.3, respectively.

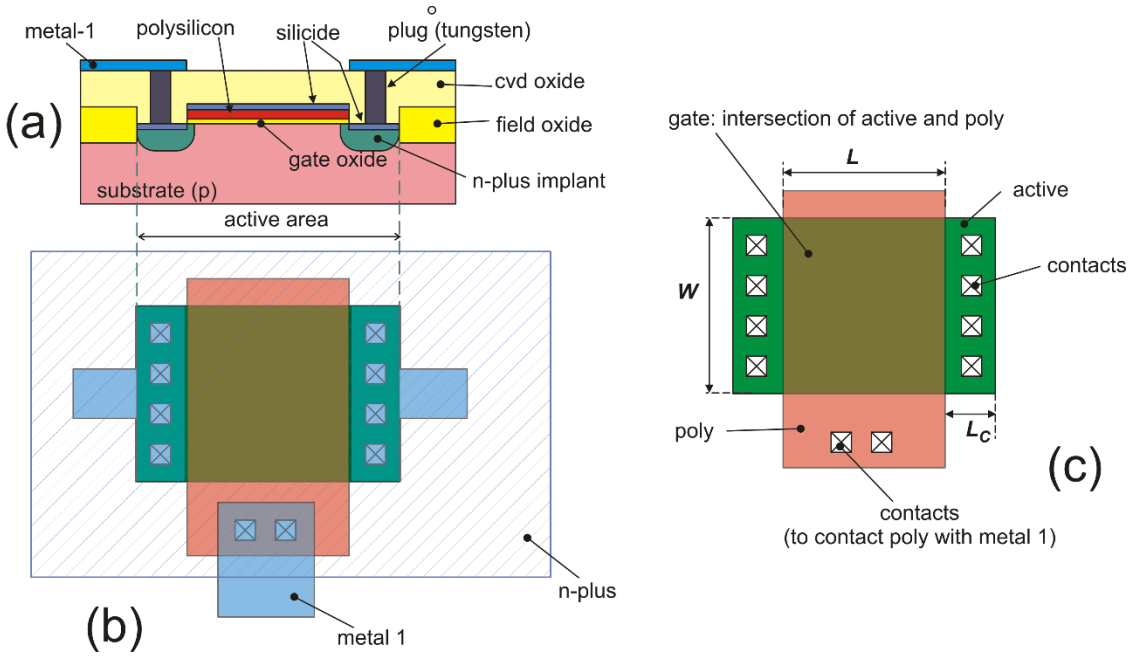

Fig.1.2. Simplified cross-section (a) and layout (b) of an n-MOSFET. An extract of the layout showing only active, poly and contact layer is proposed in (c), with indication of the most relevant lengths.

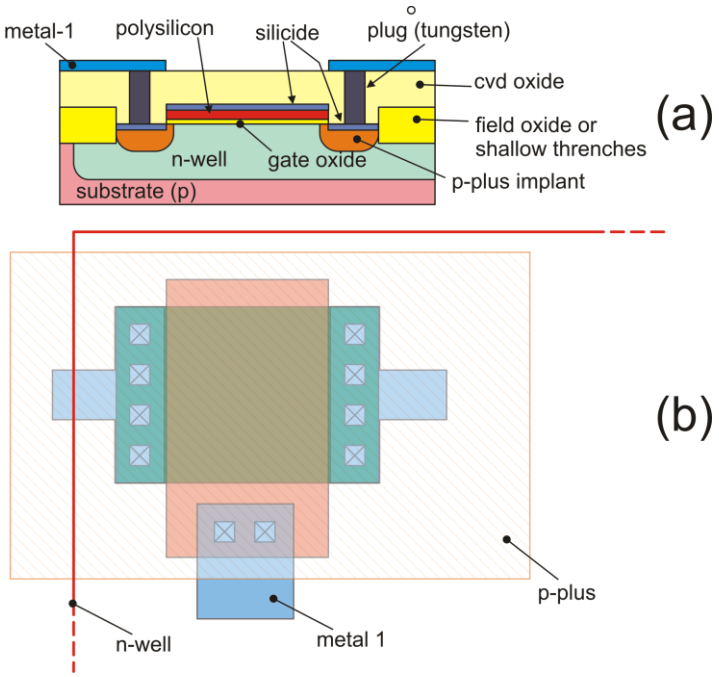

Fig.1.3. Simplified cross-section (a) and layout (b) of a p-MOSFET. The difference with respect to the n-MOSFET layout is the p-plus doping (in place of n-plus) and enclosure into an n-Well.

Giving the designer the minimum number of layers that are necessary to define a standard layout is not the only possible choice. There are foundries that give more control to the designer, including also layers that refer to process steps that can be generated automatically.

In a PDK oriented to analog design, MOSFETs are placed into the layout by creating an instance of the corresponding p-cell, which coincides with the layout view of the device. Through a graphical interface, it is possible to choose the *W* and *L* of the device being placed. This procedure generates all required layers, including contacts. The role of the layout designer is then arranging the devices inside the assigned area and making all connections.

Designer options

Processes may offer different MOSFET families as an option: for example, low threshold devices, or devices with increased oxide thickness for handling higher logical levels may be available as an alternative to the regular complementary (n and p) devices. For any MOSFETs that the designer places into the circuit, the width (W) and length (L) should be specified. Another degree of freedom that is under control of the designer is the type of layout. For example, instead of the simple layouts shown in Fig.1.2 and 1.3, it is possible to use fingered layouts, which are particularly convenient for MOSFETs with large W values.

1.2 Mosfet models

In this part, we will consider an n-MOSFET (enhancement type), since most quantities (currents and voltages) are positive for this type of device. Type-p devices will be briefly reviewed at the end of this

document. The large signal model of an n-MOSFET is shown in Fig.1.4. The core of the device is the I_{DS} controlled source. Resistors R_S and R_D, connect the edge of the channel, S' and D', to the actual source and drain contacts, S and D, respectively. Diodes D_{BS} and D_{BD} represent the junctions that isolate the source and drain from the body (B). The various capacitances are generally characterized by a non-linear relationship between voltage and charge. In the following part of this document, the series resistance will be neglected for simplicity. Then, we will consider that S≡S' and D≡D'.

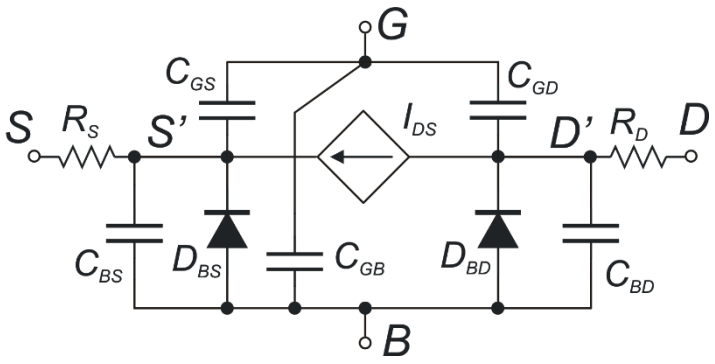

Fig. 1.4. Large signal equivalent model of an n-MOSFET

Control voltages and operating regions.

The drain-current I_{DS} depends on the voltages across the MOSFET terminals. There are two possible ways to define these voltages: source-referred or body-referred. Source referred potentials are more commonly used and consists of V_{GS}, V_{BS} and V_{DS}. Body-referred voltages are V_{GB}, V_{SB} and V_{DB}. Obviously, the two approaches are equivalent, since it is always possible to transform body-referred potentials into source referred ones and vice versa. However, source-referred potentials are simpler and more intuitive. The only real drawback of source-referred potentials is that, for a symmetrical MOSFET, it is necessary to determine which terminal is working as the source before writing down the current equations. We have to consider the voltages of the two terminals that are candidate to be the source or the drain and apply the following rule:

- in an n-channel device, the source is the terminal that has the lower potential;
- in a p-channel device, the source is the terminal that has the higher potential.

The voltage V_{BS} should be such that both the body-source and body-drain diodes do not carry a significant current. The ideal situation is to make sure that both diodes are reverse-biased or, at least zero-biased. For an n-MOSFET, this means:

$$V_{BS} \le 0 \Rightarrow V_S > V_B \tag{1.1}$$

Voltage V_{BS} affects the threshold voltage V_t through the body effect [1]:

$$V_t = V_{t0} + \gamma\left(\sqrt{\phi_s - V_{BS}} - \sqrt{\phi_s}\right) \tag{1.2}$$

where V_{t0} is the threshold voltage for V_{BS}=0, whereas γ and ϕ_s are equal to:

$$\phi_s = 2\psi_B = 2\frac{k_B T}{q}\ln\left(\frac{N_A}{n_i}\right); \qquad \gamma = \frac{\sqrt{2q\varepsilon_{Si}N_A}}{C_{ox}} \tag{1.3}$$

and: k_B=Boltzmann constant, T=absolute temperature, q = electron charge, N_A= dopant concentration of the body, e_{Si} the silicon permittivity and C_{ox} the gate capacitance-per-unit area.

It is useful to calculate the sensitivity of V_t with respect to V_{BS}. Due to (1.1), it is convenient to express the sensitivity with respect to $V_{SB}= -V_{BS}$, obtaining [2]

$$\frac{dV_t}{dV_{SB}} = m-1; \quad \text{with} \qquad m = 1+\frac{C_{dm}}{C_{ox}} \tag{1.4}$$

where m is the body-effect factor and C_{dm} is the depletion layer capacitance given by:

$$C_{dm} = \sqrt{\frac{q\varepsilon_{Si}N_A}{2\left(2\phi_f + V_{SB}\right)}} \tag{1.5}$$

The value assumed by m for V_{BS}=0 varies from 1.2 to 1.3, so that dV_t/dV_{SB} is in the range 0.2-0.3. The threshold voltage to body bias sensitivity decreases for increasing reverse voltages.

As far as V_{GS} and V_{DS} are concerned, six regions can be roughly distinguished, as shown in Table 1.1, where V_T is k_BT/q. Note that V_{GS} appears through the quantity (V_{GS}-V_t) which is called "overdrive voltage"

Table 1.1. MOSFET operating regions on the basis of V_{GS} and V_{DS}.

	$V_{GS}-V_t \le 0$	$0 \le V_{GS}-V_t \le 4V_T$	$V_{GS}-V_t \ge 4V_T$
$V_{DS} \le V_{DSAT}$	Triode – Weak Inversion	Triode – Moderate Inversion	Triode – Strong Inversion
$V_{DS} \ge V_{DSAT}$	Saturation – Weak Inversion	Saturation – Moderate Inversion	Saturation – Strong Inversion

Independently of the condition of strong or weak inversion, saturation region is characterized by reduced dependence of the I_{DS} on the V_{DS}. On the contrary, in triode region, the drain current shows a strong dependence on V_{DS}. For very small V_{DS} values (V_{DS}<<V_{DSAT}) the I_{DS} vs V_{DS} dependence is linear.

Strong inversion is a region where the inversion layer is well formed and the depletion layer in the MOSFET body is not affected by V_{GS}. In strong inversion, the dependence of I_{DS} on V_{GS} and V_{DS} can be approximated by square laws (MOSFET parabolic equations). Weak inversion, also indicated with "subthreshold region", is marked by exponential dependence of I_{DS} vs V_{GS}. Moderate inversion is the transition region between weak and strong inversion.

For V_{GS} << V_t , I_{DS} becomes so small that the MOSFET can be considered turned off (off state). This occurs when the I_{DS} current becomes of the same order of the junction leakage currents.

Drain current equations in strong inversion

In strong inversion, the usual equations used in triode and saturation regions are:

$$I_{DS} = \beta_n \left(V_{GS} - V_t - \frac{V_{DS}}{2} \right) V_{DS} \qquad \text{(Triode)} \tag{1.6}$$

$$I_{DS} = \beta_n \frac{\left(V_{GS} - V_t\right)^2}{2} \left[1 + \lambda \left(V_{DS} - V_{DSAT} \right) \right] \qquad \text{(Saturation)} \tag{1.7}$$

where:

$$\beta_n = \mu_n C_{OX} \frac{W_{eff}}{L_{eff}} . \quad W_{eff}=W-2W_D, \quad L_{eff}=L-2L_D , \tag{1.8}$$

and μ_n is the electron mobility in the channel. Parameters W_D and L_D are the reduction that the channel width and length, respectively, suffer from each side of the gate with respect to the corresponding drawn geometries *W* and *L*.

Lambda (λ) is an important parameter that determine the dependence of I_{DS} on V_{DS} in saturation region. This dependence is due to multiple phenomenon, such as channel length modulation or drain-induced barrier lowering (DIBL). The ideal situation would be λ=0, i.e. no dependence on V_{DS}. In practice, it is difficult to obtain $\lambda < 0.01\ V^{-1}$. The most important parameter that affects λ is the channel length. As a first order approximation it possible to express the inverse of lambda on channel length with a linear relationship, valid for lengths somewhat larger than the minimum L of the process:

$$\lambda^{-1} = k_\lambda L_{eff} \tag{1.9}$$

where k_λ (V/μm) is a constant.

Drain current in weak inversion

Weak inversion is also indicated as sub-threshold region. Strictly speaking, sub-threshold region would require that V_{GS}-V_t<*0.* In practice, the terms weak inversion and sub-threshold are both used to indicate a region where the I_{DS} dependence on both V_{GS} and V_{DS} becomes exponential, according to the equation:

$$I_{DS} = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{\frac{-V_{DS}}{V_T}} \right) \left[1 + \lambda \left(V_{DS} - V_{DSAT} \right) \right] \tag{1.10}$$

where I_{SM} is given by:

$$I_{SM} = \mu_n C_{dm} \frac{W_{eff}}{L_{eff}} V_T^2 = \mu_n C_{ox} (m-1) V_T^2 \frac{W_{eff}}{L_{eff}} \tag{1.11}$$

When V_{DS} is high enough, the exponential term exp(-V_{DS}/V_T) can be neglected with respect to one and the drain current shows a reduced dependence on V_{DS} (limited to the λV_{DS} term as in strong inversion. In this condition, the MOSFET is in saturation and (1.10) reduces to:

$$I_{DS} = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left[1+\lambda\left(V_{DS}-V_{DSAT}\right)\right] \tag{1.12}$$

For low V_{DS} value, the exponential term in the round parentheses is no more negligible and the current begins to show a strong dependence on V_{DS}. This is the analogue of the triode region defined for the strong inversion.

Moderate inversion

In moderate inversion, I_{DS} dependence progressively changes from parabolic to exponential. Simple equations for this region do not exist. However, the EKV (Enz-Krummenacher-Vittoz) model [3] consists of a single I_{DS} equation for all three regions (weak-moderate-strong inversion). This equation is rather complex for hand calculation and require the voltage to be expressed with the body-referred method.

Saturation voltage, V_{DSAT}

The saturation voltage V_{DSAT} can be approximated by the following expressions:

$$V_{DSAT} \cong \begin{cases} \left(V_{GS}-V_t\right) & \text{in strong inversion} \\ 4V_T \ \ (100 \ \ \text{mV}) & \text{in moderate and weak inversion} \end{cases} \tag{1.13}$$

These, are empirical expressions that set a lower limit to V_{DS} for having a small dependence of I_{DS} on V_{DS}. Definitions that are more related to physical phenomena can be also used. For example, textbooks on electron devices [4] often propose the expression (V_{GS}-V_t)/m for V_{DSAT} in strong inversion. Since m is slightly greater than one, using this definition lead to a lower V_{DSAT} value. On the other hand, the definitions given in (1.13) are simpler to use for design purposes and give a good representation of the experimental MOSFET behavior.

Junction currents

The source-body and drain-body junctions are normally reverse biased. As a result, these junctions, represented by the D_{BS} abd D_{BD} diodes in Fig.1.4, carry only the inverse saturation current, which is given by:

$$I_J = A_J J_S \tag{1.14}$$

where I_J is the current flowing through the junction, A_J is the area of the junction (source or drain area) and J_S the inverse saturation current density. Considering the layout of Fig.1.4, A_D and A_S (drain a source areas, respectively) are both given by $W \cdot L_C$. Typical values of J_S at room temperature are of the order of 0.1 fA/μm^2.

Temperature effects

The effect of temperature on the MOSFET dc characteristics can be represented by the temperature dependence of β_n and V_t. Temperature affect b through the mobility, which generally decreases with temperature. The following equation can be used to represent the temperature dependence of β_n:

$$\beta_n(T) = \beta_n(T_0)\left(\frac{T}{T_0}\right)^{-\alpha_\mu} \tag{1.15}$$

where T_0 is a reference temperature (for example 300 K) and α_μ a process-dependent constant that can vary in the 1.2-2 interval.

The temperature dependence of the threshold voltage is often approximated with a linear expression:

$$V_t(T) = V_t(T_0) - \alpha_{VT}\left(T - T_0\right) \tag{1.16}$$

where α_{VT} is of the order of 1 mV/K. As temperature increases, both V_t and β_n decreases producing opposite effects on I_{DS}: the V_t reduction increases I_{DS} while the β_n reduction causes a proportional decrease of I_{DS}. These effects generally do not compensate each other:

- At low V_{GS}-V_t, the threshold voltage dominates and the current increases with temperature.
- At high V_{GS}-V_t, it is β_n to dominate and the current increases with temperature.

Compensation of the two effects occurs only for a particular value of V_{GS}-V_t., which typically falls well into the strong inversion range. The presence of an operating point with low temperature sensitivity (ZTC, zero temperature coefficient) can be used for reference voltage generation [5].

The behavior of p-channel devices is similar to n-channel ones. Expression (1.15) can be used also for $\beta_p=\mu_p C_{ox} W_{eff}/L_{eff}$, while (1.16) is applicable when the absolute value of p-channel threshold voltage is considered. Then, for p-channel devices, the absolute value of V_t decreases with temperature, increasing the magnitude of I_{DS}, just as for n-channel devices [5].

As far as leakage is concerned, it should be observed that the junction saturation currents of all junctions (e.g. body-drain, body-source junctions) roughly doubles for every temperature increment of 10 °C. This means that these currents are multiplied by about a factor of 1000 for a temperature increment of 100 °C.

Small signal model and parameters

The small signal model of the MOSFET is shown in Fig. 1.5. Notice that, for simplicity, the series resistances R_S and R_D have been neglected in the small signal circuit. The reduced circuit for dc signals is obtained from the circuit of Fig. 1.5 by removing all parasitic capacitors. The result is shown in Fig.1.6.

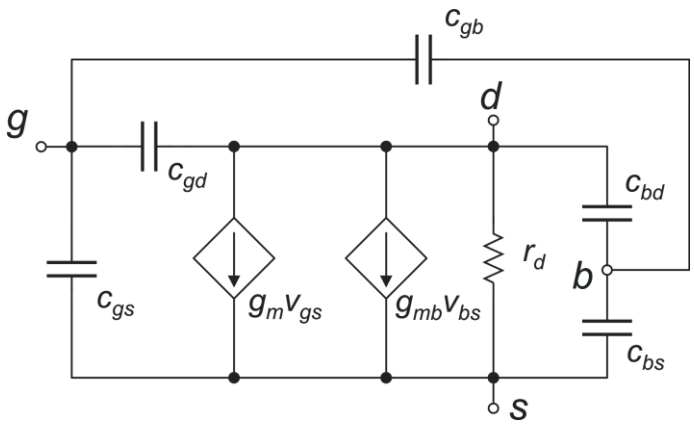

Fig.1.5. MOSFET small signal model.

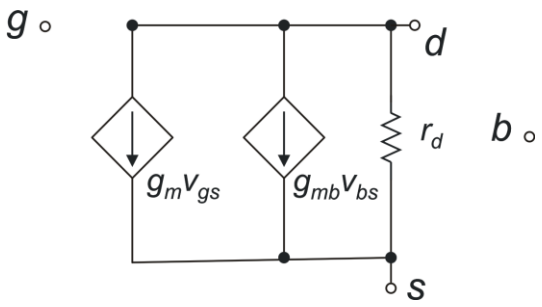

Fig.1.6. MOSFET small signal model for dc signals.

Let us start from the dc equivalent circuit. It is important to study how the parameters vary with the operating point. We will focus on g_m and r_d, since g_{mb} is related to g_m by the following relationship:

$$g_{mb} = (m-1) g_m \cong 0.2 g_m \tag{1.17}$$

which is a direct consequence of (1.4).

It is interesting to consider the following two aspects:

- How g_m and r_d varies as a function of V_{DS} for a fixed overdrive voltage (V_{GS}-V_t).
- How g_m varies in saturation region as a function of V_{GS}-V_t.

The first point will be analyzed considering (V_{GS}-V_t)>4V_t, i.e. strong inversion. Then, considering (1.6), we can calculate the values of g_m and g_{ds}=1/r_d in triode region:

$$g_m \equiv \left(\frac{\partial I_{DS}}{\partial V_{GS}} \right)_{V_{DS}, V_{BS}=const} = \beta_n V_{DS} \tag{1.18}$$

$$\frac{1}{r_d} = g_{ds} \equiv \left(\frac{\partial I_{DS}}{\partial V_{DS}}\right)_{V_{GS},V_{BS}=const} = \beta_n\left(V_{GS} - V_t - V_{DS}\right) \tag{1.19}$$

In saturation region, the same parameters can be find using (1.7), and neglecting the dependence of V_{DSAT} on V_{GS}. :

$$g_m \equiv \left(\frac{\partial I_{DS}}{\partial V_{GS}}\right)_{V_{DS},V_{BS}=const} = \beta_n\left(V_{GS} - V_t\right)\left[1 + \lambda\left(V_{DS} - V_{DSAT}\right)\right] \cong \beta_n\left(V_{GS} - V_t\right) \tag{1.20}$$

$$\frac{1}{r_d} = g_{ds} \equiv \left(\frac{\partial I_{DS}}{\partial V_{DS}}\right)_{V_{GS},V_{BS}=const} = \lambda\frac{\beta_n}{2}\left(V_{GS} - V_t\right)^2 \cong \lambda I_{DS} \tag{1.21}$$

For $V_{DS}=V_{DSAT}=V_{GS}-V_t$, triode expression (1.19) yields $g_{ds}=0$, while, for the same V_{DS}, the saturation formula (1.21) gives $g_{ds}=\lambda I_D$. The reason is the discontinuity (of both I_{DS} and its first derivative) obtained for $V_{DS}=V_{DSAT}$ using (1.6) and (1.7). Assuming that the correct value of g_{ds} for $V_{DS}=V_{DSAT}$ is given by the saturation equations, the simplified behavior of gm and g_{ds} across both triode and saturation region are represented in Fig.1.7.

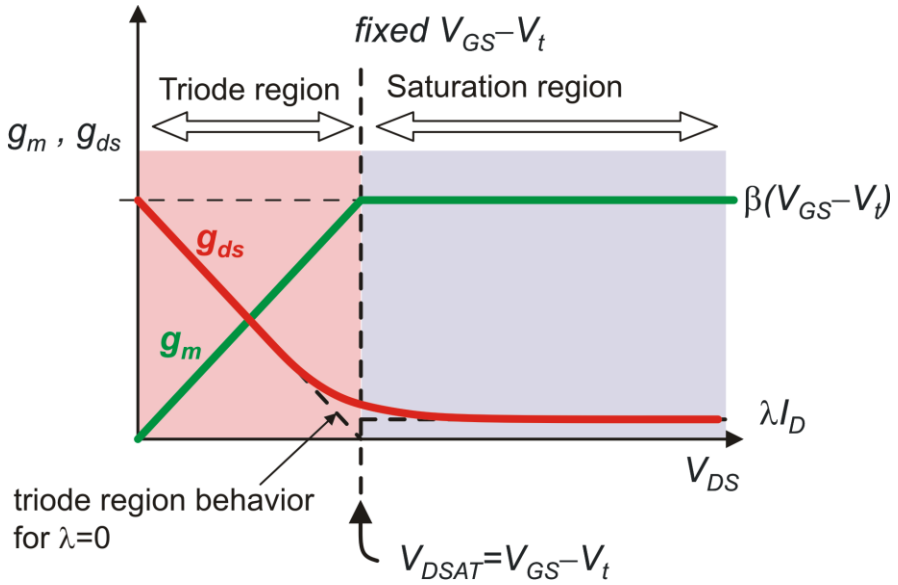

Fig. 1.7. Simplified plots of g_m and g_{ds} as a function of V_{DS} for constant V_{GS}-V_t, in strong inversion.

A similar behavior can be derived for V_{GS}-V_t values that set the device in weak or moderate inversion, but with different values for g_m, g_{ds} and V_{DSAT}.

In all cases, it is important to observe that in saturation both g_m and $r_d=1/g_{ds}$ assume their maximum value in saturation. When V_{DS} gets lower than V_{DSAT} and the device gets into triode region, both g_m and r_d start

decreasing progressively, down to their minimum values that are reached for V_{DS}=0. In particular, for V_{DS}=0 g_m is zero.

Transconductance models in saturation region

The second point can be analyzed considering three different equivalent expressions for g_m in strong inversion:

$$g_m = \beta_n (V_{GS} - V_t) \tag{1.22}$$

$$g_m = \sqrt{2\beta_n I_D} \tag{1.23}$$

$$g_m = \frac{2I_D}{(V_{GS} - V_t)} \tag{1.24}$$

The first one simply coincides with (1.20), The second one can be obtained from (1.22) considering that, neglecting the λV_{DS} term in (1.7), then the overdrive voltage in strong inversion is given by:

$$V_{GS} - V_t = \sqrt{\frac{2I_D}{\beta_n}} \tag{1.25}$$

The third expression, (1.24), can be obtained considering that, neglecting λV_{DS} in (1.7), $V_{GS}-V_t$ is equal to $2I_D/\beta_n(V_{GS}-V_t)$.

In weak inversion, using (1.12), it is possible to find the following expression for g_m:

$$g_m = \frac{I_D}{mV_T} \tag{1.26}$$

Considering that for a BJT $g_m=I_C/V_T$, it is possible to use the following g_m expression for a MOSFET in strong, moderate and weak inversion and extend it to the BJT:

$$g_m = \frac{I_D}{V_{TE}} \tag{1.27}$$

where V_{TE} (equivalent V_T) is a voltage, defined just by equation (1.27), which assumes the following value:

$$V_{TE} = \begin{cases} (V_{GS} - V_t)/2 & \text{MOSFET in strong inversion} \\ mV_T & \text{MOSFET in weak inversion} \\ V_T & \text{BJT} \end{cases} \tag{1.28}$$

The typical behavior of V_{TE} is represented in Fig.1.8

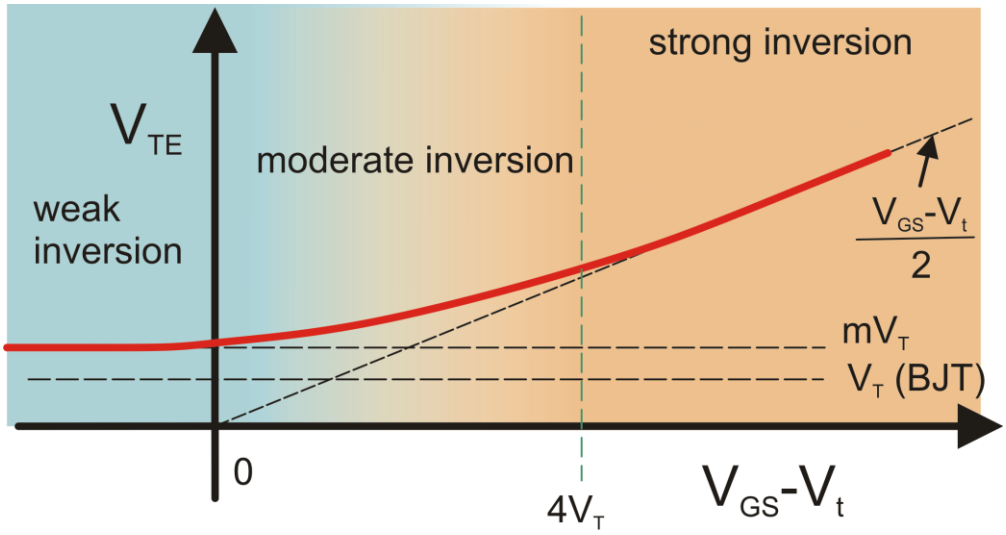

Fig. 1.8. Behavior of the V_{TE} parameter as a function of V_{GS}-V_t, compared with the BJT value.

A parameter that is sometimes quoted in scientific articles on analog design [6] is g_m/I_D, equal to $1/V_{TE}$. This parameter is sometimes referred to as "transconductance efficiency", since a high g_m/I_D value means that it is possible to obtain large transconductances with relatively small I_D. Notice that it is often necessary to obtain a large g_m from selected devices in a circuit, in order to achieve low noise and/or wide bandwidth performances. With high values of g_m/I_D it is possible to meet noise and bandwidth specifications with lower current consumption. Considering that V_{TE} is the inverse of g_m/I_D, Fig.1.8 means that, for a MOSFET, the best efficiency is found at the lowest overdrive voltages, ideally in weak inversion. Due to the factor m>1, a BJT achieves a V_{TE} lower than a MOSFET in weak inversion. Then BJT are superior to MOSFETs in this respect.

Capacitance models

Parasitic capacitances shown in the circuit of Fig. 1.5 are the small signal equivalent of non-linear capacitors shown in Fig. 1.4.

Each one of the capacitances that involve the gate, (c_{gs}, c_{gd}, c_{gb}) can be divided into two components, named "overlap" and "intrinsic" capacitances, according to:

$$\begin{aligned} c_{gs} &= c_{gs}^{(ov)} + c_{gs}^{(i)} \\ c_{gd} &= c_{gd}^{(ov)} + c_{gd}^{(i)} \\ c_{gb} &= c_{gb}^{(ov)} + c_{gb}^{(i)} \end{aligned} \tag{1.29}$$

where the apex (ov) stands for "overlap" and (i) for "intrinsic". Overlap capacitances derive from the gate partial superimposition on the other electrodes. For example, examining Fig.1.1, it is possible to recognize the gate-source and gate-drain overlap areas. Gate-body overlap is due to the boundary regions, where the gate layer (polysilicon in Fig.1.1) progressively gets away from the substrate. In these regions, the gate is still close enough to the substrate to create a significant capacitance, but not so close to be able to invert the substrate surface and create the channel. Overlap capacitances are practically independent of the state of the channel (i.e. depletion, inversion, etc.) and do not depend on the MOSFET terminal voltage. Their value is given by:

$$\text{overlap capacitances: } \begin{cases} c_{gs}^{(ov)} = c_{gso} \cdot W \\ c_{gd}^{(ov)} = c_{gdo} \cdot W \\ c_{gb}^{(ov)} = c_{gbo} \cdot L \end{cases} \tag{1.30}$$

where c_{gso}, c_{gdo}, c_{gbo} are constant capacitance-per-unit length coefficients (unity: F/m). Inspection of Fig.1.1 shows that the gate-source and gate-drain overlap capacitances are located along the gate width , whereas the gate-body capacitance is distributed along the gate length. For this reason the overlap components of c_{gs}, c_{gd}, are proportional to W, while the gate-body one is proportional to L.

The intrinsic capacitances are strictly related to the charge accumulation in the channel. These capacitances are strongly non-linear. Simplified expressions, which are often used for hand calculations consist in the so-called Meyer model. According to this model the first order approximation of the three intrinsic capacitances in off-region ($V_{GS}<<V_t$), triode and saturation regions are given in table 1.2.

Table 1.2. Meyer Model for the mosfet intrinsic capacitances

	Off –state	Triode	Saturation
$c_{gs}^{(i)}$	0	$\frac{1}{2}C_{ox}WL$	$\frac{2}{3}C_{ox}WL$
$c_{gd}^{(i)}$	0	$\frac{1}{2}C_{ox}WL$	0
$c_{gb}^{(i)}$	$\frac{1}{C_{ox}WL}+\frac{1}{C_{dm}}$	0	0

A major drawback of the Meyer model is that it does not respect charge conservation. Such a model fails to represent transient phenomena where a MOSFET crosses different operating region, as occurs with logical gates or switches (pass – transistors). A more accurate description requires that the intrinsic capacitances be replaced by "capacitance coefficients" generically defined by:

$$c_{ij} \equiv \frac{\partial Q_i}{\partial V_j} \tag{1.31}$$

where Q_i is the charge accumulated on terminal "i" and V_j is the voltage of terminal "j". The terminal taken into consideration are drain, gate and source. The body is used as a reference for the voltage of the other electrodes. Then we have three self-capacitance coefficients (c_{dd}, c_{gg}, c_{ss}) which are all positive, and six mutual capacitance coefficients (c_{gd}, c_{dg}, c_{gs}, c_{sg}, c_{sd}, c_{ds}), which are negative, since the charge displaced on an electrode by a positive voltage change applied to a different electrode is negative (as happens even in an ideal parallel plate capacitor). Capacitances c_{sd} and c_{ds} are generally negligible. Finally, it should be observed that, generally, c_{ij} is different from c_{ji}, due to the nonlinear behavior of the MOSFET. As a result, the mutual charge induction between the gate and the other two electrode cannot be represented by a simple capacitance, but two distinct coefficients are required. This model is called charge-oriented model and was introduced by Dutton and Ward [7]. All modern simulation model adopt the charge-oriented model for the intrinsic capacitances.

Finally, c_{bd} and c_{ds} are junction capacitances, marked by a strong dependence on voltage. Normally, the drain-body and source-body junctions are reverse biased. The larger the (reverse) bias, the smaller the capacitance. Commonly used expressions for these capacitances are the following:

$$c_{bs} = \frac{C_J A_S}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_j}} + \frac{C_{JSW} P_S}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_{jsw}}} \tag{1.32}$$

$$c_{bd} = \frac{C_J A_D}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_j}} + \frac{C_{JSW} P_D}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_{jsw}}} \tag{1.33}$$

where: A_S=source area, P_S= source perimeter, A_D=drain area, P_D= drain perimeter. Parameters C_J and C_{JSW} are the (zero-bias) capacitances per unit area and unit perimeter, respectively. V_0 is the built-in potential of the junction whereas m_j and m_{jsw} exponents, called "grading coefficients", depend on the doping profiles across the bottom and sidewalls of the source/drain implants, respectively. These exponents are generally in the range 0.33-0.5.

The overlap capacitance and the junction capacitances are classified as extrinsic capacitances.

Matching parameters

The most frequently used expressions for the standard deviation of the threshold voltage and transconductance factor (β) matching errors of two matched MOSFETs, are the following:

$$\sigma_{\frac{\Delta\beta}{\beta}} = \frac{C_\beta}{\sqrt{WL}}; \qquad \sigma_{Vt} = \frac{C_{Vt}}{\sqrt{WL}} \tag{1.34}$$

1.3 Bipolar Transistor Layouts.

Layout descriptions

Typical BJTs that are available in a standard bipolar or BiCMOS process are the vertical NPN and the lateral PNP transistors. Both devices are created inside a moderately n-doped region, which is isolated from the surrounding p-doped substrate. These isolated n-type areas are generally called “pockets”. In earlier pure bipolar process, the pockets were obtained from an n-doped epitaxial layer, grown onto a p-substrate. The epi-layer was divided into isolated pockets (epi-pockets) by means of isolation p+ implants, or trenches filled by silicon dioxide. Modern BiCMOS process are generally derived by simpler CMOS process. Therefore, the n-pockets are simply n-wells diffused through a p^- epitaxial layer. In order to reduce the resistivity of the pocket, without altering the superficial moderately doped region, a highly n-doped buried layer (or buried well) is created on the bottom of pocket. A possible cross-section and layout of a vertical NPN BJT built inside an n-well are shown in Fig. 1.9. The n-pocket coincides with the collector. Creation of the base region requires a special doping (p-base), which is normally not available in a standard CMOS process. The emitter can be obtained with an n+ implant, or by simply putting an n-doped polysilicon strip in contact with the p-base diffusion [2].

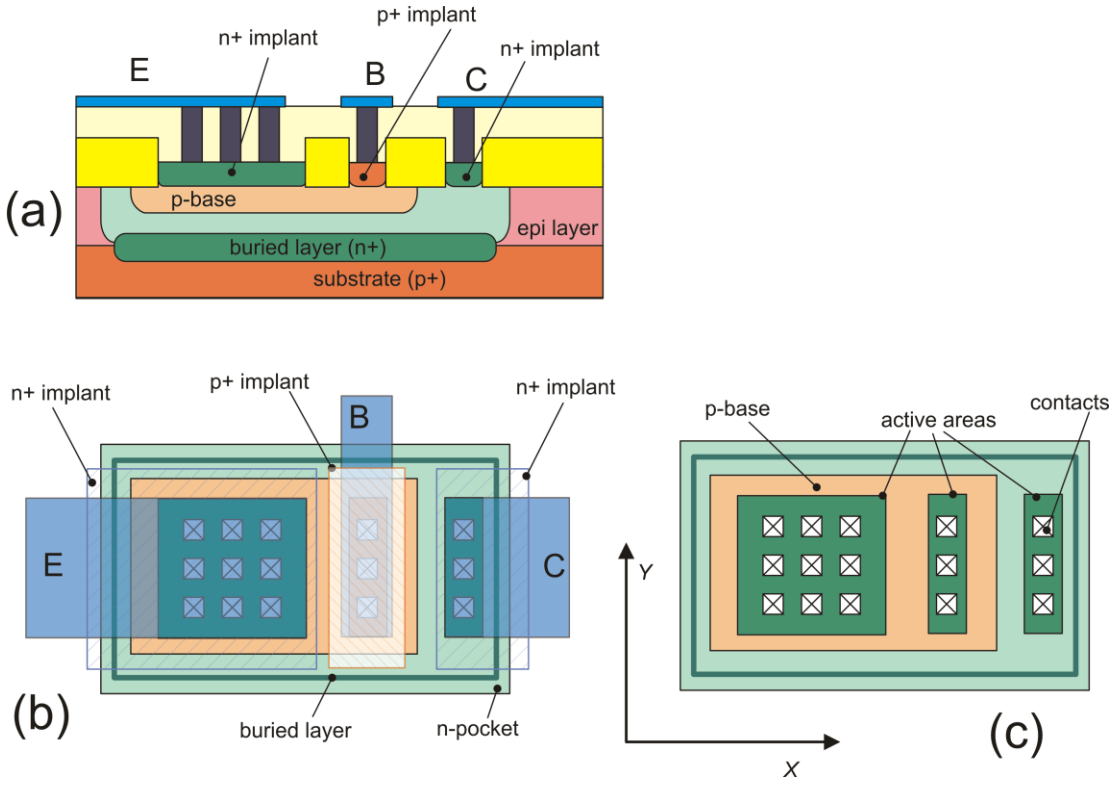

Fig. 1.9. Cross section (a) and Layout (b) of a vertical npn BJT. A layout excluding n-plus, p-plus and metal layers is shown in (c).

The cross-section and layout of a lateral PNP BJT are shown in Fig. 1.10 (a). The pocket is the base, while the emitter and collector are p-diffusions (p-base in the example of Fig.1.10). Transport of holes from the emitter to the collector occurs laterally, under control of the thin n-well (n-pocket) layer that separates them and acts as the base. In early Bipolar processes, lateral PNP BJTs had a very low beta,

due to the difficulty of reducing the base length. With modern photolithography resolutions, this no more a limiting factor and PNP transistors with beta of the order of one hundred or more can be easily obtained. One of the major limitations of lateral PNP transistors is the possibility to control the doping profiles of the emitter and collector independently.

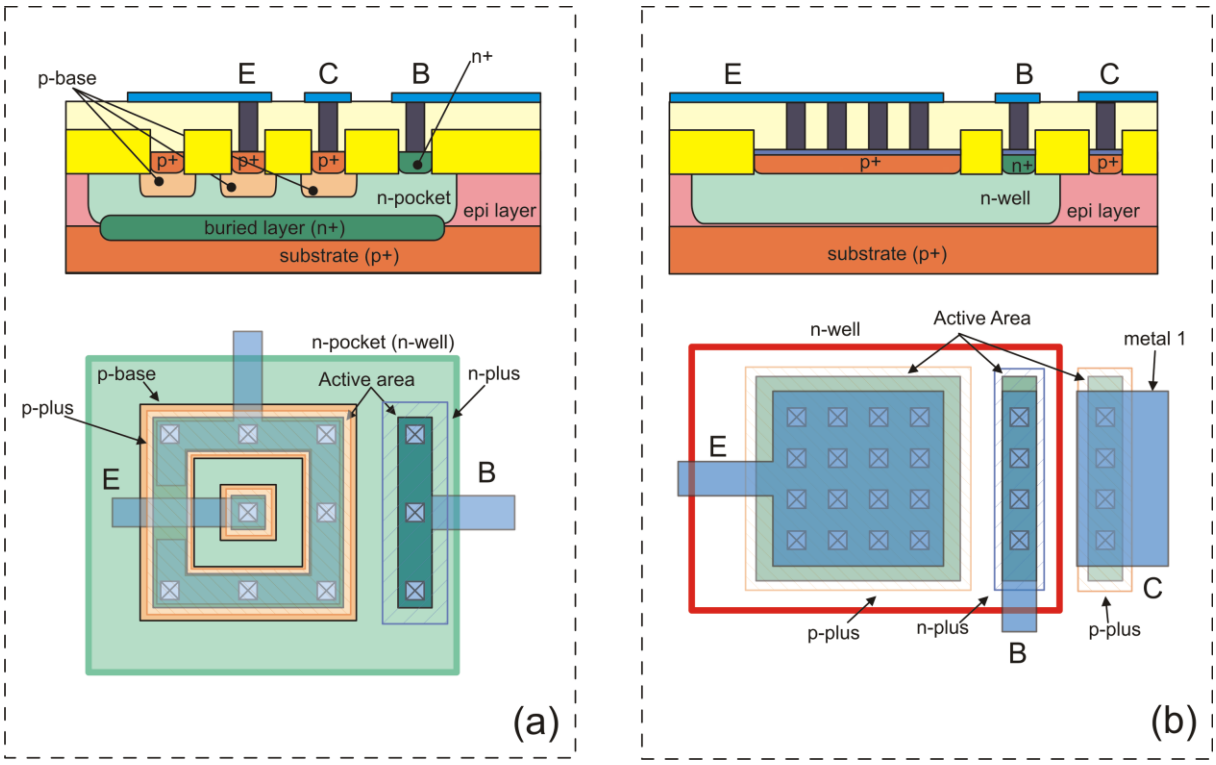

Fig.1.10. Cross-section and layout of a lateral BJT (a) and of a substrate PNP (b).

The high resistivity of the base (n-pocket) introduce a parasitic base resistance that adversely affect noise and high frequency performances. In spite of these limitations, lateral PNP transistors are widely used in general-purpose analog circuits, such as operational and instrumentation amplifiers. Bipolar analog circuit requiring NPN and PNP transistors with similar characteristics should be based on both NPN and PNP vertical transistors (Complementary Bipolar Processes).

A vertical PNP BJT, which is available even in standard n-Well CMOS technology with no need of additional masks or process steps, is shown in Fig.1.10 (b). The base is formed by an n-Well, while the emitter is a p+ region (doped active area) created into the n-Well. The emitter doping is introduced with the same step used to create the drain/source of p-MOSFETs. A major limitation of this kind of device is that the collector the substrate, thus it should be mandatorily fixed to the smallest supply voltage. Substrate BJTs in CMOS processes are typically used to build reference voltage circuit based on the Band-Gap principle.

Designer options for BJTs

Process design kits (PDK) generally give the designer less freedom to customize bipolar transistors with respect to MOSFEts. We have seen that the designer may assign both the width and length of a MOSFET within a very wide interval of values. The PDK offers a set of BJTs both npn and pnp, that differ for the

type of layout. These devices are called elementary transistors. The designer can place an elementary device directly into the circuit or personalize it through a dimensionless parameter called *area*. The area acts as a multiplier for many parameters of the elementary device. The saturation current (I_S) and parasitic capacitances are multiplied by the *area* factor, while the series resistances are divided by *area*. These transformations of parameters are taken into account by the electrical simulator. The parameter area must be greater or equal to one, meaning that we can modify an elementary transistor only making it bigger, not smaller. As far as the layout is concerned, the parameter *area* can be implemented in different ways. The simplest method, which is always available, is placing a number N of elementary devices in parallel. In this way $area=N$. The main limitation is that area must be an integer. A different approach that allows also factionary area values (but still $area \geq 1$) is stretching the layout of the elementary transistor along a selected direction. For example, for the vertical transistor of Fig.1.9, it is possible to stretch the layout by a factor equal to *area* along the Y direction. This increases the effective emitter area, proportionally increasing I_S . Layout stretching is not possible for the lateral PNP of Fig.1.10 (a), thus the only option for this type of device is paralleling elementary transistors (integer areas only).

1.4 Bipolar transistor models

Large signal model

Large signal dc analysis of bipolar transistors is performed using the Ebers Moll equivalent circuit, properly modified to take into account the Early effect. As far as large-signal transient analysis is concerned, the charge-control model is generally used [3]. Here, we will limit to recall the simplified collector current equation in the forward-active region, which is used by SPICE-like simulators [8]:

$$I_C = I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB}}{V_A} \right) \tag{1.35}$$

where I_S is the saturation current, $V_T=k_BT/q$ and V_A is the Early voltage.

This equation is often simplified by considering that $V_{CB}=V_{CE}-V_{BE}$ and that V_{BE} is nearly constant in the forward-active region. By this simplification, this leads the following equation, which is generally referred to in most textbook on electronic design:

$$I_C \cong I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CE}}{V_A} \right) \tag{1.36}$$

The base current can be estimated by means of the well-known relationship:

$$I_B \cong \frac{I_C}{\beta} \tag{1.37}$$

where β is the dc current gain.

The active region is characterized by $V_{CE} \geq V_{CESAT}$. Textbooks on electron devices define saturation as a condition where both the BE and BC junctions are forward-biased. From this definition, V_{CESAT} would

be such that V_{BC}=0, then V_{CESAT}= V_{BE}. For practical purposes, saturation is considered the region where I_C starts to exhibit a strong dependence on V_{CE} and the base current gets much larger than the value given by (1.37). Typical values of V_{CESAT} that correspond to the practical definition are in the range 100-200 mV.

Equations (1.36) and (1.37) represent a good approximation of I_C and I_B over a large interval of collector currents. In particular, the exponential behavior represented by (1.36) is generally maintained across an I_C range of several decades. Outside this range, the collector current deviates from the exponential behavior. This is well represented by the so-called Gummel plot, which is generally included in the process manuals for any elementary device. An example of Gummel plot is shown in Fig. 1.11. The plot gives a representation of the collector and base currents as a function of the base-emitter voltage, for a fixed collector-emitter voltage. The latter is chosen to keep the device in forward-active region. Considering the semi-logarithmic scale (linear voltage, logarithmic currents), an exponential behavior turns out into a straight line.

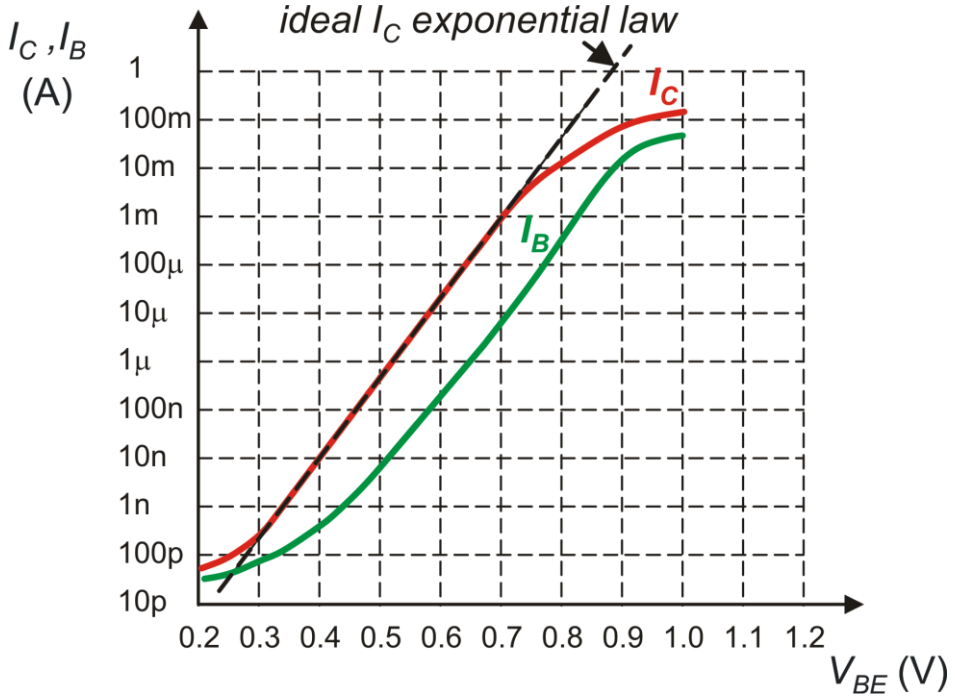

Fig.1.11. Example of Gummel plot.

Equation (1.36) holds true region where the I_C curve is well approximated by a straight line. Deviation from the straight-line at low currents is mainly due to leakage currents (e.g. currents from the reverse-biased CB junction). At high currents, the discrepancy is due to high-injection effects and to the base series resistance. Due to the logarithmic scale, the distance between the I_C and I_B curve is proportional to log(β). At low and high collector currents, the I_C and I_B curves get closer, meaning a reduction of β. This is well represented by the curve of β (see Fig.1.12), which is also generally included into the process manual.

The fall of beta at low currents is mainly due to generation-recombination and tunneling current associated to the base-emitter junction. The beta decrease at high currents is due to base widening effects occurring at high current densities (e.g. Kirk effect) [2].

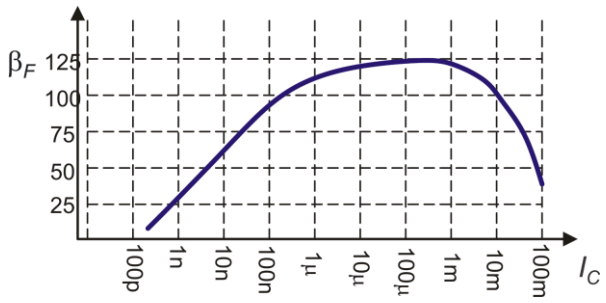

Fig.1.12 . Beta as a function of I_C.

For a transistor having *area* > 1, the same Gummel plot (or same beta plot) of the corresponding elementary device can be used by simply dividing the actual device current by the *area* factor. As an example, let us consider an elementary BJT having the Gummel plot of Fig. 1.11. If we bias the elementary device (*area*=1) with a current of 10 mA, then we are out of the region where the exponential equations hold. This may have several adverse consequences. For example, the V_{BE} cannot be considered as a quasi-constant voltage (Vγ) anymore since its increases with I_C gets larger than in the exponential region. Furthermore, there are circuits that rely on the exponential behavior, such as band-gap voltage references or analog multipliers. If we cannot reduce the bias current (e.g., due to noise or bandwidth constraints), we can still set the area parameter to a value greater than one. Using *area*=10, the equivalent I_C current to be used in the Gummel plot of the elementary device is the actual I_C current (10 mA) by 10. The resulting value (1 mA) is well within the exponential region. Obviously, the maximum collector current of a device is *area* times as large as the maximum current of the elementary transistor.

BJT small signal model

A simplified small-signal equivalent circuit of a vertical BJT is shown in Fig.1.13. This circuit is similar to the MOSFET equivalent circuit (with the exception of the input resistance r_{be}), allowing straightforward porting of circuit analysis results across the two device types. For this reasons the circuit of Fig.1.13 is more frequently used than h-parameter one for integrated circuit design.

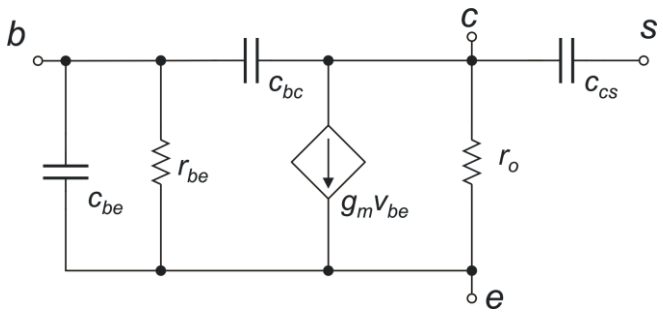

Fig.1.13. BJT small signal equivalent circuit

Terminal "s" represents the substrate. In a vertical npn BJT (see Fig.1.9), the substrate (p-doped) forms a junction with the collector (n-doped). This junction is kept reverse-biased by setting the substrate to the smaller supply voltage. Nevertheless, due to the junction, a capacitance (c_{cs} in Fig.1.13) is present between the collector and the substrate. Due to the large extension of the collector body, this capacitance can be relatively large, affecting the performances of circuit that use npn BJTs. In the case of a lateral

structure such as the pnp BJT of Fig. 1.10 (a), it is the base to be coupled to the substrate with a junction capacitance.

The dc parameters, namely g_m, r_0 and r_{be} can be derived from (1.36) and (1.37), obtaining the following expressions:

$$g_m \equiv \left(\frac{\partial I_C}{\partial V_{BE}} \right)_{V_{CE}=const} = \frac{I_C}{V_T} \tag{1.38}$$

$$r_0 \equiv \frac{1}{\left(\frac{\partial I_C}{\partial V_{CE}} \right)_{V_{BE}=const}} = \frac{1}{\frac{1}{V_A} \cdot I_S e^{\frac{V_{BE}}{V_T}}} \cong \frac{V_A}{I_C} \tag{1.39}$$

$$r_{be} \equiv \left(\frac{\partial V_{BE}}{\partial I_B} \right)_{V_{CE}=const} = \left(\frac{\partial I_B}{\partial V_{BE}} \right)^{-1}_{V_{CE}=const} = \left[\frac{1}{\beta} \left(\frac{\partial I_C}{\partial V_{BE}} \right)_{V_{CE}=const} \right]^{-1} = \beta \frac{1}{g_m} \tag{1.40}$$

Notice that expression (1.40) is an approximation, which is acceptable only if β can be considered independent of I_C. The dc version of circuit in Fig. 1.13 is equivalent to the well-known hybrid parameter circuit with $h_{re}=0$. Equivalence with the remaining h-parameters are: $h_{ie}=r_{be}$, $h_{oe}=1/r_o$, $h_{fe}=g_m r_{be}$.

The capacitances that appear in the equivalent circuit are all due to junctions, and then are strongly voltage dependent. In active-forward region, the base-collector and collector-substrate junctions are generally reverse biased. The only exception occurs for $V_{CESAT} \leq V_{CE} \leq V_{BE}$, where the base-collector is be weakly forward biased. In all cases, c_{bc} and c_{cs} are dominated by the depletion-layer capacitance, given by:

$$c_{bc} = \frac{C_{JC}}{\left(1 + \frac{V_{CB}}{V_{JC}} \right)^{m_{jc}}}, \qquad c_{cs} = \frac{C_{JS}}{\left(1 + \frac{V_{CS}}{V_{JS}} \right)^{m_{js}}} \tag{1.41}$$

where C_{JC} and C_{JS} are the corresponding zero-bias capacitances, V_{JC} and V_{JS} are the built-in potentials of the two junctions and m_{jc}, m_{js} the corresponding grading coefficients.

On the other hand, the base-emitter junction is forward biased in the forward-active region. Therefore, both the depletion-layer (c_{te}) and diffusion (c_{de}) capacitances must be considered. Then, $c_{be}=c_{te}+c_{de}$, with:

$$c_{te} = \frac{C_{JE}}{\left(1 - \frac{V_{BE}}{V_{JE}} \right)^{m_{je}}} \tag{1.42}$$

$$c_{de} = \tau_F g_m \tag{1.43}$$

where τ_F is the forward transit time [2]. At sufficently high I_C values, c_{de} is much larger than c_{te} and c_{bc}. In these conditions, the forward transit time can be related to the transition frequency (f_T) of the bipolar transistor through the following equation:

$$f_T = \frac{1}{2\pi\tau_F} \tag{1.44}$$

1.5 Considerations about small-signal resistances seen across device terminals or from one terminal to ground

Classification of resistances according to their magnitude

Depending on the configuration, the small-signal resistances seen across terminal pairs or netween a terminal and ground may vary by several orders of magnitude. Clearly, resistances depend also on the device bias, thus comparison should be made only in condition of equal bias. For a given bias point, resistances can be classified in small, medium and large, and very large, according to the following table:

Table 1.3. Classification of small-signal resistances according to magnitude

	Small	Medium-large	Large	Very large
MOSFETs	$1/g_m$	-	r_d	$(g_m r_d) r_d$
BJTs	$1/g_m$	h_{ie} (r_{be})	r_0	$h_{fe} r_o$

Notice that, for the MOSFET, from one category to the successive the resistance increases by a factor $g_m r_d$, which roughly corresponds to nearly two orders of magnitude. For the BJT it is necessary to use one more resistance class (medium), in order to arrange h_{ie} which, on average, is only 5-10 times smaller than r_o.

Notable cases of small-signal resistances

The knowledge of approximate expressions for small-signal resistances can be very useful to understand the operating principle of a large number of circuits. In many cases, it is sufficient to classify resistances on the basis of the categories represented in table 1.3. Such a knowledge may allow calculating the order of magnitude of an amplifier gain or deciding, which path a small-signal current follows.

The cases that are treated in this document are shown in Fig.1.14. The first three cases, namely r_{V1}, r_{V2} and r_{V3} concern MOSFET circuits. These resistances have been obtained using the small signal circuit of Fig. 1.6, considering g_{mb}=0.

Case r_{v1}: Resistance seen from source to ground.

$$r_{v1} = \frac{R_D + r_d}{1 + g_m r_d} \quad \text{(exact result)} \tag{1.45}$$

Approximate result:

$$\text{for } g_m r_d >> 1 \quad \text{and} \quad R_D << r_d \;\Rightarrow\; r_{v1} \cong \frac{1}{g_m} \tag{1.46}$$

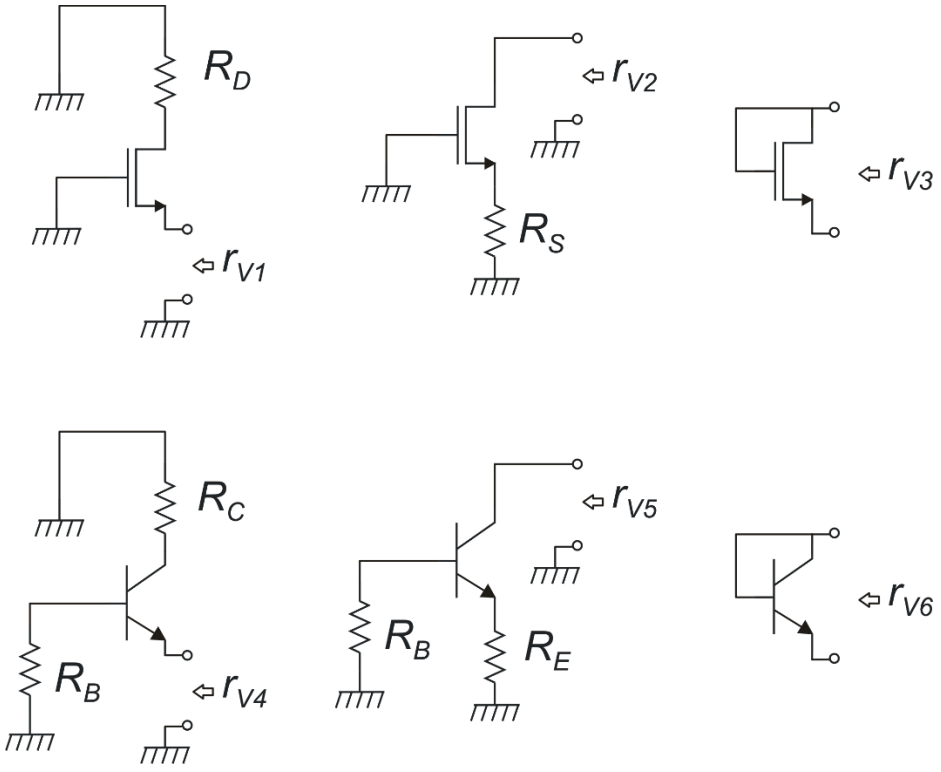

Fig.1.14. Notable cases of small-signal resistance.

Case r_{v2}: Resistance from the drain to ground, with source degeneration (R_S is in series to the source).

Exact result:

$$r_{v2} = R_S + r_d\left(1 + g_m R_S\right) \tag{1.47}$$

If $R_S << r_d$

$$r_{v2} \cong r_d\left(1 + g_m R_S\right) \tag{1.48}$$

The factor $1+g_m R_S$ is a multiplier that is often used to increase the resistance seen from the drain of a MOSFET.

Case r_{v3}: Resistance of a diode-connected MOSFET.

$$r_{v3} = \quad r_D \,//\, \frac{1}{g_m} \quad \cong \frac{1}{g_m} \tag{1.49}$$

Resistances r_{v4}, r_{v5} and r_{v6} concern circuits using bipolar transistors. They correspond to the three MOSFET resistances r_{v1}, r_{v2}, r_{v3}, respectively. Expressions for BJTs are slightly more complicated due to the presence of the base-emitter resistance r_{be}. The following resistances have been calculated using the dc version of the small signal circuit of Fig. 1.13 (same circuit as in Fig.1.13 but with no capacitors). Notice that, from the circuit of Fig.1.13:

$$h_{fe} = \left(\frac{i_c}{i_b} \right)_{v_{ce}=0} = g_m r_{be} \tag{1.50}$$

Furthermore, due to resistance $R_{B,}$ which is placed between ground and the BJT base, it is convenient to consider an equivalent g_{m}, indicated as g_{meq}:

$$g_{meq} = \frac{r_{be}}{R_B + r_{be}} g_m \tag{1.51}$$

If $R_B << r_{be}$, g_{meq} coincides with g_m.

Case r_{v4}: Resistance seen from emitter to ground.

Exact result:

$$r_{v4} = \frac{R_C + r_o}{1 + g_{meq} r_o} // \left(r_{be} + R_B \right) \tag{1.52}$$

Approximate results:

$$g_{meq} r_o >> 1 \quad \text{and} \quad R_{\mathrm{C}} << r_o \Rightarrow \quad r_{v4} \cong \frac{1}{g_{meq}} // \left(r_{be} + R_B \right) = \frac{\left(r_{be} + R_B \right)}{h_{fe} + 1} \cong \frac{1}{g_{meq}} \tag{1.53}$$

Case r_{v5}: Resistance seen from collector to ground, in the presence of emitter degeneration.

Exact result:

$$r_{v5} = R_{Eeq} + r_o \left(1 + g_{meq} R_{Eeq} \right) \tag{1.54}$$

where:

$$R_{Eeq} = \left(R_B + h_{ie} \right) // R_E \tag{1.55}$$

Useful approximations can be found for two opposite conditions:

$$\begin{aligned} &1) \quad R_E << (r_{be} + R_B) \Rightarrow r_{v5} \cong r_o\left(1 + g_{meq} R_E\right) \\ &2) \quad R_E >> (r_{be} + R_B) \Rightarrow r_{v5} \cong r_o\left(1 + h_{fe}\right) \end{aligned} \tag{1.56}$$

Case r_{v6}: Resistance of a diode-connected BJT.

$$r_{v6} = \; r_o \,//\, \frac{1}{g_m} \,//\, h_{ie} \quad \cong \quad \frac{1}{g_m} \tag{1.57}$$

1.6 References

[1] P.R. Gray, P. J. Hurst, S. H. Lewis, R. G. Meyer, "Analysis and Design of Analog Integrated Circuits", 4th edition, John Wiley & Sons, New York, 2001.

[2] Y. Taur, T.H. Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, 2nd edition, 2009.

[3] C. C. Enz, F. Krummenacher and E. A. Vittoz. "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications", Analog Integrated Circuits and Signal Processing, vol. 8, pp. 83-114, 1995.

[4] Y. Tsividis, "Operation and Modeling of the MOS Transistor", 2nd edition, Oxford University Press, New York, 1998.

[5] I. M. Filanovsky, A. Allam "A Mutual compensation of mobility and threshold voltage temperature effects with applications in CMOS circuits". IEEE Trans Circuits and Syst I, vol. 48, pp. 876–884, 2001.

[6] F. Silveira, D. Flandre, and P. G. A. Jespers, "A g_m/I_D Based Methodology for the Design of CMOS Analog Circuits and Its Application to the Synthesis of a Silicon-on-Insulator Micropower OTA", IEEE J. Solid State Circuits, vol. 31, pp. 1314-1319, September 1996

[7] D. E. Ward and R.W. Dutton, " A Charge-Oriented Model for MOS Transistor Capacitances", IEEE J. Solid State Circuits, vol. SC-13, No 5, pp. 703-708, October 1978.

[8] A.S. Sedra, K.C. Smith "Microelectronic Circuits", 7th edition, - 2016 - Oxford University Press, appendix B: SPICE Device Models And Design Simulation Examples Using Pspice And Multisim, available on line: https://global.oup.com/us/companion.websites/9780199339136/student/app/