14.1 OVERVIEW: THE MOSFET DEVICE AND ITS SPICE MODELS

The MOSFET is a charge–control, field–effect device. The acronym "MOS" refers to its construction, which consists of a series of layers forming a metal–oxide–semiconductor (MOS) sandwich. By means of a voltage bias across the oxide layer, which is thin, of thickness typically around 50 nm, electric fields on the order of 10^5 to 10^6 V/cm will be created. These are formidable E–fields, and will have strong effects on the charge and conductance properties of the semiconductor substrate.

The principal consequence of this strong E–field is that it induces a highly–conductive layer of mobile charge in the surface region of the semiconductor substrate. This surface charge layer forms a conductive channel, between two end terminals, usually identified as the "source" and "drain" terminals, with its properties directly controlled by the transverse E–field. The terminal which applies the transverse E–field is called the "gate" terminal, and is the principal control terminal.

The MOSFET can be fabricated at dimensions of microns or less, and therefore lends itself well to the fabrication of high–density VLSI circuits. Most integrated circuits are constructed with the MOSFET as the principal component, not just as a transistor device, but even one that supplants resistances, since the MOSFET device can be constructed at much smaller dimensions than those needed for the typical resistive paths.

Consistent with the concept of a control element, we like to apply the MOSFET as if it were a 3–terminal component. This concept is consistent with the perspective of a transistor as an electrically–controlled 'valve' for electric current, as represented by figure 14.1–1.

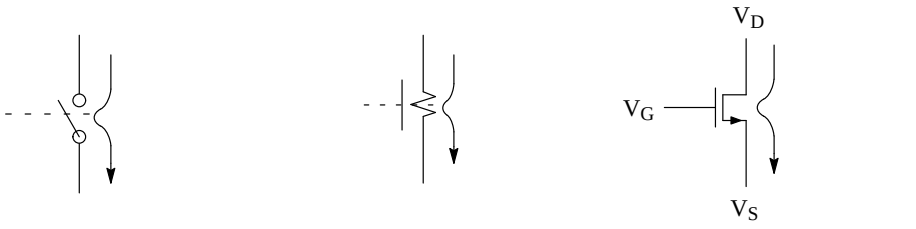

Figure 14.1–1: Conceptual and circuit models of the (n–channel) MOSFET.

These models are all a little too ideal for robust circuit design, but are adequate for first–order conceptual purposes. In order to make a circuit that will actually work, it is necessary to take a closer look at the device, identify the physical mode(s) of its operation, and deploy mathematical models that yield a realistic representation of its operating characteristics,

The conductance of the FET is most directly controlled by the gate–source bias V_{GS}. There is a threshold level of V_{GS}, usually labelled as $V_{TH,}$, which must be reached in order to create the conductive charge layer. In fact, if the bias between gate and any point within the channel drops below this threshold, the device self–limits the level of conducting current, reaching a state usually referred to as "saturation". When the drain end of the channel approaches this limit condition, the gate–drain bias $V_{GD} \rightarrow V_{TH}$, and the conducting charge hypothetically $\rightarrow 0$. Due to this pinching effect on the level of conducting charge, this condition is usually referred to as "pinch–off".

Of course the charge does not really pinch–off to zero, but it makes the concept of a conduction–limiting effect within the channel more graphic. What really happens is that the charge layer self–limits itself to a very small but finite level as the carriers approach terminal velocity, somewhere near the drain end of the channel.

Since $V_{GD} = V_{GS} - V_{DS}$, the pinch–off condition may also be stated in terms of drain–source bias V_{DS} , for which $V_{DS} = V_{DSAT} = V_{GS} - V_{TH}$ as represented by figure 14.1–2. Since there are three terminals to this device, this is a necessary perspective, since the I_D vs V_{DS} characteristics describe the *output* properties of the current channel. Figure 14.1–2 shows these properties for a fixed value of V_{GS}.

It might be noted that the drain–source characteristics are manifested by a finite conductance when V_{DS} = small, consistent with the idea of a conductive bridge layer in between source and drain. But as V_{DS} increases, this channel gets more constricted at the drain end, and conductance rolls off to a zero slope as $V_{DS} \rightarrow V_{DSAT}$. This behavior is consistent with an approximately parabolic form for the $I_D - V_{DS}$ characteristics of the transistor, and makes it reasonable, to first–order, to model the MOSFET electrical behavior as a quadratic equation fitted to the I–V drain characteristics, as represented by Figure 14.1–2.

This quadratic fit gives us the equations:

$$\text{for } V_{DS} < V_{GS} - V_{TH}: \quad I_D = K[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2] \tag{14.1–1}$$

$$\text{for } V_{DS} > V_{GS} - V_{TH}: \quad I_D = K(V_{GS} - V_{TH})^2 \tag{14.1–2}$$

These equations are for the n–channel (nMOS) transistor, for which V_{TH} is (usually) a positive value. Note that the saturation condition, for which $V_{DS} > V_{GS} - V_{TH}$, is also the same as the "pinch–off" condition, $V_{GD} < V_{TH}$. We also should note that the transistor conducts only when $V_{GS} > V_{TH}$, necessary for formation of the charge–layer in the first–place, and if this condition is not met, we say that the transistor is in "cut–off".

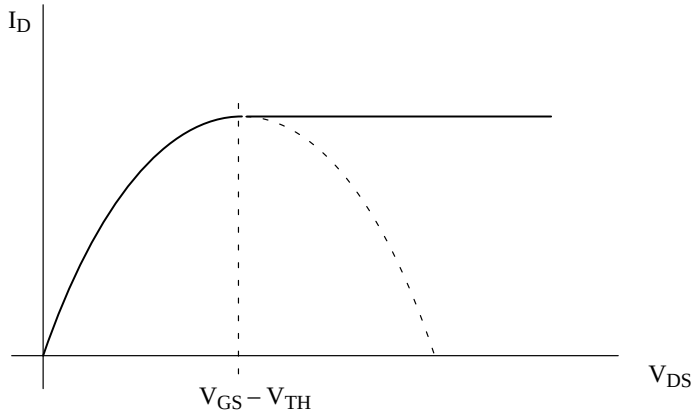

Figure 14.1–1 Fitting of a parabolic model to drain characteristics of the MOS transistor

Equations (14.1–1) and (14.1–2) are also called the *Shichman–Hodges* model[14.1], and are used as the LEVEL–1 model for SPICE.

The conduction characteristics of the p–channel transistor are the same as those of the nMOS transistor, approximately parabolic in form, except that the currents and bias polarities are opposite in polarity. Equations (14.1–1) and (14.1–2) are therefore equally appropriate to the pMOSFET as they are to the nMOSFET. The main change is that all of the junctions are reversed, so that all voltage polarities are reversed. Therefore V_{GS} is negative, V_{DS} is negative, V_{TH} is negative. The conditions for conduction and saturation are also reversed. Therefore for the pMOSFET, the transistor conducts when $V_{GS} < V_{TH}$, and the saturation condition (or pinch–off) is $V_{DS} < V_{GS} -$

V_{TH}, (or $V_{GD} > V_{TH}$).

Once you get the polarities and the conditions straightened out, these equations are very easy to use, and therefore are readily applicable to first–order hand calculations and first–order analysis of a circuit. Unfortunately, physical reality is not quite as accommodating, and we must adjust our circuit analysis accordingly if we are to have a robust design. Equations (14.1–1) and (14.1–2) are compromised by the fact that the MOSFET is actually NOT a three–terminal device but a *four*–terminal device, as represented by Figure 14.2–1 (two pages ahead). The fourth terminal belongs to the substrate. This terminal has value such that the substrate junctions are kept in reverse bias. Although substrate bias V_B has an effect on transistor conductance, it is not usually used for control of the circuit. But it does exercise a strong influence on the uncovered charge layer below the conducting channel, which is a major effect in the definition of the threshold V_{TH}. Therefore the assumption that V_{TH} is a constant, on which equations (14.1–1) and (14.1–2) rely, is not valid. For the sake of simplicity, we often assume a constant V_{TH}, in order to make an approximate assessment of the device behavior using the Shichman–Hodges equations. But it is not a physically good assumption, and if the circuit is not re–evaluated, we may well find ourselves with an unworkable circuit and a localized disturbance in the Force.

As a matter of grace, the Shichman–Hodges equations are adequate for the rough, analytical, analysis of circuit performance needed to initiate the circuit design. But we must relinquish much of this simplicity in order to gain a mathematical model that (1) considers the more complete charge effect of the MOS sandwich and (2) lets our simulation software converge. This upgrade implies a few more parameters. Table 14.1–1 lists parameters that define the LEVEL–2 model of the MOSFET. This model makes a reasonably complete assessment of the device physics underlying the MOS transistor and therefore is used as a basis for most adjustments we might use in reshaping or remodeling a circuit with MOS transistors.

Software is also our main means by which we refine our circuit design. If the circuit is of VLSI form, where it is difficult or impossible to electrically probe the circuit, it may be our only means of assessing critical aspects of its performance.

In the table there are at least six parameters (PHI, VTO, GAMMA, XJ, DELTA) that are needed to define the threshold. They are a result of the charge–control effects which defines V_{TH}. It emphasizes that the regrettable fact that V_{TH} is not constant, but a parameter which is dependent on at V_S, V_D, and V_B. As a consequence, the equation for drain current will be considerably more of a mess than the quadratic Shichman–Hodges model given by equations (14.1–1) and (14.1–2).

The physical model of *I(V)*, generally identified as the gradual–channel, strong–inversion (GCSI) model, is not particularly useful for hand calculations, unless disciplinary mathematical exercises just happen to strike your fancy. It is for use by software, which will apply it in an iterative, Newton–Raphson process. The model of *I(V)* is subject to the corollary that it must include a reasonably good assessment of all of the physical effects without overburdening the iterative analysis. Your task, should you choose to accept it, is to identify the physical effects and the parameters needed thereto, to be able to assess the SPICE simulation of the circuit, and make knowledgeable adjustments in the circuit design.

In this respect the engineer is more of an executive designer, using and applying his/her understanding of the way that the MOS junction and the MOS transistor works to implement a circuit design. Knowledge of transistor effects, and how they are set by the parameters, represent a more and more significant part of the design process. Circuit design, particularly of VLSI circuits, is a process that follows a design cycle, with the simulation of the circuit being a critical step iterating the design to meet physical and tolerance criteria.

Table 14.1–1 SPICE parameters

Symbol	Description	Keyword	Units	Default	Typical
Level	model index	LEVEL			
V_{TO}	zero–bias threshold voltage	VTO	V	0	1.0
K_P	transconductance parameter	KP	A/V^2	2×10^{-5}	3.1×10^{-5}
γ	body–effect parameter	GAMMA	$\sqrt{V}$	0	1.3
ϕ_B	inversion potential	PHI	V	0.6	0.8
λ	channel–length modulation	LAMBDA	1/V	0.0	0.02
R_D	drain ohmic resistance	RD	Ω	0.0	2.0
R_S	source ohmic resistance	RS	Ω	0.0	2.0
C_{BD}	zero–bias B–D junction capacitance	CBD	F	0.0	20F
C_{BS}	zero–bias B–S junction capacitance	CBS	F	0.0	20F
I_S	bulk–junction saturation current	IS	A	1×10^{-14}	1×10^{-15}
ϕ_0	bulk–junction potential	PB	V	0.8	0.8
C_{GSO}	gate–source overlap capacitance	CGSO	F/m	0.0	4×10^{-11}
C_{GDO}	gate–drain overlap capacitance	CGDO	F/m	0.0	4×10^{-11}
C_{GBO}	gate–body overlap capacitance	CGBO	F/m	0.0	4×10^{-11}
R_{SH}	drain & source diffusion resistance	RSH	Ω/sq	0.0	10.0
t_{OX}	oxide thickness	TOX	m	100nm	50nm
N_{SUB}	substrate doping	NSUB	$\#/cm^3$	0.0	4×10^{15}
N_{SS}	surface state density	NSS	$\#/cm^2$	0.0	1×10^{10}
TPG	type of gate material	TPG	–	1	1
	(opposite to substrate)				1
	(same as substrate)				–1
	(aluminum gate)				0
N_{FS}	fast state density	NFS	$\#/cm^2$	0.0	1×10^{10}
X_J	metallurgical junction depth	XJ	m	0.0	1 μm
L_D	lateral diffusion	LD	m	0.0	0.8 μm
μ_S	surface mobility	UO	cm^2/Vs	600	400

14.2 THE MOS JUNCTION

We also need a view of the MOSFET along the drain–source cross–section, as represented by Figure 14.2–1, in order to assess the effect of the E–field and induced charge layers. The cross–section of an nMOS transistor is represented. Note that the substrate is *p–type* semiconductor, and that the connection between source and drain must therefore be an induced channel of n–type carriers, to form the n–channel MOSFET (≡ nMOSFET).

We see that the polysilicon gate lies over a gap in the diffusion path, which, in this case is an n^+ implant. The gap is of length *L*. When the transistor is in its conducting mode, this gap is bridged by field–induced charge–layers.

Note that the basic structure of the active transistor region is of the form of an 'MOS junction', as represented by the inset to Figure 14.2–1, which shows a slice across the transistor structure. The acronym "MOS" is for (M)etal–(O)xide–(S)emiconductor, which is the basic form of the junction "sandwich". For modern transistors the acronym is not completely correct, since polycrystalline silicon or an alloyed form of semiconductor is usually applied in place of a metal (M). But the acronym "SOSFET" is not in the common vernacular, so we will identify these types of transistor all as "MOSFET"s regardless of their religious convictions.

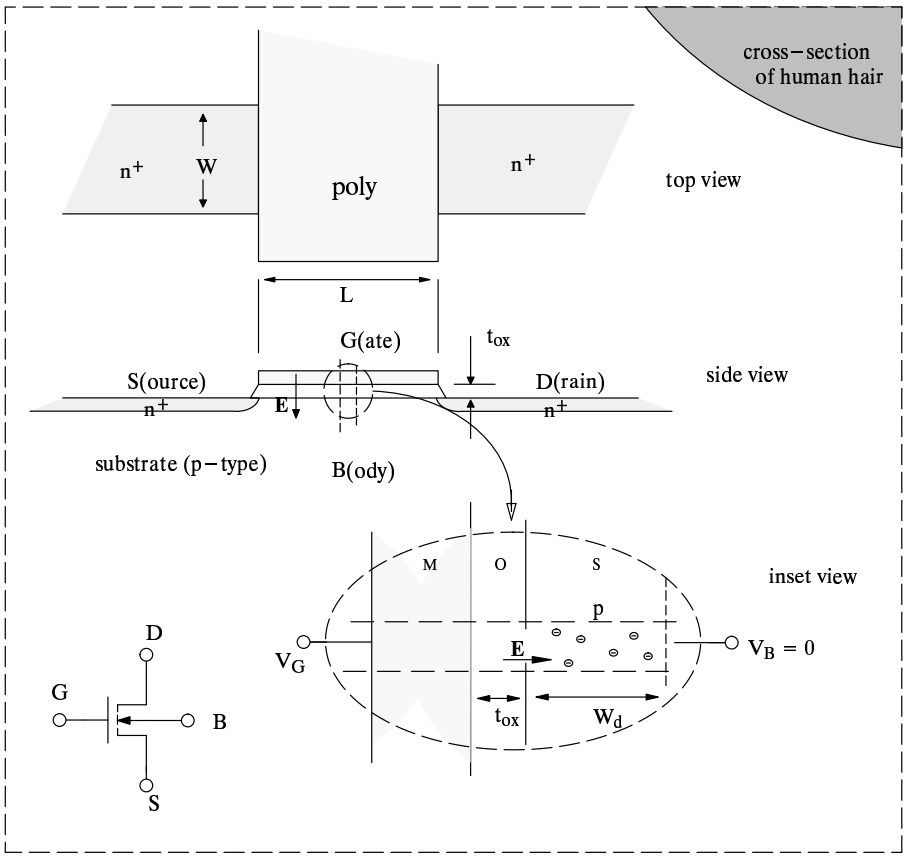

Figure 14.2–1 The nMOS transistor in cross–sectiion

Looking at the MOS 'sandwich', particularly when the (M) is replaced by semiconductor, we see that it is very much like an *np* junction. In this case, the *np* junction has a thin layer of insulating material sandwiched between the *n* and the *p* materials. If the *n* material is very heavily doped, i.e. n^+, which is usually the case, it acts almost like a metal in its conduction and charge properties. Almost. The comparison of the MOS junction to the *np* junction is informative, and is represented by Figure 14.2–2.

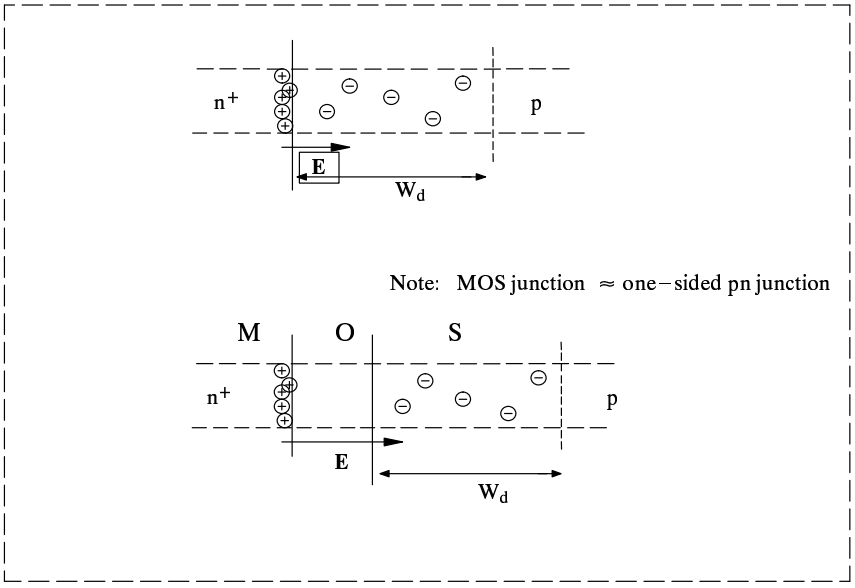

Figure 14.2–2 Comparison between the MOS junction and the *pn* junction

Note that both the MOS junction and the *pn* junction have (i) a built–in potential associated with the work–function difference between the two sides, and (ii) a depletion layer, related to the E–field within the junction. The E–field will be defined by the gate–to–body potential, $V_G - V_B = V_{GB}$.

The layer of charge on the semiconductor side of the MOS junction uncovered by the E–field is of the same form as that uncovered in the *np* junction. This uncovered charge, which is also called the "depletion layer" since mobile charges have been pushed away by the E–field, is of thickness:

$$W_d = L_B\sqrt{2\phi_S/V_T} \tag{14.2–1}$$

where L_B is the extrinsic Debye length, given by

$$\boxed{L_B = \sqrt{\epsilon_S V_T/qN_B}}$$

The parameters ε_S and N_B are the permittivity and the substrate doping, respectively, of the semiconductor. ϕ_S is the potential of the surface relative to the substrate potential V_B.

Since we have distributed charge, we have capacitance. The capacitance per area associated with this layer of uncovered charge is the same as that of the *pn* junction. It is given by the equation

$$C_S = \frac{\epsilon_S}{W_d} = \frac{\epsilon_S}{L_B\sqrt{2\phi_S/V_T}} \tag{14.2–2}$$

Equation (14.2–2) is usually called the "depletion capacitance" of the semiconductor since it is associated with the layer of "depletion" charge. The ratio e_s / L_B is of the form capacitance/area.

Equation (14.2–2) is sometimes called the *depletion approximation*. It assumes that the charges are uniformly uncovered to a finite depth, at which point the effect abruptly terminates. The depletion approximation is reasonably good for ψ_S large, but it fails as $\phi_s \rightarrow 0$. If we make a more exact analysis using Boltzmann statistics, which is done in section 14.9, we would find that when $\psi_s \rightarrow 0$, $C_s \rightarrow e_s / L_B$. Of course as $\phi_s \rightarrow 0$, the E–field also goes to zero,

and the semiconductor bands E_C and E_V are no longer 'bent'. Therefore e_s/L_B is given the name "flat–band" (= zero field) capacitance of the semiconductor. We give it the label

$$C_{FBS} = \frac{\epsilon_s}{L_B}$$

From these definitions we can assess the voltage–induced behavior of the capacitance/area of the MOS junction. Knowledge of the capacitance behavior tells us how the charges are distributed under the influence of the gate field. If we think about the MOS sandwich as if it were two capacitances in series, one associated with the oxide and the other associated with the semiconductor, as shown by Figure 14.2–3, then

$$C_{MOS} = \frac{C_{OX}}{1 + C_{OX}/C_S} \tag{14.2–3}$$

where C_{MOS} is the capacitance/area of the MOS junction and *Cox* is the capacitance/area across the oxide, $C_{ox} = \varepsilon_{ox} / t_{ox}$.

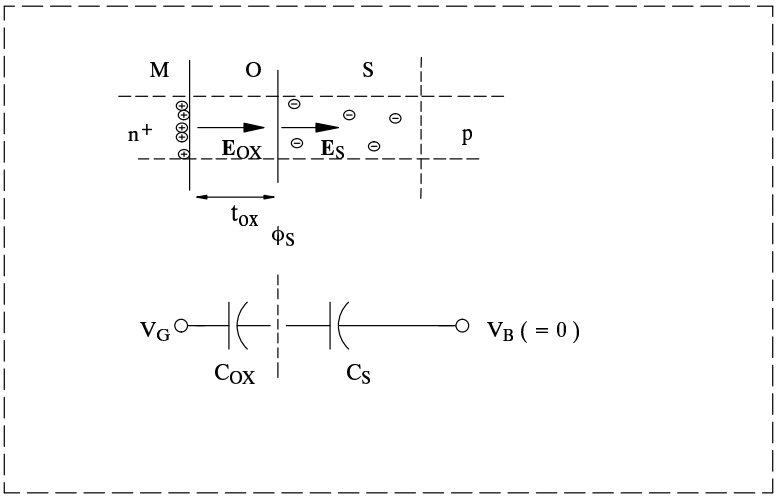

Figure 14.2–3 Capacitance of the MOS junction. The layers are equivalent to capacitances/area in series.

Equation (14.2–3) is not greatly enlightening if we merely relate C_s to ϕ_S. We need its behavior in terms of the potential across the junction $V_G - V_B$. We can make use of Gauss' law to to develop this relationship, as follows:

$$\epsilon_{OX}E_{OX} = \epsilon_{OX}\frac{(V_G - \phi_S)}{t_{OX}} = Q_S = qN_BW_d \tag{14.2–4}$$

The parameters ε_{ox} and E_{ox} are the permittivity and the E–field, respectively, for the oxide layer. We have identified the thickness of the oxide layer as t_{ox}, usually on the order of *50 nm*. Q_s is the charge/area in the semiconductor substrate that is uncovered by the E–field. For lower–level gate fields this charge is the "depletion" charge, for which $Q_s = qN_BW_d$. For stronger gate fields, however, Q_s may include other charge effects. We will check out the stronger field–effects in section 14.4.

It might be noted that a relatively–high gate–field strength is necessary to induce charge effects in the semiconductor. E–fields must be on the order of strength 10 kV/cm to push mobile (+) charges (holes) away from their home sites. After these mobile (+) charges are evicted by the E–field they leave a "depleted" zone of uncovered doping sites. Each empty site is left as a net (–) site. This space–charge layer is the "depletion" layer.
Equation (14.2–4) can be rewritten in the form

$$C_{OX}(V_G - \phi_S) = V_T \frac{\epsilon_S}{L_B}\sqrt{2\phi_S/V_t} \qquad (14.2\text{–}5)$$

Which, with a little manipulation, using equation (14.2–2), gives

$$\frac{C_{OX}^2}{C_S^2} + 2\frac{C_{OX}}{C_S} - 2\frac{C_{OX}^2}{C_{FBS}{}^2}\frac{V_G}{V_T} = 0 \qquad (14.2\text{–}6)$$

Note that this is an equation in C_{ox}/C_S, which is what we need for use in equation (14.2–3). Equation (14.2–6) is quadratic, and only the positive root is applicable since negative capacitance would make no sense. Taking the positive root and applying it to equation (14.2–3), we get

$$C_{MOS} = C_{OX} \Big/ \sqrt{1 + \frac{2C_{OX}^2}{C_{FBS}^2}\frac{V_G}{V_T}} \quad = C_{OX} \Big/ \sqrt{1 + \frac{4}{\gamma^2}V_G} \qquad (14.2\text{–}7)$$

where we have defined a parameter γ, which will turn out to be useful when we get to section 14.4. It is of the form,

$$\gamma = \frac{C_{FBS}}{C_{OX}}\sqrt{2V_T} = \frac{\sqrt{2\epsilon_S q N_B}}{C_{OX}} \qquad (14.2\text{–}8)$$

This parameter is usually called as the “body–effect” coefficient since it is associated with the layer of depletion charge in the ”body” or “bulk” of the semiconductor substrate. It reappears in a number of places in analysis of MOS devices, so you might consider adding it to your analytical menu.

Equation (14.2–7) tells us about the distribution of depletion charge for the MOS junction since capacitance = $\Delta Q/\Delta V$. More about the nature of this junction is represented by Figure 14.2–4.

There are two curves represented by Figure 14.2–4. One curve is equation (14.2–7), which is discontinuous at $V_G = 0$, and for which the depletion approximation is no longer valid. The smooth curve is a more detailed analysis, representing the equilibrium behavior of the electron gas under the influence of the E–field. We see that the equation for the electron gas departs from equation (14.2–7) at the special value $V_G = V_{TH}$. This point is called the “threshold” V_{TH}. At this point it is apparent the strong E–field is not just depleting the substrate but causing another effect on the charge distribution. Since the junction capacitance, *CMOS*, makes a sharp increases to *Cox* at this point, it informs us that charges are accumulating at the surface layer rather than continuing to deplete at greater depths.

Instead of more loose (+) charges being pushed away by the strong E–field, and uncovering more negative doping sites, the increased E–field at $V_G > V_{TO}$ is beginning to induce a major accumulation of loose (–) charges at the oxide–semiconductor surface. Inasmuch as this effect is equivalent to having a thin n–type layer of charge at the surface, it is called “inversion”, as if the p–type substrate at the surface had somehow been changed (inverted) into an n–type material within this thin surface layer.

Figure 14.2–4 Capacitance characteristics of the MOS junction vs V_{GB}.

The charge characteristics of the MOS junction are represented by Figure 14.2–5, which show the various types of effects. They include the case for which the field direction is reversed *($V_G < 0$).* Under this circumstance there is an accumulation of loose (+) charges at the surface. Therefore the capacitance *CMOS* appears as a separation of charges on each side of the oxide, = *Cox*.

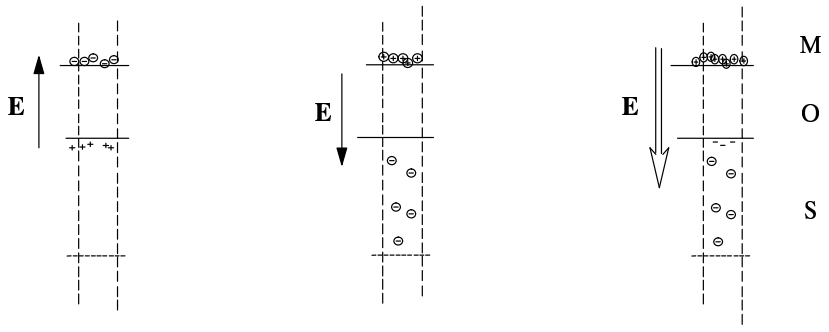

Figure 14.2–5 Charges induced by the gate–body bias V_{GB} for the MOS junction

14.3 BUILT–IN POTENTIALS IN THE MOS JUNCTION

Like most junctions between dissimilar materials, the MOS junction has potentials which are built–in. In the MOS junction these potentials result from two different types of effects:

(1) the work–function potential, (or, electron potential), and
(2) potentials resulting from trapped charges in the oxide.

These effects give the transistor its own personal contribution to the gate potential, ΔV_G. This potential represents an offset which becomes a significant contribution to the device threshold, V_{TH}. Our mission, should we choose to accept it, is to determine the nature of these effects, and see how their contributions affect the V_{TH} in our circuits.

The work–function potential:

The work–function potential is a simple effect that we can make more complicated, if we so choose, by introducing such concepts as the electron affinity of the insulator and vacuum energy–levels. Although such concepts are often included in more comprehensive treatments, these complications are not necessary. All that we need is to identify the electron potentials on either side of the oxide layer, since it is the *difference* that defines the electron (or work–function) potential.

The nature of the work–function potential is more evident when we make a comparison of the MOS junction and the *pn* junction. Assuming an nMOS transistor (has p–type substrate) with an n–type polysilicon gate, it is evident that we have an *np* junction, if we ignore the small matter of the oxide layer in the middle. We know that at equilibrium, the Fermi energy, which is our index, must be the same everywhere, across the junction, across the *n*, the *p* and oxide, the connecting wires, and the rest of the world if we so desire, provided no external biases are applied. For the *pn* junction, this equilibrium condition identifies a "built–in" potential ϕ_0 across the junction, which the electrons can easily see and feel. Comparison between the band diagrams for the MOS junction and the *np* junction, as represented by Figure 14.3–1, shows that the *same* built–in potential, ϕ_0 exists for the MOS junction, being merely the difference in electron potential between the two sides. The difference potential will be of value = $\phi_p - \phi_n$, independently of what material exists in middle of the sandwich.

We note that the orientation of the built–in potential is important. For the nMOS structure represented by Figure 14.3–1, $\Delta V_G = \phi_p - \phi_n$. Note that if we happened to have a metal instead of an n–type semiconductor as gate material, we might identify the electron potential of the gate as ϕ_m. Since the substrate is always semiconductor material, we might identify its electron potential, more generally, as ϕ_s. Therefore we specify the built–in work–function potential as

$$\Delta V_G = \phi_s - \phi_m = -\phi_{ms} \tag{14.3–1}$$

where ϕ_m is the work–function (or electron potential) of the gate and ϕ_s is the work–function (or electron potential) of the semiconductor.

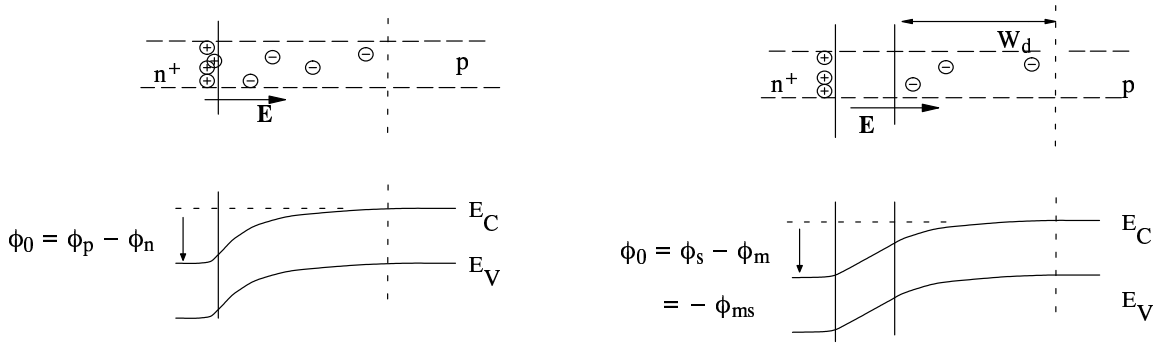

Figure 14.3–1 The work–function potential ϕ_{ms} for the MOS junction, in comparison to *pn* junction potential ϕ_0.

****

EXAMPLE 14.3–1: Suppose we have a gate material of polysilicon, doped at concentration 10^{19}*#/cm*^{3} of donors, and a semiconductor substrate doped at density of 10^{15}*#/cm*^{3} of acceptors.

SOLUTION: $\phi_m = -V_T \ln(N_D/n_i) = -0.525$ *V.* † $\phi_s = +V_T \ln(N_A/n_i) = 0.287$ *V.*

We have used $n_i = 1.5 \times 10^{10}$#/ cm^3 and $V_T = .02585$V (at 300 K).

RESULT: $\phi_{ms} = -0.525\ V - 0.287\ V = -0.812$ *V.*

****

† Note: If the gate is heavily doped, as was represented by this example, and which is often the case, then the gate work–function potential is taken to be approximately +0.56 V or –0.56 V, default. This option is controlled by the SPICE parameter TPG, which is +1 if the gate is of doping opposite to that of the substrate, e.g. n–type gate and p–type substrate, and is –1 if the gate is of doping which is like to that of the substrate, e.g. p–type gate and p–type substrate. SPICE will establish a default gate work function of $\phi_m = \pm 0.56$ *V,* according to the substrate type and this parameter.

Charges trapped in the oxide:

The other built–in effect is due to trapped charges. When ionic charges are trapped in the oxide layer, which is usually unavoidable, then the fact that these charges are proximal to the semiconductor surface can induce a relatively strong field in the semiconductor material. This effect is represented by Figure 14.3–2.

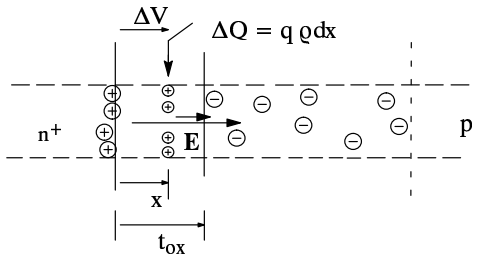

Figure 14.3–2 Effect of a thin layer of charge in the oxide on threshold

Analysis is straightforward, using the known relationship between charge and voltage. From our definition of capacitance we do know that

$$\Delta V = \frac{\Delta Q}{C}$$

where $C = \varepsilon_{ox}A/x$, x being the distance from the gate plane to the plane of the thin charge layer, as shown by Figure 14.3–2. A is the area of the gate. This view is sufficient to generalize the process, because we can assume that we have a distribution of infinitesimally thin layers, i.e. of thickness dQ. Then the effect can be identified in terms of a charge density $q\rho(x)$:

$$\Delta \frac{Q}{A} \rightarrow q\rho(x)dx$$

so that

$$\Delta V = \int dV = \frac{1}{C}\int dQ \qquad \boxed{= \frac{q}{C_{OX}}\int_0^{t_{OX}} \frac{x}{t_{OX}}\rho(x)dx} \tag{14.3–2}$$

We have taken the convenience of factoring out the $1/C_{ox}$ from the integrand, which we accomplished by multiplying and dividing by t_{ox}. This modification makes the end analysis much simpler. We can be a little more sophisticated with our mathematics to confirm that ΔV is a positive quantity for trapped positive charges, but it should be apparent that positive charge in the oxide will induce an E–field from left–to–right, corresponding to a positive built–in potential from gate to substrate.

****

EXAMPLE 14.3–2: What is the resulting ΔV if a fabrication process which creates a distribution of charge in the oxide of density $\rho(x) = \rho_0 (1 - x^2/t_{ox}^2)$. The total dose of charge/area is $N_{ox} = 10^{11}\#/cm^2$. Assume $t_{ox} = 69\ nm$.

SOLUTION: Note that ρ_0 is not a given. It must be determined from N_{ox}, the dose/area. This dose/area is related to density by

$$N_{OX} = \int_0^{t_{OX}} \rho(x)dx \tag{14.3–3}$$

where $\rho(x) = \rho_0(1 - x^2/t_{ox}^2)$.

Carrying out the mathematics of (14.3–3) we eventually will get $N_{ox} = \rho_0 \times t_{ox} \times (1 - 1/3)$. Now we can use the result of our previous analysis, (14.3–2), to give

$$\Delta V = \frac{3}{2}\frac{N_{OX}}{t_{OX}} \times \frac{q}{C_{OX}} \times t_{OX}(\frac{1}{2} - \frac{1}{4}) = \frac{3}{8}\frac{qN_{OX}}{C_{OX}}$$

For oxide thickness given, $C_{ox} = \varepsilon_{ox}/t_{ox} = 5 \times 10^4\ pF/cm^2$,

where we have used $\varepsilon_{ox} = 3.9 \times (8.85 \times 10^{-14})F/m = 0.345\ pF/cm$.

Therefore:

RESULT:

$$\Delta V = \frac{3}{8} \times \frac{(1.6 \times 10^{-7} pC)}{5 \times 10^{4} pF/cm^2} = 0.12V$$

****

Note that we generally identify N_{ox}, the charge density per area, rather than ρ_0. This choice is made because we may choose to *implant* ionic charges, and we implant them as a charge "dose". This implantation gives us a way to adjust the built–in potential term.

Combining effects, which are given by equations (14.3–1) and (14.3–2) we see that the total built–in potential term is then

$$\Delta V_G(bi) = -\phi_{ms} + Q_{OX}/C_{OX} \tag{14.3–4}$$

where, for convenience, we have lumped all of the distribution of the oxide charge into a single term Q_{ox}. A more strict analysis may choose to break the lump Q_{ox} up into several parts. For the SPICE circuit simulator, fixed trapped charges in the oxide are indicated by the parameter N_{SS}. Strictly speaking,

$$Q_{OX} = q \int_0^{t_{OX}} \frac{x}{t_{OX}} \rho(x) dx = q\gamma_M N_{OX} \tag{14.3–5}$$

which tells us that, in general, we should NOT just take Q_{ox} to be $= qN_{ox}$. The distribution factor, γ_M, is, in general, somewhere between *0* and *1*. SPICE will assume that $NSS = \gamma_M N_{ox}$.

A more sophisticated analysis would look at types and stabilities of these trapped charges[14.3–1], some of which, called "fast–states", will migrate as result of the gate field. SPICE, in fact, defines a parameter NFS which represents the fast–state dose present. But in this case we will generalize to just those that are fixed within the oxide as result of impurities incurred during the gate–oxide process, which are usually positive ions.

As a matter of convention, we usually identify equation (14.3–4) in terms of the voltage that we need to apply to the gate to bring the E–field in the semiconductor to zero. This voltage is called the "flat–band" voltage, V_{FB}. A zero–field is equivalent to the situation where the energy bands are not bent (note that the presence of an E–field always bends the bands). Hence

$$\Delta V_G(bi) = -V_{FB} = -\phi_{ms} + Q_{OX}/C_{OX}$$

or,

$$\boxed{V_{FB} = \phi_{ms} - Q_{OX}/C_{OX}}$$

The use of V_{FB} is a handy way to relate all of the built–in effects to a reasonably direct physical measurement. As it turns out, V_{FB} is not used as a SPICE parameter since it can be embedded under another parameter, V_{TO}. But is not improper to assume that some other circuit simulation software or some different version of SPICE may elect to make more direct use of it as a parameter.

14.4 THE CHARGE–CONTROL MODEL OF THE MOSFET

So far, we have seen that the MOS junction has many similarities to the *pn* junction, in that it includes such effects as work–function potentials and depletion charge. In this respect, we have discovered that, for every charge layer that can be identified in the junction, we can also identify a potential. This type analysis is called "charge–control" analysis, since the charge effects can be interpreted as being under control of applied voltages.

In examining the MOS capacitance, we found (equation 14.2.5) that

$$C_{OX}(V_G - \phi_S) = Q_S \tag{14.4–1}$$

where $Q_S = V_T C_{FBS}\sqrt{2\phi_S/V_T}$. With a little manipulation, and use of the definition given by equation (14.2.8) for γ, we can change this to

$$C_{OX}(V_G - \phi_S) = \gamma C_{OX}\sqrt{\psi_S} \tag{14.4–2}$$

where, if we also include the reference level V_B, would read

$$C_{OX}(V_G - \phi_S) = \gamma C_{OX}\sqrt{\phi_S - V_B} \tag{14.4–3}$$

In section 14.2 we identified Q_s as being entirely the uncovered depletion charge Q_B. But then we realized that at some potential $V_G = V_{TH}$, (minority type) charges begin accumulating at the surface, as shown by Figures 14.2.4 and 14.2.5. Therefore, for strong fields, we need to subdivide charge Q_s into two different types,

$$Q_S = Q_B + q_I \tag{14.4–4}$$

where q_I is the thin sheet of minority–carrier charge that is accumulated at the surface by the "pull" of the strong gate field. Highly conductive, this layer of charge is easily analyzed by modifying equation (14.4–1) as follows:

$$C_{OX}(V_G - \phi_S) = Q_B + q_I \tag{14.4–5}$$

q_I is usually referred to as the "inversion" charge layer. The condition for which this inversion layer charge begins to be of significant conductivity is approximately at the state when $\phi_S \approx 2\phi_F$, ϕ_F being the Fermi potential of the substrate, given by

$$\boxed{\phi_F = V_T \ln(N_{SUB}/n_{i)}}$$

where we have indicated the substrate doping as N_{SUB} rather than the less explicit form N_B used in section 14.2. The value of ϕ_S at which inversion occurs we will call ϕ_B. This condition can be seen by Figure 14.4–1, which shows the band–bending when $V_G \approx V_{TH}$. At this point the bands are bent so that the Fermi level (which defines the equilibrium level of carrier concentration) is approximately as close to the conduction band E_C as it is to the valence band E_V deep within the semiconductor, far from the junction fields. This condition is represented by Figure 14.4–1.

A more accurate value of ϕ_B results from analyzing the effect of the fields on the Boltzmann statistics, as will be done by section 14.9, for which[14.4–1]

$$\boxed{\phi_B = 2.1\phi_F + 2.08V_T} \tag{14.4–6}$$

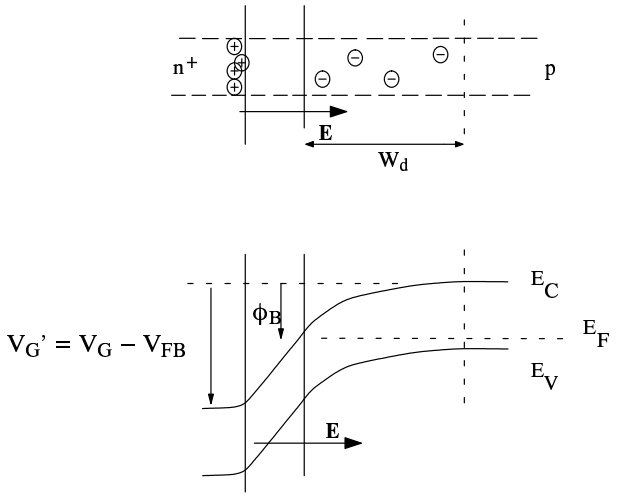

Figure 14.4–1 Band–bending at the onset of inversion.

At the value ϕ_B, whether we use the traditional value $2\phi_F$, or equation (14.4–6), it is assumed that at or about $\phi_S = \phi_B$, the highly–conductive inversion layer q_I is formed.

The SPICE parameter corresponding to ϕ_B is PHI.

Now, when the MOS junction has source and drain nodes attached to either side, as shown by Figure 14.4–2, then they will make conductive contact with the inversion charge layer when the inversion condition $\phi_S = \phi_B$ is met. Since we expect that the voltage will change gradually from V_S to V_D, then we identify the behavior of the surface potential as

$$\phi_S = \phi_B + V \tag{14.4–7}$$

where $V_S < V < V_D$. Equation (14.4–7) is called the *gradual–channel approximation* (GCA).

If we now apply equation (14.4–5) to the gradual–channel approximation we then get

$$C_{OX}(V_G - \phi_B - V) = Q_B + q_I \tag{14.4–8}$$

In order to accommodate the built–in contributions to the gate voltage, we need to make a correction to V_G of the form,

$$V_G \rightarrow V_G + \Delta V_G(BI) = V_G - V_{FB}$$

Furthermore, from (14.4–3), which is the case where $Q_S = Q_B$, we can identify the depletion contribution Q_B of (14.4–8) as

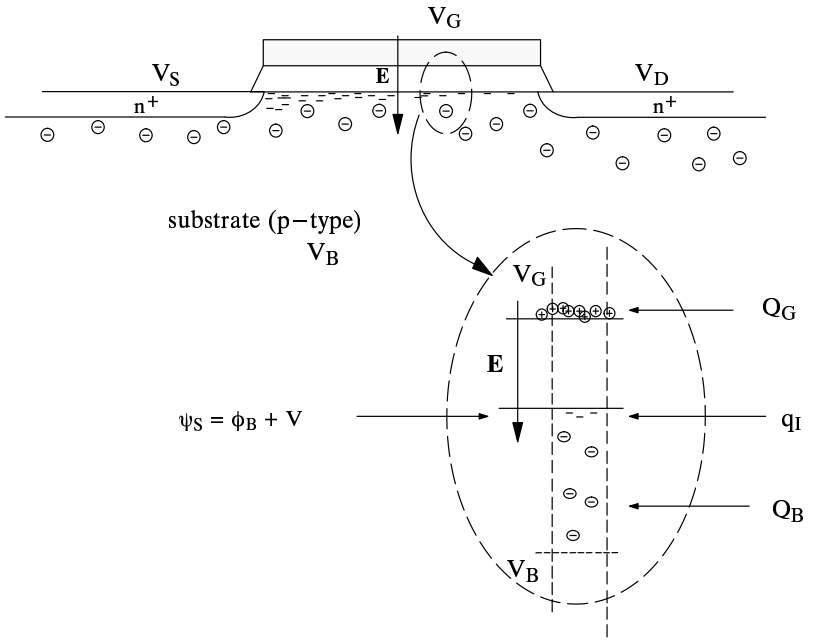

Figure 14.4–2 The MOS transistor and the gradual–channel approximation.

$$Q_B = \gamma C_{OX}\sqrt{\phi_S - V_B} = = \gamma C_{OX}\sqrt{\phi_B + V - V_B} \tag{14.4–9}$$

where we have taken used (14.4–7) to specify ψ_S for the case for which the transistor is in a conducting state.

Now what we see that equation (14.4–8) is a means of defining q_I. Combining equations (14.4–8) and (14.4–9), and solving for q_I, we get

$$q_I = C_{OX}(V_G - V_{FB} - \phi_B - V) - \gamma C_{OX}\sqrt{\phi_B + V - V_B}$$

which can be rewritten as

$$q_I = C_{OX}(V_G - (V_{FB} + \phi_B + \gamma\sqrt{\phi_B + V - V_B}) - V) \tag{14.4–10}$$

At $V = V_S$, we can see how q_I relates to the source potential:

$$q_I = C_{OX}[V_G - V_S - (V_{FB} + \phi_B + \gamma\sqrt{\phi_B + V_S - V_B})] = C_{OX}(V_G - V_{TH} - V_S)$$

This condition defines threshold for the MOS transistor, which is associated with formation of a conducting inversion layer at the *source* end, as

$$\boxed{V_{TH} = V_{FB} + \phi_B + \gamma\sqrt{\phi_B + V_S - V_B}} \tag{14.4–11}$$

where we usually write $V_S - V_B$ as V_{SB}.

Equation (14.4–11) is an interpretation of threshold V_{TH} in terms of charge and junction effects. It shows that threshold is associated with the onset the highly conductive *Inversion* layer at the oxide–semiconductor interface.

This inversion layer is of the form of a sheet charge, and so the analysis which we have used in deriving (14.1.11) is also called the “charge–sheet” analysis.

SPICE uses a parameter VTO, the zero–bias threshold defined as

$$V_{TO} = V_{FB} + \phi_B + \gamma\sqrt{\phi_B} \tag{14.4–12}$$

corresponding to $V_{SB} = 0$. This parameter eliminates the need to include V_{FB} in the SPICE parameter list, since V_{TH} can be expressed as

$$\boxed{V_{TH} = V_{TO} + \gamma\,(\sqrt{\phi_B + V_{SB}} - \sqrt{\phi_B})} \tag{14.4–13}$$

14.5 THRESHOLD ADJUST

The first term of the threshold equation (14.4.11) represents a potential that is built into the junction

$$V_{FB} = \phi_{ms} - Q_{OX}/C_{OX}$$

Note that excess trapped charges help to define threshold voltage. In many respects this effect is a hindrance, and it is necessary to purge the MOS junction of impurities which cause excess charges. The environment must therefore be of extreme cleanliness, with high quality and high–purity of materials and environment. It is therefore not likely that MOS transistors can be made in the back of your garage, if so equipped with furnaces, etc.

However, the built–in charges are also the means by which we may adjust the threshold up or down. Charges can be implanted through the gate and gate oxide into the oxide–semiconductor interface by means of a high–voltage ion gun, also referred to as an ion–implanter. This process is represented by Figure 14.5–1. The thin gate oxide may suffer a little, but the damage will be annealed out by the high temperatures used in a later step of the fabrication process.

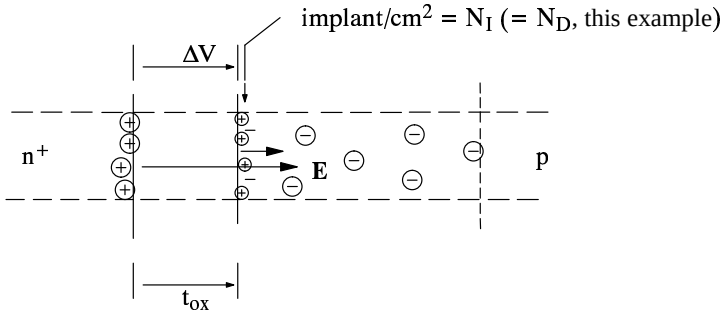

Figure 14.5–1 Effect of an implant layer of charge approximately at the oxide–semiconductor interface

Whether or not these ions are in the oxide or in a layer close to the oxide, the effect can be treated by the same analysis as used to derive equation (14.3.2). Implanted ions assume that a layer of the form

$$\rho(x) = N_I\delta(x - t_{ox})$$

is created by implant, where N_I represents *implant* dose/area of charges, driven in at a depth localized at or near to the oxide–semiconductor interface.

Note that, using (14.3.2) this gives us a *threshold adjust* of

$$\boxed{\Delta V_{TH} = -\frac{qN_I}{C_{OX}}} \qquad (14.5\text{–}1)$$

In general, we implant donor and acceptor impurities, so that N_I will be either of the form $N_D{}^+$ or $N_A{}^-$. Note that a negative shift of threshold results from an implant of donor ions, and a positive shift from an implant of acceptor ions.

14.6 KEEPING TRACK OF THE POLARITIES

The threshold equation (14.4.11) shows that there are three basic terms which define threshold voltage:

1. V_{FB} = the flatband built–in junction voltage
2. ϕ_B = the potential needed to create inversion
3. $\gamma\sqrt{\phi_B + V_{SB}}$ = the body effect

The previous sections have shown that these terms are based either on the presence of a charge distribution or on work functions. Each therefore has a polarity.

For example, the potential needed to create inversion ϕ_B, is of polarity defined by the type substrate material,

$$\phi_B = \pm\ (2.1\phi_F + 2.08V_T) \qquad (14.6\text{–}1)$$

which takes the (+) sign if the substrate is p–type ($\Rightarrow$ nMOS transistor) and the (–) sign if the substrate is n–type ($\Rightarrow$ pMOS transistor). This equation is derived from a field condition, and the polarity of the field defines the sign.

Note that the body–effect term likewise is dependent on the substrate. As we saw in section 14.2, (Figure 14.2.5) a positive gate potential has to be applied in order to induce the (negative) depletion charge in the p–type substrate. If we had analyzed a pMOS junction, with n–type substrate, then a negative potential would have had to be applied.

We can indicate the polarity of the body–effect contribution by means of

$$\Delta V_{TH} = \pm\ \gamma C_{OX}\sqrt{|(\phi_B + V - V_B)|} \qquad (14.6\text{–}2)$$

where the (+) sign corresponds to a p–type substrate and the (–) sign corresponds to an n–type substrate. Note that an nMOS transistor requires a p–type substrate, and this term represents a large positive fraction of V_{TH} for the nMOS enhancement transistor, and conversely for the pMOS transistor.

The flat–band term and the threshold adjust term contribute to the threshold according to equations (14.3.5) and (14.5.1) as:

$$\Delta V_{TH} = -qN_I/C_{OX} + (\phi_{ms} - qN_{SS}/C_{OX})$$

where we have taken liberty of indicating that the charge Q_{ox} distributed in the oxide can be expressed as qN_{SS}. The polarity of the implant ions N_I and the oxide–trapped ions N_{SS} may be considered in terms of *donor* impurities,

which form *positive* ions, and make a *negative* contribution to the threshold. The converse is true for acceptor impurities.

Note that for the threshold terms addressed so far, i.e. inversion, body–effect, oxide ions, and implant ions, we see that *n–type* impurities yield terms of (–) polarity, and *p–type* impurities yield terms of (+) polarity. All terms except ϕ_{ms} can follow this rule. As we saw in example 14.3.1, a heavily doped gate silicon gate material will usually have $\phi_m = \pm$ 0.56V, with n+ doping taking the (–) sign and p^+ doping taking the (+) sign.

However, the substrate, in this case, is a subtractive term, so that a p–type substrate will subtract, and an n–type substrate will add, to the ϕ_{ms} term. The SPICE terminology uses the parameter TPG, toggling $\phi_m = \pm$ 0.56V according to whether the transistor is designated as nMOS or as pMOS. If TPG = 1, and the transistor is nMOS, then ϕ_m = –0.56V. If TPG = –1 and the transistor is nMOS, then ϕ_m = + 0.56V, resulting in a much smaller ϕ_{ms}. If TPG = 0, then SPICE assumes that ϕ_m will take a default value, usually $\phi_m = 0$. It may take a negative value if the work function for a metal is inserted in the default list.

14.7 CHARGE–SHARING AND NARROW–CHANNEL EFFECTS

In the analysis of the threshold, it should be apparent that one of the major terms of V_{TH}, if not the dominant one, is the "body–effect". Behavior of channel conductance g_I as a function of V_{GS} and V_{BS} is indicated by Figure 14.7–1, which shows the effect of the "body–effect coefficient" γ.

Figure 14.7–1 Plot of g_I vs V_{GS}. In this case $\gamma = 1\sqrt{V}$, $V_{FB} = -1V$, and $\phi_B = 1V$, realistic, but simplified approximate values for the nMOS transistor parameters. We see that when we have a V_{SB} of as little as *3V*, the threshold V_{TH} will approximately double in value.

There is a tendency to make transistors at dimensions on the order of microns and less. Therefore it is important that we see what effect these shrinking dimensions will have on the transistor. These effects are not easily modeled, and therefore the analysis is qualitative as much as it is quantitative.

Figure 14.7–2 shows the effect of a reduced channel length on the threshold. We see that the source–drain junctions have some influence over the depletion charge under the gate. For long–channel devices, this influence is negligible, since the source–drain depletion regions are only a small fraction of the channel region. For short–channel devices, the source–drain ends are a large fraction of the depletion region, and consequently "share" a larger portion of this part of the body–effect. We therefore usually identify this effect as "charge–sharing".

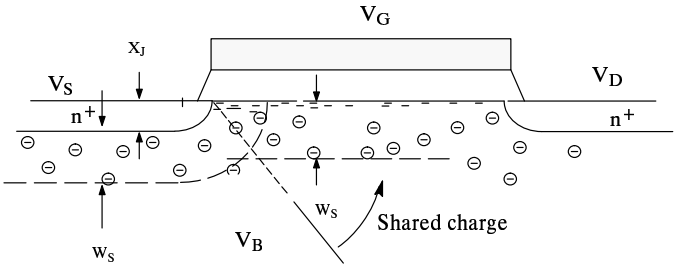

Figure 14.7–2 Charge–sharing effects in the short–channel MOSFET. An nMOS transistor is shown. Note that the effect is related to the junction depth of the source–drain regions, indicated in the Figure as X_J.

There are a number of different ways to approach this particular effect. The one which used by SPICE takes a geometrical, approach in which γ is reduced by two end terms, α_S and α_D, as follows:

$$\gamma_S = \gamma(1 - \alpha_S - \alpha_D) \tag{14.7–1}$$

where

$$\alpha_S = \frac{X_J}{2L}\left(\sqrt{1 + 2W_S/X_J} - 1\right) \tag{14.7–2}$$

and

$$\alpha_D = \frac{X_J}{2L}\left(\sqrt{1 + 2W_D/X_J} - 1\right) \tag{14.7–3}$$

where L is the channel length, and where W_S and W_D are the depletion depth of the source and drain junctions, respectively, given by

$$W_S = L_B\sqrt{2(V_S - V_B + \phi_0)/V_T}$$

and

$$W_S = L_B\sqrt{2(V_D - V_B + \phi_0)/V_T}$$

where ϕ_0 is the built–in potential of these junctions.

These terms can be obtained geometrically from Figure 14.7–2. We will not attempt to derive (14.7–2) and (14.7–3), even though their derivations are relatively straightforward. The main intent is to indicate that the "charge–sharing" effect can be defined by the junction depth X_J, or the parameter XJ, as used by SPICE. We see that as L is reduced, then α_s and α_d increase in magnitude, diminishing the body effect and reducing the magnitude of V_{TH}. This is represented by Figure 14.7–5.

We see that even the threshold is of a somewhat more complicated form than we would want to calculate by hand. From equation (14.7.3), we see that the threshold depends on V_D. In this sense, V_{TH} is more a *conduction* threshold rather than a simple *inversion* threshold, and depends on the biases V_S, V_D, and V_B associated with the MOSFET.

It also depends on the width of the gate, as represented by Figure 14.7–3. This Figure represents the “narrow–channel” effect. As we see from the figure, the *fringing* lateral fields also command a finite fraction of depletion charge. If the gate is wide, this fraction is small. If the gate is narrow, this fraction is large.

This effect can also be analyzed geometrically by including the lateral areas indicated by Figure 14.7–3. But it also will vary along the channel since channel potential *V,* and hence depletion effects, will vary from source–to–drain. The “narrow–channel” effect therefore adds a term to the depletion charge Q_B, equation (14.4.9) of the form

$$\delta \frac{\epsilon_S \pi}{4WC_{OX}}(\phi_B + V - V_B) \tag{14.7–4}$$

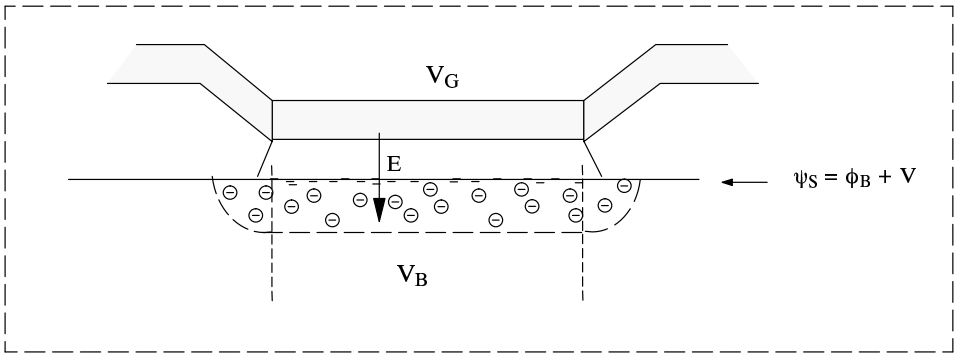

Figure 14.7–3 Narrow–channel effects in the MOSFET. An nMOS transistor at an end–view cross–section is shown.

Note that the magnitude of this term is defined by the factor δ, which under SPICE, is called DELTA. For $V = V_{S,}$, we see that the “threshold” will therefore increase as *W* decreases. This is represented by Figure 14.7–5.

Figure 14.7–4 Representative plots showing the effect on threshold of (a) short–channel and (b) narrow–channel effects in the MOSFET.

14.8 THE MEYER MODEL OF THE MOSFET

We can determine link the conductance of the channel to charge–control analysis by means of the conductivity within the channel, given by

$$\sigma_I = q\mu_S n_I \tag{14.8–1}$$

where μ_S is the mobility of the carriers in this surface layer. Note that n_I varies monotonically from source to drain, since it is affected by the channel voltage V as it varies from V_S to V_D. Therefore all that we need to do to evaluate the *I–V* behavior is to define conductance between source and drain in terms of the conductivity along the channel. The details of the charge layers and coordinate framework are indicated by Figure 14.8–1.

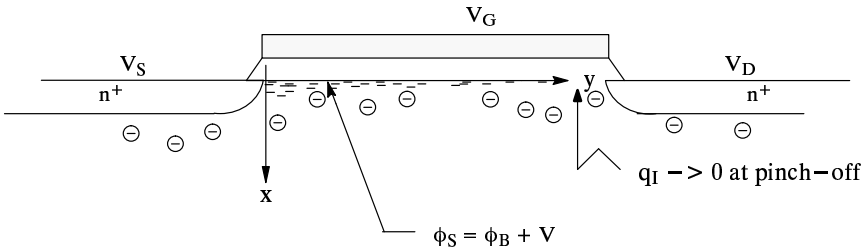

Figure 14.8–1 The nMOS transistor in cross–section.

Referring to the figure for the coordinates, current density in the y–direction is given by

$$J_y = \sigma_I E_y = -q\mu_S n_I \frac{\partial V}{\partial y} \tag{14.8–2}$$

Then the current in the y–direction is given by

$$I_y = \int J_y dA = -\int_0^W \int_0^\infty \left(q\mu_s n_I \frac{\partial V}{\partial y}\right) dx dz \tag{14.8–3}$$

If we make the approximation

$$\int_0^\infty \mu_s q n_I \frac{\partial V}{\partial y} \approx \mu_s \frac{\partial V}{\partial y} \int_0^\infty q n_I dx \tag{14.8–4}$$

This approximation is equivalent to the assertion that most of the inversion charge is concentrated in a thin "sheet–charge" layer at the surface, and that the transverse mobility at the surface, μ_S and field *dV/dy* do not change much over the depth (x–direction) of this thin conductive layer.

It also lets the charge per area be a separable function, i.e.

$$q_I = \int_0^\infty q n_I dx$$

The sheet charge interpretation avoids any need for defining a depth for the highly conductive layer of inversion charge. Since it is of the form of a electron gas accumulated at the surface, any attempt to define a depth may require us to determine whether or not this layer may have "condensed" into a *Fermi liquid* form rather than a Fermi gas. We have identified a reasonably good form for q_I in section 14.4, and can apply it to (14.8–4) without need for any additional qualification:

$$\begin{aligned} I_y &= -\int_0^W \left(\mu_s q n_I \frac{\partial V}{\partial y}\right) dz = -\left(\mu_s q n_I \frac{\partial V}{\partial y}\right) \int_0^W dz \\ &= W\mu_s C_{OX}\left(V_G - V_{FB} - \phi_B - \gamma\sqrt{\phi_B + V - V_B} - V\right)\frac{\partial V}{\partial y} \end{aligned} \tag{14.8–5}$$

Note that there is no dependence of the integrand on *z*, and therefore the integral in *z* merely returns the width of the device, *W*, as a cross–section factor.

Recognizing that $I_y = -I_D$, and evaluating this differential equation gives

$$\int_0^L I_D dy = W\mu_s C_{OX} \int_{V_s}^{V_D} \left(V_G - V_{BI} - V - \gamma\sqrt{\phi_B + V - V_B}\right) dV \tag{14.8–6}$$

where, we have used for convenience, $V_{BI} = V_{FB} + \phi_B$

Evaluating (14.8–6) we get

$$I_D = \frac{W}{L} K_p \left\{\frac{1}{2}\left[(V_{GS} - V_{BI})^2 - (V_{GD} - V_{BI})^2\right] - \frac{2}{3}\gamma\left[(\phi_B + V_{DB})^{3/2} - (\phi_B + V_{SB})^{3/2}\right]\right\}$$

$$\boxed{I_D = K_p \frac{W}{L}\left\{(V_{GS} - V_{BI})V_{DS} - \frac{1}{2}V_{DS}^2 - \frac{2}{3}\gamma\left[(\phi_B + V_{DB})^{3/2} - (\phi_B + V_{SB})^{3/2}\right]\right\}} \tag{14.8–7}$$

where $K_p = \mu_S C_{ox}$ is the SPICE parameter KP. The conduction coefficient of equation (14.1.1) is $K = \frac{1}{2}\mu_s C_{OX}\frac{W}{L}$.

Some treatments of the Meyer model may elect to use $\beta = \mu_s C_{OX}\frac{W}{L}$ instead of *K*.

Equation (14.8–7) is the charge–control equivalent to equation (14.1.1). It is the form used by the LEVEL–2 of SPICE.

Since the threshold is voltage–dependent, the condition for saturation is not as concise as $V_{DS} = V_{GS} - V_{TH}$. Assuming that saturation corresponds approximately to the classical "pinch–off" where $q_I = 0$ for some $V = V_D$, equation (14.4.10) gives

$$V_G - V_{FB} - \phi_B - \gamma\sqrt{\phi_B + V_D - V_B} - V_D = 0$$

This equation is quadratic in $\phi = \phi_B + V_D - V_B$. With a little manipulation, the quadratic equation will be

$$\phi^2 - 2\phi\left(V_{GFB} + \frac{\gamma^2}{2}\right) + V_{GFB}^2 = 0$$

where, for simplification, we have let $V_G - V_{FB} - V_B = V_{GFB}$. Solving this equation, we get

$$\phi = \left(V_{GFB} + \frac{\gamma^2}{2}\right) - \frac{\gamma^2}{2}\sqrt{1 + \frac{4}{\gamma^2}V_{GFB}}$$

so that saturation occurs at

$$V_D(sat) = V_G - V_{FB} - \phi_B + \frac{\gamma^2}{2}\left[1 - \sqrt{1 + \frac{4}{\gamma^2}(V_G - V_{FB} - V_B)}\right] \qquad (14.8\text{–}8)$$

We can replace V_D in equation (14.8–7) by $V_D(sat)$ to get an analytical expression for saturation current, but the expression would be a lengthy and unhelpful mess. It is sufficient to use (14.8–7) and (14.8–8) concurrently for definition of drain behavior I_D *vs* V_{DS}. Figure 14.8–2 shows a plot of the drain characteristics as defined by these equations in comparison to equations (14.1.1) and (14.1.2), the parabolic model. Both have the same V_{TH} and K. The plot shows that that the parabolic model will usually overestimate the current level unless we compensate it in some other way.

It should be clear that transistors with different body effects will have considerably different drain characteristics. A comparison of drain characteristics with the same V_{TH} and same K, but different body effects, is shown by Figure 14.8–3.

Figure 14.8–3 shows that, in general, it is *not* correct for us to assume that two transistors have the same behavior if they have the same threshold V_{TH} and the same conduction coefficient K. The body–effect makes a huge difference. In defense of the past use of this assumption, however, it is very likely that two transistors with the same V_{TH} and K will be fabricated on the same substrate, and therefore will have the same body effect, and consequently approximately the same drain characteristics.

Figure 14.8–2 Comparison of drain characteristics for the two models, Shichman–Hodges and Meyer, with the same K and V_{TH}. In this case, we have assumed that $\gamma = 1\sqrt{V}, \phi_B = 1V$ and $V_{FB} = -1V$.

Figure 14.8–3 Comparison of drain characteristics for two transistors with the same K and V_{TH}, but different body effects. In this case we have assumed that $\gamma_1 = 1\sqrt{V}, \phi_{B1} = 1V$ and $V_{FB1} = -1V$, and $\gamma_2 = 2\sqrt{V}, \phi_{B2} = 1V$, and $V_{FB2} = -2V$.

14.9 THE INVERSION CONDITION

This section qualifies some of the statements that we made in sections 14.2 and 14.3, where we "interpreted" our conditions for inversion on the basis of the behavior of the electron gas, and identified that inversion can be tagged in terms of a particular value of the surface potential $\phi s = \phi_B$. If we look at the interaction between E–fields and carrier levels in terms of the basic Boltzmann statisitics, then this condition for inversion can be identified.

For an extransic semiconductor, the levels of n–type and p–type charge carriers are related to the energy levels by :

$$p = n_i e^{(E_i - E_F)/kT} = e^{\phi_F/V_T} \quad (14.9\text{–}1)$$

$$n = n_i e^{(E_F - E_i)/kT} = e^{-\phi_F/V_T} \quad (14.9\text{–}2)$$

where n_i is the intrinsic carrier density. If an E–field is applied to the semiconductor, as represented by figure 14.9–1, then the energy changes with respect to distance due to the E–field, and the potentials also change with respect to position. For example the intrinsic potential of the semiconductor $\phi_i = qE_i/kT$ will decrease with respect to position

$$\phi_i = \phi_{io} - \phi$$

where ϕ_{iO} is the potential in the semiconductor far from the influence of the E–field. This added potential subtracts from the difference between E_F and E_i, as represented by figure 14.9–1, resulting in a change of carrier levels as a function of postion. For E–field polarity as shown, the reduction of charge–carrier levels corresponds to an uncovering of the doping sites, which is why we sometimes say that the depletion region is also the "uncovered" region.

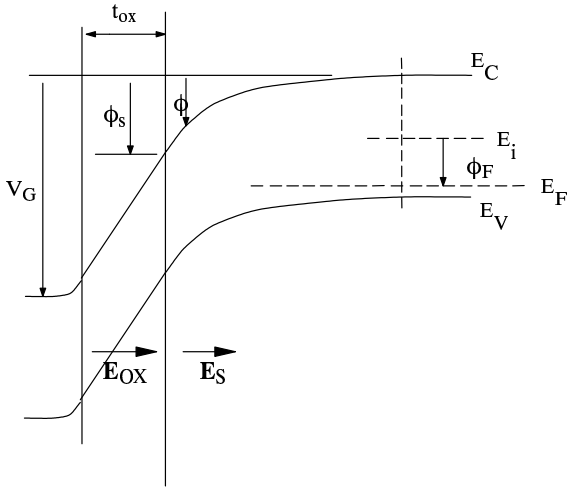

Figure 14.9–1 The nMOS junction under influence of the gate–field, Eox. The band–bending is induced by the effect of the E–field on the semiconductor.

In the semiconductor, both genders of charge–carriers always exist. The charges within the region influenced by the E–field therefore includes the mobile charges as well as the charge centers uncovered by the E–field, i.e.

$$\Delta p(x) = p(x) - p_0 \tag{14.9–3a}$$

represents the depletion of the majority (+) charge carriers (within the p–type substrate), due to effect of the E–field pushing them away from the surface. This difference $\Delta p(x)$ can be expressed in terms of potentials and Boltzmann statistics by

$$\Delta p(x) = p(x) - p_0 = n_i e^{(\phi_F - \phi(x))/V_T} - n_i e^{(\phi_F/V_T} \tag{14.9–3b}$$

Where the potential difference $q\phi_F$ represents the difference of energy between E_i and E_F, as represented by figure 14.9–1, and $\phi(x)$ represents the potential due to the "band–bending", (which is the effect of the E–field).

Similarly,

$$\Delta n(x) = n(x) - n_0 = n_i e^{(\phi(x) - \phi_F)/V_T} - n_i e^{\phi_F} \tag{14.9–4}$$

corresponds to the (–) charge carriers and represents the enhancement of the minority–carrier levels due to the E–field. The overall charge density at any point within the semiconductor field region is then

$$q\rho(x) = q(\Delta p(x) - \Delta n(x)) \tag{14.9–5}$$

Where, using equations (14.9–3b) and (14.9–4),

$$\begin{aligned} \rho(x) &= n_i\left[\left(e^{(\phi_F - \phi(x))/V_T} - e^{\phi_F}\right) - \left(e^{-(\phi_F - \phi(x))/V_T} - e^{-\phi_F/V_T}\right)\right] \\ &= n_i\left[(e^{(u_F - u)} - e^{u_F}) - (e^{u - u_F} - e^{-u_F})\right] \end{aligned} \tag{14.9–6}$$

where, for simplification of the form, we have chosen to let $u_F = \phi_F / V_T$, and $u = \phi(x) / V_T$

Gauss' law requires that

$$\frac{dE}{dx} = \frac{q\rho(x)}{\epsilon_s} \tag{14.9–7}$$

So that

$$\frac{dE}{dx} = -\frac{2qn_i}{\epsilon_s}\left[\left(e^{(u_F-u)} - e^{u_F}\right) - \left(e^{u-u_F} - e^{-u_F}\right)\right] \tag{14.9–8a}$$

The electric field can be expressed in terms of the parameter u, since $E = -d\phi/dx = -V_T\, du/dx$. We can also make use of the "trick" of multiplying both sides by $2E = 2V_T\, du/dx$ to set up this differential equation for easy solution. Equation (14.9–8a) then becomes

$$2E\frac{dE}{dx} = \frac{d}{dx}(E^2) = \frac{2qn_i}{\epsilon_s}\left[\left(e^{u-u_F} - e^{-u_F}\right) - \left(e^{(u_F-u)} - e^{u_F}\right)\right] \times V_T\frac{du}{dx} \tag{14.9–8b}$$

This allows us to find a solution of (14.9–8) in terms of ϕ. Assuming that $E = 0$ deep within the substrate where $\phi = 0$, *and* $E = E_S$ when $\phi = \phi_S$ = surface potential, then

$$E_s = \frac{V_T}{L_I}[e^{-u_F}(e^{u_s} - u_s - 1) + e^{u_F}(e^{-u_s} + u_s - 1)]^{1/2} \tag{14.9–9}$$

where L_I is the *intrinsic* Debye length, given by

$$L_I = \sqrt{\frac{\epsilon_S V_T}{2qn_i}}$$

Since $E_S = Q_S/\epsilon_S$, then equation (14.9–9) also gives us a means of evaluating the charge/area Q_S as a function of u_S. The first parenthesis represents the contribution to the field due to the (–) carriers and the second parenthesis represents the contribution to the field due to the (+) carriers. When the level of (–) "inversion" carriers becomes dominant, then

$$e^{u_s} - u_s - 1 > e^{2u_F}(e^{-u_s} + u_s - 1) \tag{14.9–10}$$

Neglecting terms which are vanishingly small, the boundary on ϕ_S will be when

$$e^{u_s - 2u_F} = u_s - 1 \tag{14.9–11}$$

is met. This equation is transcendental in u_S. It can be solved iteratively. If the result is plotted vs u_F it is nearly linear. As an approximation[14.9–1],

$$u_s = \phi_B/V_T = 2.1u_F + 2.08 \tag{14.9–12}$$

This is the condition for inversion (14.4.6). A plot comparing u_S, the iterated solution of (14.9–11), to u_S as given by equation (14.9–12), is shown by figure 14.9–2. The two results overlap to the extent that they appear to be the same.

Figure 14.9–2 Comparison of inversion conditions vs N_{sub}

Since $Q_S = \varepsilon_S E_S$, equation (14.9–9) is also appropriate for determination of the capacitance of the semiconductor, $C_S = \partial Q_S/\partial \phi_S$. From equation (14.9–9) we have

$$Q_s = \epsilon_s E_s = V_T \times \frac{\epsilon_s}{L_I}[e^{-u_F}(e^{u_s} - u_s - 1) + e^{u_F}(e^{-u_s} + u_s - 1)]^{1/2} \tag{14.9–13}$$

This can also be expressed as

$$Q_s = V_T \times \frac{\epsilon_s}{L_B}\sqrt{2}\left[e^{-2u_F}(e^{u_s} - u_s - 1) + (e^{-u_s} + u_s - 1)\right]^{1/2} \tag{14.9–14}$$

If we use the relationship $e^{u_F} = N_{SUB}/n_i$. As was seen by section 14.2, this information is sufficient to determine the capacitance of the MOS junction C_{MOS} as a function of V_G by means of the parametric equations

$$C_{ox}(u_G - u_s) = \frac{\epsilon_s}{L_B}\left[e^{-2u_F}(e^{u_s} - u_s - 1) + (e^{-u_s} + u_s - 1)\right]^{1/2} \tag{14.9–15}$$

and

$$C_S = \frac{1}{V_T}\frac{\partial Q_S}{\partial u_S} \tag{14.9–16}$$

where $u_G = V_G/V_T$. Since (14.9–15) is transcendental in u_S, it is not possible to obtain an analytical form for u_S in terms of V_G, or C_s in terms of V_G. They can be used parametrically to plot C_s *vs* V_G (using u_S as the parametric variable). The capacitance of the MOS junction plotted in figure 14.2.4 uses equations (14.9–15), (14.9–16) and (14.2.3).

14.10 CAPACITANCES FOR THE MOS TRANSISTOR

The MOS transistor, by nature of its construction, is a capacitative structure. The structure has four terminals, and charge is controlled by these terminals. In this respect, we have to recognize that we have a more than just a transistor. We also have a capacitance *matrix*, with components of the form

$$C_{JK} = \frac{dQ_J}{dV_K} \tag{14.10–1}$$

where J and K are the nodes of the transistor, *G, S, D, B* .

The major circuit effects are associated with the active charge layer q_I and its dynamics. The total inversion charge controlled by the gate is

$$Q_G = W\int_O^L q_I dy \tag{14.10–2}$$

For the sake of simplicity we take the parabolic model assumption that the threshold V_{TH} and the conductance coefficient K_P are both constant. Then

$$q_I = C_{OX}(V_G - V_{TH} - V) \tag{14.10–3}$$

A link between current I and the inversion charge/area q_I is given by equation (14.8.5), written in a more compact form as

$$I = W\mu_S q_I \frac{\partial V}{\partial y} \tag{14.10–4}$$

Note that when equation (14.10–4) is integrated from V_S to V_D, we get the parabolic model equation

$$I = \frac{1}{2}\mu_S C_{OX}\frac{W}{L}\left[(V_{GS} - V_{TH})^2 - (V_{GD} - V_{TH})^2\right] \tag{14.10–5}$$

We can apply (14.10–4) and (14.10–3) to equation (14.10–2) to get Q_G as a function of V_S and V_D. This gives

$$Q_G = \frac{W^2 C_{OX}^2}{I} \int_{V_S}^{V_D} (V_G - V_{TH} - V)^2 dV \tag{14.10–6}$$

When this integration is carried out, and (14.10–5) is included, the form reduces to

$$Q_G = \frac{2}{3} WL\, C_{OX} \frac{V_{GST}^3 - V_{GDT}^3}{V_{GST}^2 - V_{GDT}^2} \tag{14.10–7}$$

where we have defined $V_{GST} = V_{GS} - V_{TH}$ and $V_{GDT} = V_{GD} - V_{TH}$, to keep the form as uncomplicated as possible. This result can be reduced by means of factoring to a relatively simple form:

$$Q_G = \frac{2}{3} C_0 \frac{V_{GST}^2 + V_{GST} V_{GDT} + V_{GDT}^2}{V_{GST} + V_{GDT}} \tag{14.10–8}$$

For convenience, we have let $C_O = WLC_{ox}$. Capacitances can be obtained by derivatives of (14.10–8) with respect to each of the voltages V_G, V_S, and V_D. These capacitances are called the *static gate* capacitances, since the derivation of (14.10–8) assumes steady–state current flow. These capacitances are given by table 14.8.1.

Table 14.10–1 STATIC MOSFET CAPACITANCES

TERM	FORM
$C_{GG} = \frac{\partial Q_G}{\partial V_G}$	$\frac{2}{3} C_O \frac{a^2 + 4a + 1}{(1 + a)^2}$
$C_{GS} = \frac{\partial Q_G}{\partial V_S}$	$\frac{2}{3} C_O \frac{2a + 1}{(1 + a)^2}$
$C_{GD} = \frac{\partial Q_G}{\partial V_D}$	$\frac{2}{3} C_O \frac{a(a + 2)}{(1 + a)^2}$

where we have used $a = V_{GDT}/V_{GST}$ to keep the forms simple. Note that a defines the triode and saturation regimes in the limits:

saturation: a → 0 for V_{DS} → large, or as $V_{GS} \rightarrow V_{TH}$

triode: a → 1 as $V_{DS} \rightarrow 0$, or as V_{GS} → large

Table 14.10.1 is called the *Meyer model* for MOSFET capacitances.

In the case where we do not have steady–state current, and inasmuch as the MOSFET is subject to charge and discharge of the inversion layer by the biases at the V_G, V_S and V_D nodes, dQ/dt terms must be considered, for which the continuity equation applies:

$$\frac{\partial I}{\partial y} + W \frac{\partial q_I}{\partial t} = 0 \tag{14.10–9}$$

Continuity identifies the operation of the MOSFET when the current I_S is not equal to I_D. It represents the conditions that must be met under *quasi–static* conditions, where charge and discharge of the inversion layer occurs.

This situation is represented by figure 14.10–1, which shows the MOSFET as a device with a "core" of free charge that is supplied and/or drained by source and drain currents I_S and I_D. Integration of (14.10–8) yields

$$\int_{V_S}^{V} dI + W\frac{d}{dt}\int_0^y q_I\, dy' = 0$$

for which, assuming that $I_S = I(V_S)$, and $I(V)$ given by (14.10–4), gives

$$I_S = \mu W q_I \frac{\partial V}{\partial y} + W\frac{d}{dt}\int_0^y q_I dy' \tag{14.10–9}$$

I_S is the current out of the source, and is independent of y. When (14.10–9) is integrated with respect to y, for $0 < y < L$, we get

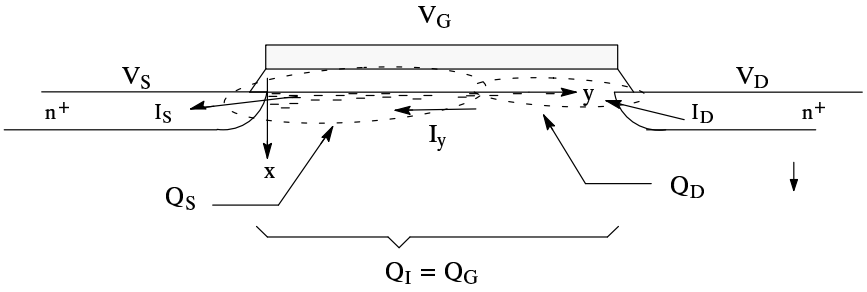

Figure 14.10–1. Quasi–static charges and currents

$$I_S = \mu\frac{W}{L}\int_{V_S}^{V_D} q_I dV + \frac{dQ_S}{dt}$$

The first term on the left–hand side represents the channel current and the second term represents the effect of the current on the inversion charge. Q_S is then a partition of the charge Q_G, of measure

$$Q_S = \frac{1}{L}\int_0^L W\int_0^y q_I dy' dy$$

This equation can be integrated by parts to yield

$$Q_S = W\int_0^L \left(1 - \frac{y}{L}\right) q_I dy \tag{14.10–10}$$

If we make use of (14.10–4) to integrate from 0 to y while V goes from V_S to V, then

$$\frac{y}{L} = \frac{V_{GST}^2 - V_{GT}^2}{V_{GST}^2 - V_{GDT}^2} \tag{14.10–11}$$

where we have let $V_G - V_{TH} - V = V_{GT}$ for convenience. If (14.10–11) and (14.10–4) are applied to (14.10–10) then we will get Q_S in term of node voltages V_S, V_D, and V_G, as follows:

$$Q_S = \frac{2}{15} Co\left[\frac{3V_{GST}^5 - 5V_{GST}^3 V_{GDT}^2 + 2V_{GDT}^5}{(V_{GST}^2 - V_{GDT}^2)^2}\right]$$

This equation can be reduced by factoring to

$$Q_S = \frac{2}{15} Co \left[\frac{3V_{GST}^3 + 6V_{GST}^2 V_{GDT} + 4V_{GST} V_{GDT}^2 + 2V_{GDT}^3}{(V_{GST} + V_{GDT})^2} \right] \qquad (14.10\text{–}12)$$

where $C_O = WL\, C_{ox}$, as before.

In like manner, equation (14.10–8) can be integrated from y to L, corresponding to the channel voltage from V to V_D. This analysis leads to definition of quasi–static channel charge associated with the drain end of the channel

$$Q_D = W \int_0^L \frac{y}{L} q_I dy \qquad (14.10\text{–}13)$$

which, using equations (14.10–11) and (14.10–4), reduces to

$$Q_D = \frac{2}{15} Co \left[\frac{2V_{GST}^3 + 4V_{GST}^2 V_{GDT} + 6V_{GST} V_{GDT}^2 + 3V_{GDT}^3}{(V_{GST} + V_{GDT})^2} \right].$$

Charges Q_S and Q_D given by (14.10–10) and (14.10–13), respectively, add up to (14.10–2), the channel charge Q_G. In this respect, the channel charge is said to be *partitioned* in terms of a source partition Q_S and a drain partition Q_D. This partitioning is also called the 60/40 partition since, in the saturation limit where $V_{GDT} \rightarrow 0$, $Q_S/Q_G \rightarrow 2/3$ and $Q_D/Q_G \rightarrow 1/3$.

These charge partitions also define six more terms of the MOSFET capacitance matrix, as represented by table 14.8.2.

Table 14.10–2. QUASI–STATIC MOSFET CAPACITANCES

TERM	FORM
$C_{GG} = \frac{\partial Q_G}{\partial V_G}$	$\frac{2}{3} C_O \frac{a^2 + 4a + 1}{(1 + a)^2}$
$C_{GS} = \frac{\partial Q_G}{\partial V_S}$	$\frac{2}{3} C_O \frac{2a + 1}{(1 + a)^2}$
$C_{GD} = \frac{\partial Q_G}{\partial V_D}$	$\frac{2}{3} C_O \frac{a(a + 2}{(1 + a)^2}$
$C_{SG} = \frac{\partial Q_S}{\partial V_G}$	$\frac{2}{15} C_O \frac{2a^3 + 14a^2 + 11a + 3}{(1 + a)^3}$
$C_{SS} = \frac{\partial Q_S}{\partial V_S}$	$\frac{2}{15} C_O \frac{8a^2 + 9a + 3}{(1 + a)^3}$
$C_{SD} = \frac{\partial Q_S}{\partial V_D}$	$\frac{2}{15} C_O \frac{2a(a^2 + 3a + 1)}{(1 + a)^3}$
$C_{DG} = \frac{\partial Q_D}{\partial V_G}$	$\frac{2}{15} C_O \frac{3a^3 + 11a^2 + 14a + 2}{(1 + a)^3}$
$C_{DS} = \frac{\partial Q_D}{\partial V_S}$	$\frac{2}{15} C_O \frac{2(a^2 + 3a + 1)}{(1 + a)^3}$
$C_{DD} = \frac{\partial Q_D}{\partial V_D}$	$\frac{2}{15} C_O \frac{a(3a^2 + 9a + 8)}{(1 + a)^3}$

The quasi–static gate charge Q_G is the same, whether considering the static case, or the quasi–static case. A plot of each of these capacitances is shown by Figure 14.10–2. Note that the capacitances are *non–reciprocal*, i.e. C_{SG} is *not* equal to C_{GS}.

Figure 14.10–2. MOSFET capacitances from Table 14.10–2. Capacitance vs V_{DS} is plotted for the case where $V_{GS} = 5.0V$ and $V_{TH} = 0.8V$.

Features of these capacitances that are of importance are their values in the limit as $a \rightarrow 0$, corresponding to saturation. In this limit

$$C_{GG} \rightarrow C_{GS} \rightarrow C_{SG} \rightarrow C_{SS} \rightarrow \frac{2}{3} C_O$$

$$C_{GD} \rightarrow C_{SD} \rightarrow C_{DD} \rightarrow 0$$

$$C_{DG} \rightarrow C_{DS} \rightarrow \frac{4}{15} C_O$$

If we are operating in the saturation regime, as is usually the case for linear circuits, then the only capacitance from gate to drain is the *overlap* capacitance

$$C_{OL} = W \times C_W \tag{14.10–15}$$

where C_w is the overlap capacitance per cm along the gate edge. This term must also be added to each of the other terms where either the gate node V_G or the gate charge Q_G is concerned. SPICE uses CGSO for gate–source overlap per meter, and CGDO for gate–drain overlap per meter. In saturation, $C_{GS} \approx W \times$ CGDO.

14.11 HIGH–FIELD EFFECTS – VELOCITY SATURATION

At high E–fields, the linear relationship between voltage and current, Ohm's law, begins to deteriorate. Since, in the analysis of current I_D in the FET channel we relied upon Ohm's law to define the current–voltage relationships, we need to re–evaluate this analysis when we consider short–channel devices, where the driving fields are high everywhere within the channel.

Turning to the basic definition of current, we find that it relates to the velocity *v* of the charge–carriers, as

$$J = qnv \tag{14.11–1}$$

where, in this case, we have assumed n–type charge carriers. In the nMOSFET, conduction current flows through an thin inversion layer created by the gate field, of charge density n_I. Current within the channel will therefore be of the form

where *v* is the velocity of carriers. At low fields, velocity *v* is proportional to the E–field E_y, according to

$$v = \mu E_y$$

which gives us our basic definition of mobility, and eventually leads to Ohm's law, *(J = σE)*. When we have high E–fields, this Ohm's law equation is no longer applicable.

The effect of increasing the E–field is represented by figure 14.11–1. There is a natural physical limit to the velocity of the charge carriers, and as the E–field is increased, the limit is approached asymptotically. This terminal velocity is approximately the thermal velocity of the "gas" of charge carriers within the semiconductor, on the order of 10^7 cm/s.

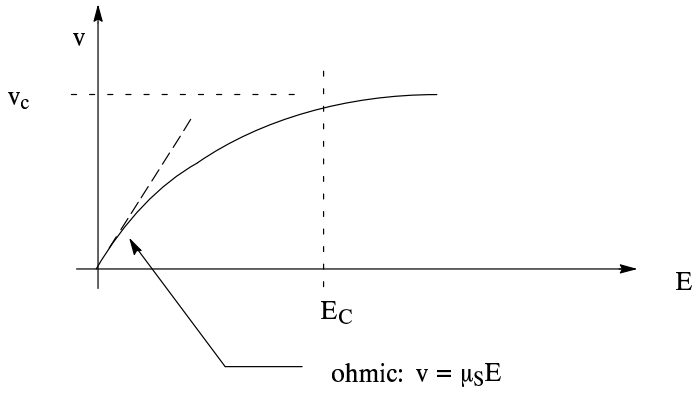

Figure 14.11–1 Velocity limiting of charge–carriers.

The velocity–limiting feature occurs within all FET transistors. When the transistor reaches its saturation condition, and $I = I_{DSAT}$, then q_I –> *0* somewhere near the drain node. As q_I –> *0*, *v* must increase. E_y which, at low fields, is ≈ *v*/μ, therefore will also increase in the vicinity of this "pinch–off", eventually reaching its own limit according to the value of *v*.

The outcome of this "velocity–saturation limit" to the current is that at or near the drain node, a high–field region and a channel charge limit are approached:

$$q_I \rightarrow q_c = \frac{I_{DST}}{v_c W} \tag{14.11–2}$$

Provided that this limit is localized to the drain region, equation (14.11–2) implies that we merely will see a small correction term to V_G as used in equation (14.8.8) for the definition of V_{DSAT}. The correction is $\Delta V_G = -q_C/C_{ox}$.

We also can assume that the mobility, which is the slope of the curve represented by Figure 14.11–1, asymptotically goes to zero. A simple interpretation of this behavior is given by

$$\mu_S = \frac{\mu_0}{1 + aE_y} = \frac{\mu_0}{1 - adV/dy} \tag{14.11–3}$$

where $a \approx \pm 1/E_c$, the sign being negative if the charge–carriers are negative. E_c represents a value of E_y at which the velocity limiting effects become dominant. When this expression for μ_S is used in equation (14.8.5) we get

$$I\left[1 + \frac{1}{E_C}\frac{\partial V}{\partial y}\right] = W\mu_0 q_I \frac{\partial V}{\partial y} \tag{14.11–4}$$

Since (14.11–4) is a linear differential equation, it can readily be integrated, which gives the result

$$I_D = W_L \frac{\mu_0 C_{OX}}{1 + V_{DS}/(E_C L)}\left[(V_{GS} - V_{BIN})V_{DS} - \frac{1}{2}V_{DS}^2 + \frac{2}{3}\gamma[(\phi_B + V_{DB})^{3/2} - (\phi_B + V_{SB})^{3/2}]\right]$$

The important distinction between equations (14.8.6) and (14.11–5) is that the mobility is voltage–dependent, as may be approximately represented by

$$\mu_S = \frac{\mu_0}{1 + V_{DS}/(E_C L)}$$

The effect is stronger for smaller channel lengths L.

However, if the this charge–limiting effect extends over a large portion of the channel, which is more likely for short–channel devices, then it is necessary to define I_{DSAT} in terms of (14.11–2), and even equation (14.8.7) is inadequate. This implies that the level of charge q_I itself is the defining quantity for current.

Therefore when I_{DSAT} is defined by the "velocity–saturation" effect, it will be approximately linear in $V_{GS} - V_{TH}$, rather than quadratic behavior indicated by equation (14.1.2).

This comparison is represented by figure 14.11–2.

Figure 14.11–2a Long–channel device. I_{DSAT} determined by charge–control analysis.

Figure 14.11–2b Short–channel device. I_{DSAT} determined by velocity limiting of charge–carriers.

14.12 A PARABOLIC APPROXIMATION – THE BSIM MODEL

The Schichman–Hodges model, given by equations (14.1–1) and (14.1–2), is sufficiently appealing in its simplicity and numerical speed so that quadratic modifications to the physically accurate charge–control model are often used. One of the more comprehensive expansions of this form, BSIM, is developed after a parabolic model, CSIM (Compact, Short–channel IGFET Model) developed at Bell Labs. The CSIM model was expanded and implemented as a table–structured model by a research group at UC Berkeley[14.12–1] and renamed BSIM. It has 54 parameters, many of which are statistical expansions in *1/L* and *1/W,* which is a means of including a wide range of short–channel effects. This approach to a comprehensive model of the MOSFET is sometimes called the "statistical" model.

The CSIM model is developed by expansion of the body–effect part of the Meyer model, equation (14.8.7). The body–effect part of equation (14.8.7) is given by

$$x_B = \frac{2}{3}\gamma[(\phi_B + V_{DB})^{3/2} - (\phi_B + V_{SB})^{3/2}] \qquad (14.12.–1)$$

If we direct our attention to the first term of the body effect, we can make the expansion

$$(\phi_B + V_{DB})^{3/2} = (\phi_B + V_{DS} + V_{SB})^{3/2} \approx (\phi_B + V_{SB})^{3/2}\left[1 + \frac{3}{2}\frac{V_{DS}}{(\phi_B + V_{SB})} + \frac{3}{8}\frac{V_{DS}^2}{(\phi_B + V_{SB})^2}\right]$$

We might feel a little unsure about this expansion since the higher–order terms in $V_{DS}{}^3$, ..etc. are not necessarily negligible as we increase V_{DS}. But assuming that corrective terms can be applied through judicious choice of coefficients, this expansion at least gives a form to equation (14.12–1) which is second–order in V_{DS}:

$$x_B = \frac{2}{3}\gamma\left[(\phi_B + V_{DB})^{3/2}\right] \approx \gamma\left[V_{DS}\sqrt{\phi_B + V_{BS}} + \frac{V_{DS}^2}{4\sqrt{\phi_B + V_{SB}}}\right] \qquad (14.12–2)$$

Where we have subtracted away the common terms in $\sqrt{\phi_B + V_{SB}}$. This equation can be combined with the first part of equation (14.8.7) to give a parabolic equation form:

$$I_D \approx K_p\frac{W}{L}\left((V_{GS} - V_{BI})V_{DS} - \frac{1}{2}V_{DS}^2 - \gamma\left[V_{DS}\sqrt{\phi_B + V_{BS}} + \frac{V_{DS}^2}{4\sqrt{\phi_B + V_{SB}}}\right]\right)$$

$$= K_p\frac{W}{L}\left(V_{GS} - V_{FB} - \phi_B - \gamma\sqrt{\phi_B + V_{BS}}\right)V_{DS} - \frac{1}{2}aV_{DS}^2\Big)$$

$$= K_p\frac{W}{L}\left((V_{GS} - V_{TH})V_{DS} - \frac{1}{2}aV_{DS}^2\right) \qquad (14.12–3)$$

where the *conductance–degradation* coefficient *a* includes the quadratic (coefficient to V_{DS}^2) component of the body–effect expansion of equation (14.12–2),

$$a = 1 + \frac{g\gamma}{2\sqrt{\phi_B + V_{SB}}} \qquad (14.12–4)$$

Note that this equation also includes the factor *g*, which is a numerical expansion coefficient which is included to accommodate the non–negligible terms neglected by the expansion (14.12–2). *g* is given by

$$g = 1 - \frac{1}{1.744 + 0.8364(\phi_B + V_{SB})} \; . \qquad (14.12–5)$$

Although it is well thought–out, this term is also known as a “fudge–factor”, and therefore the BSIM model is not without critics.

The advantage of the model is that it is of the simple form

$$I_D = \beta\left((V_{GS} - V_{TH})V_{DS} - \frac{1}{2}aV_{DS}^2\right) \qquad (14.12–6)$$

where $V_{TH} = V_{FB} + \phi_B + \gamma\sqrt{\phi_B + V_{BS}}$ is the same as the charge–control form of V_{TH} given by equation (14.4.13), and where β is the conduction coefficient.

Actually, V_{TH}, as used by BSIM, is not strictly like equation (14.4.13). It adds correction terms to the charge–control V_{TH} for short–channel and narrow–channel effects. It even includes a drain–induced barrier–lowering effect due to V_{DS}.

The BSIM model also lends itself to a parabolic form for the saturation current, I_{DSAT}, much like equation (14.1.2). This is achieved by defining

$$V_{DSAT} = \frac{V_{GS} - V_{TH}}{a} \tag{14.12–7}$$

When applied to (14.12–6) this choice for V_{DSAT} gives saturation current

$$I_{DSAT} = \frac{\beta}{2a}(V_{GS} - V_{TH})^2 \tag{14.12–8}$$

Since these equations are relatively simple they may be modified to include velocity–saturation effects by returning to the velocity–saturation limit on charge, q_I –> q_C. Since q_C is a limit on I_{DSAT} imposed by the thermal limit of carrier velocity, *vc*, it can be restated, for $V_{TH} \approx$ const, as:

$$qc = \frac{I_{DSAT}}{Wv_c} = C_{OX}(V_{GS} - V_{TH} - aV_{DSAT}) \tag{14.12–9}$$

This equation therefore gives us a link between V_{DSAT} and I_{DSAT} which we can exploit using equation (14.12.6), as follows:

$$\begin{aligned} I_{DSAT} &= \beta\left((V_{GS} - V_{TH} - rI_{DSAT})V_{DSAT} - \frac{1}{2}aV_{DSAT}^2\right) \\ &= \frac{\beta}{2a}(V_{GS} - V_{TH} - rI_{DSAT})^2 \end{aligned} \tag{14.12–10}$$

where $r = 1/(v_c WC_{ox})$.

By means of a trick, wherein we assume that I_{DSAT} is of the form:

$$I_{DSAT} = \frac{\beta}{2aK}(V_{GS} - V_{TH})^2 \tag{14.12–11}$$

we can then apply this definition of I_{DSAT} to equation (14.12–10), which gives us a quadratic equation in terms of the factor, *K*,

$$K^2 - K\left(1 + \frac{r\beta}{a}(V_{GS} - V_{TH})\right) + \left(\frac{r\beta}{2a}\right)^2(V_{GS} - V_{TH})^2$$

This equation has solution

$$K = \frac{1}{2}(1 + p + \sqrt{1 + 2p}) \tag{14.12–12}$$

Parameter *p* relates to $V_{GS} - V_{TH}$ and the effect of velocity–limited saturation by

$$p = \frac{r\beta}{a}(V_{GS} - V_{TH}) = \frac{\beta}{aWv_cC_{OX}}(V_{GS} - V_{TH}) = \frac{\mu_S}{av_cL}(V_{GS} - V_{TH}) \tag{14.12–13}$$

Admittedly, this process is somewhat manipulative, since it is hiding some of the voltage dependence of I_{Dsat} under parameter K. It means that (14.12–11) should not to be interpreted as a simple quadratic function, because parameter K relates to p, which is linear in $V_{GS} - V_{TH}$.

If $p << 1$, corresponding to L –> *large*, then

$$K \approx 1 + p = 1 + \frac{\mu_S}{v_c L} \frac{V_{GS} - V_{TH}}{a}$$

and I_{DSAT} is approximately quadratic in $V_{GS} - V_{TH}$, which is consistent with *long–channel* behavior.

If $p >> 1$, corresponding to L –> *small*, then

$$K \approx \frac{p}{2} = \frac{\mu_S}{v_c L} \frac{V_{GS} - V_{TH}}{2a}$$

then I_{DSAT} is approximately linear in $V_{GS} - V_{TH}$, which is consistent with *short–channel* behavior.